Regression

5

In regression analysis, a priori information about the experimental conditions, in the form of independent variables, is used to predict the averages, and possibly also the standard deviations, for the observed, or dependent, variables.

Section 1

Simple Regression

Regression With One Independent Variable

1. **Simple regression is linear regression with one independent variable**
2. **Simple regression uses two regression coefficients, the offset and gain, to relate the averages in the different conditions to the corresponding values of the independent variable**
3. **Simple nonlinear regression combines simple regression with a transformation on the independent variable**
4. **Generalized linear models use a nonlinear transformation on the predicted averages and also predict the variation in standard deviation**

Linear Regression

In all the examples of the previous chapters, the averages of the distributions that were used to model the observed histograms were parameters that could be optimized to improve the agreement with the observed histograms. This is customary in case of **exploratory data analysis**, i.e., when there is little or no advance knowledge on how averages change over conditions. However, experiments are frequently set up to perform **confirmatory data analysis**, i.e., to explore the evidence for a priori expected relationships between **independent variables**, i.e., a priori known information about the observed conditions, and **dependent variables** (which are summarized into observed histograms).

In **regression analysis**, the goal is to explore models for predicting the averages of the dependent variables in the distinct conditions. More specifically, in this chapter we will be interested in the important specific case of linear regression, which means that the actual averages $\{\mu_k; k = 1,\ldots,K\}$ in the K experimental conditions will be approximated by model averages

$$\left\{\hat{\mu}_k = \beta_0 + \sum_{l=1}^{L} \beta_l \cdot v_{kl}; k = 1,\ldots,K\right\}$$

that are linear combinations of the independent variables. Note that there are L independent variables in the above expression

and that $\{v_{kl}, l = 1,\ldots,L\}$ are the values of these independent variables for condition k, where $k = 1,\ldots,K$. The $L+1$ **regression coefficients**

$$\{\beta_l; l = 0,1,\ldots,L\}$$

are the parameters in such a linear regression model and replace the K model averages used in exploratory data analysis, which implies that linear regression assumes that $L+1 \leq K$, as otherwise the number of parameters would exceed the number of conditions.

In the current section, we will discuss the case of **simple regression**, where there is only one independent variables ($L = 1$), while the case of **multiple regression**, where $L > 1$, will be treated in the next section.

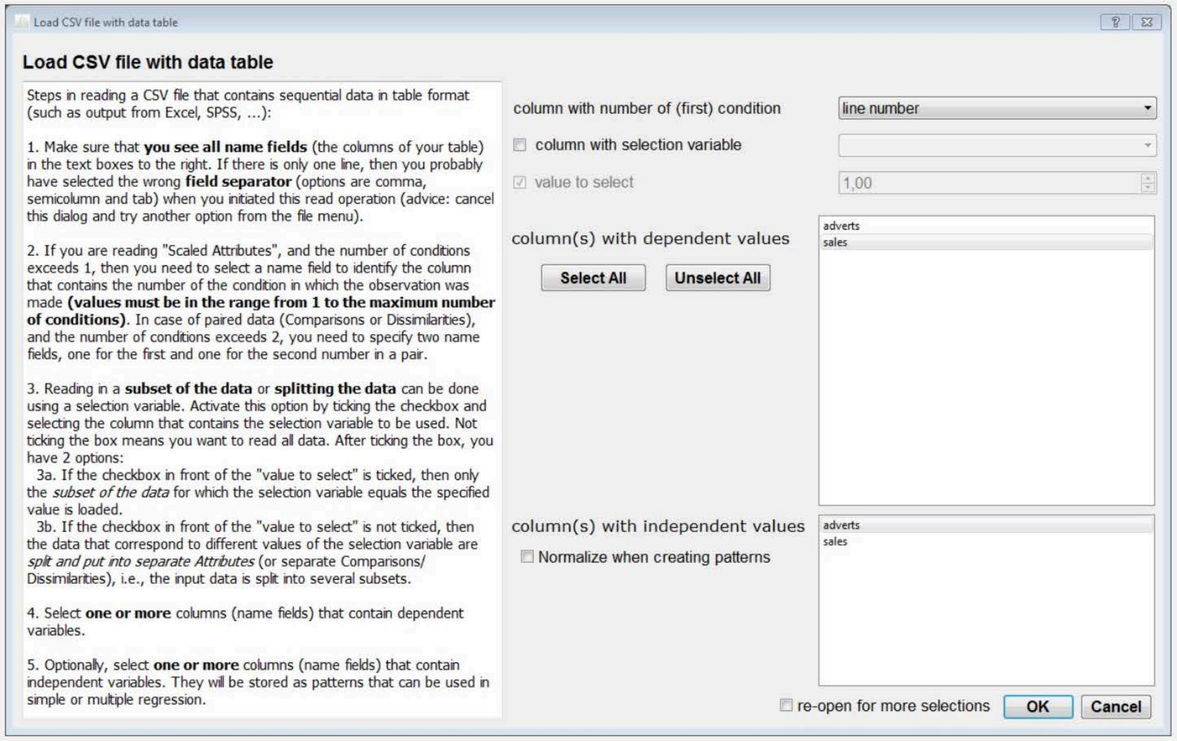

FIGURE 5.1.1 Dialog window showing how to set up simple regression with one dependent and one independent variable (data: *Record1.csv*)

Simple Regression in ILLMO

In order to illustrate simple regression, we again take the same example as used in the book(s) by Field to introduce this topic. The file *Record1.csv* contains two columns and 200 measurements. The first column entitled “adverts” contains the advertising budget (in thousand of pounds) for a record, which is treated as the independent variable, while the second column entitled “sales” contains the record sales (in thousands) as dependent variable. The a priori expectation is that the record sale will increase (linearly) with the investment in advertising. The file can be read using option “*Open CSV file (with attributes, field separator:semicolon)*” from the *File* menu, and should be interpreted as NOT containing data in separate columns. The CSV dialog window can be set as shown in Fig-

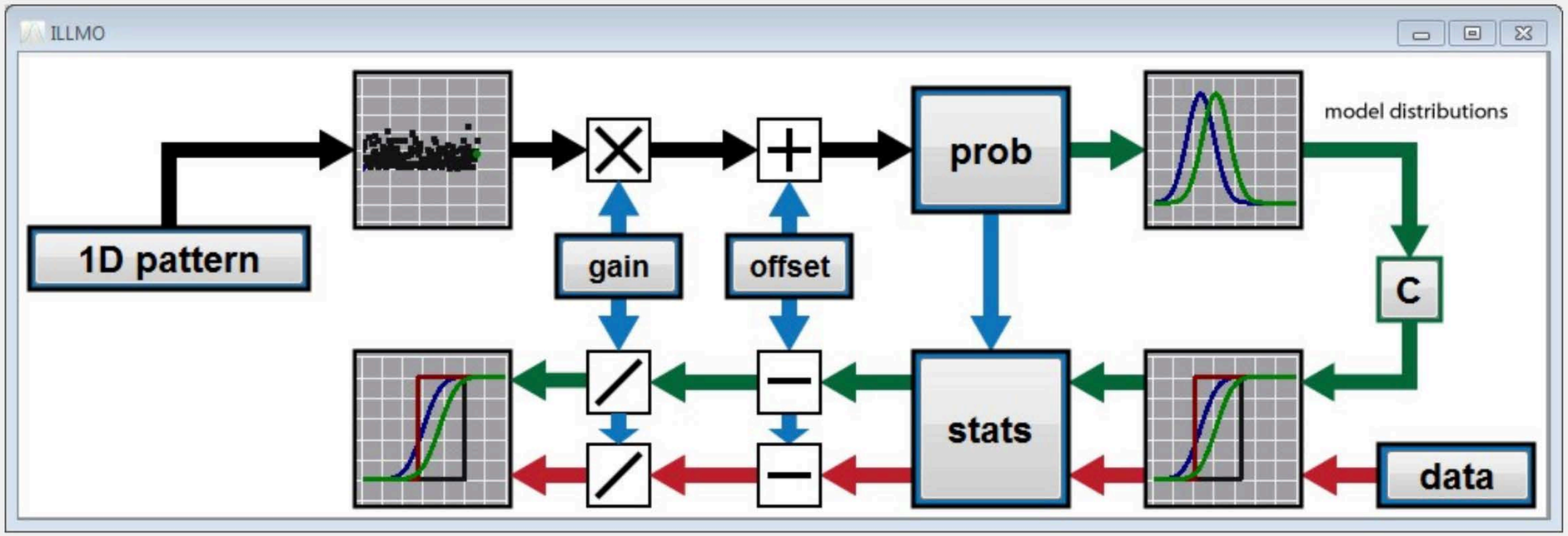

FIGURE 5.1.2 Simple regression model as visualized by ILLMO (data: *Record1.csv*)

ure 5.1.1, i.e., the line number (from 1 to 200) can be used to indicate the condition, the column "sales" contains the dependent values, while the column "adverts" contains the independent values.

The complete regression model, which can be exported from ILLMO by pressing on the "camera" icon on the right in the toolbar, or by selecting the corresponding entry in the pop-up menu that appears when right-clicking in ILLMO, is illustrated in Figure 5.1.2.

The **1D pattern** operator provides the independent variable "adverts" and can be inspected more closely by selecting the Stimulus Patterns tab in the ILLMO interface. The small graph in the upper left in Figure 5.1.2. also shows this pattern

$$\{v_k; k = 1,\ldots,K\},$$

while the **gain** and **offset** buttons allow to inspect the regression coefficients that where previously termed β_1 and β_0, respectively. The predicted averages

$$\hat{\mu}_k = \beta_0 + \beta_1 \cdot v_k$$

are used by the **prob** operator to generate Gaussian distributions with a constant standard deviation σ that is an additional parameter of the model. These Gaussian distributions are matched to the observed histograms, which are step functions in the current example, as only a single measurement is made in each of the $K = 200$ conditions. As usual, only the histograms for the reference (k=1) and the selected (k=200) condition are included in this model rendering.

As not all elements of this regression model can easily be visualized at the same time on most computer screens, ILLMO divides the model into two parts. As usual, the right part (starting from the **prob** and **stats** buttons) can be inspected by selecting the Data (Exploratory) subpage on the Scaled Attributes page of the ILLMO interface, while the left part (up to the **prob** and **stats** buttons) can be inspected by selecting the Model (Confirmatory) subpage.

A scatterplot, such as the one shown in Figure 5.1.3, is a frequently-used way to visualize the relationship between an independent and a dependent variable. This plot was generated by selecting the plot option "*Model vs data*". An alternative is to render the ***prediction error*** (option "*Model vs data-model*"), as shown in Figure 5.1.4. In this latter rendering, the prediction error (or ***residue value***) is obtained by subtracting the predicted

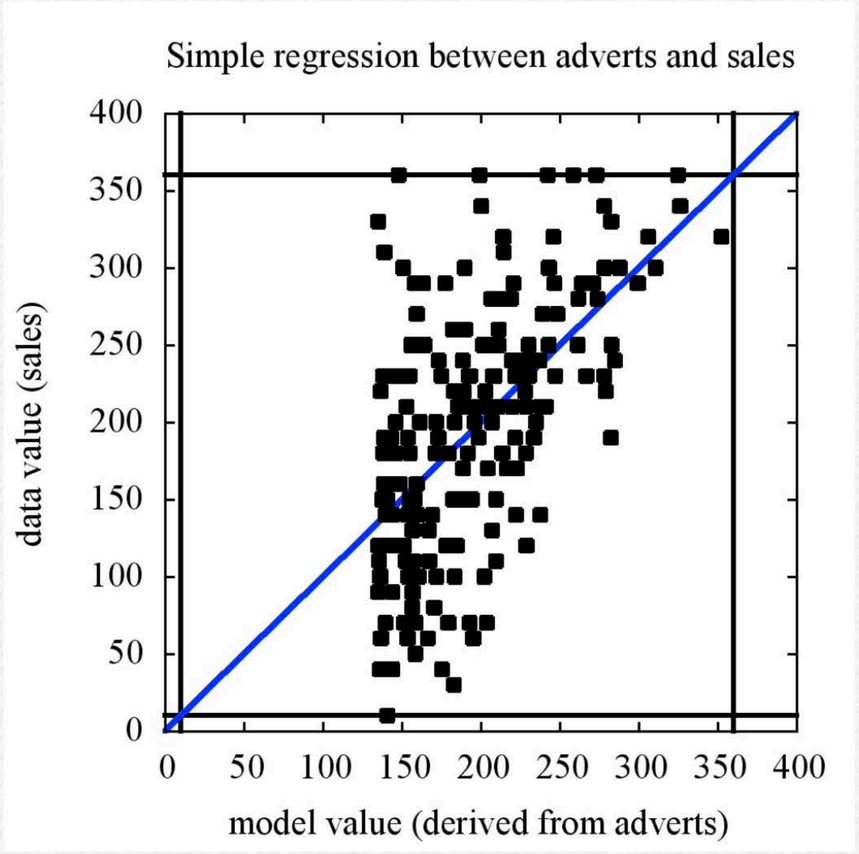

FIGURE 5.1.3 Scatterplot of data values (sales) versus the model value, which is a linear prediction derived from known adverts (data: *Record1.csv*)

average for a specific condition from the observed data values in that condition.

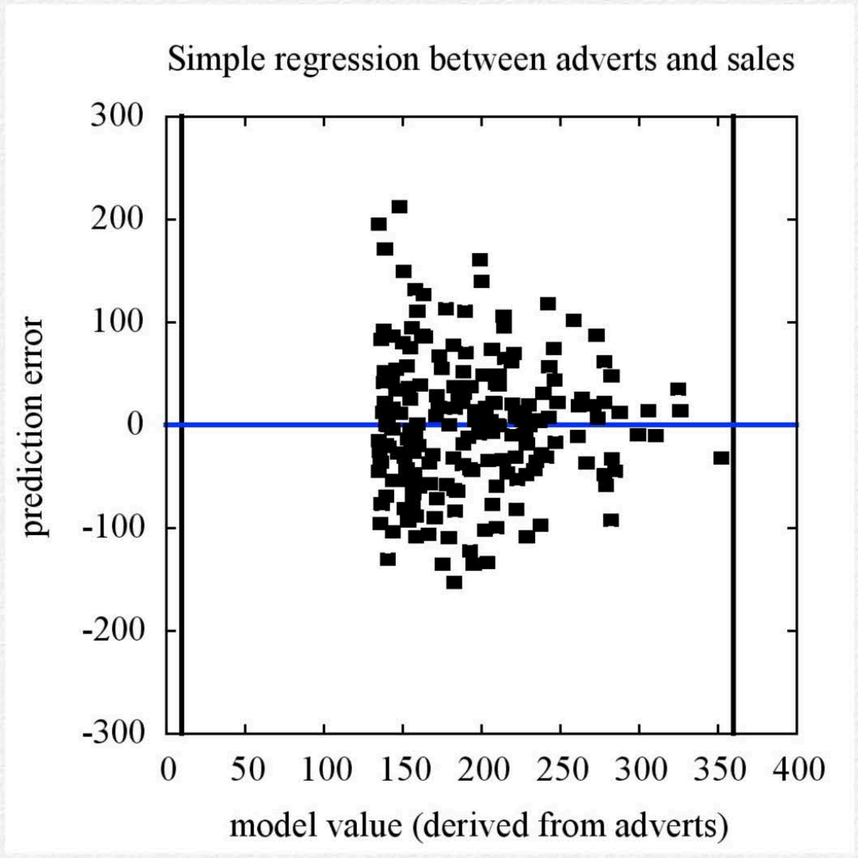

FIGURE 5.1.4 Scatterplot of prediction errors versus the model value, which is a linear prediction derived from known adverts (data: *Record1.csv*)

The statistical analysis of the simple regression can again be done in two ways, either using the general methodology of log-likelihood estimation advertised in this book, or using more traditional statistics that are, as always up to now, based on the assumption of continuous data with Gaussian distributions of constant variance. In this section, we discuss the log-likelihood approach favored by ILLMO.

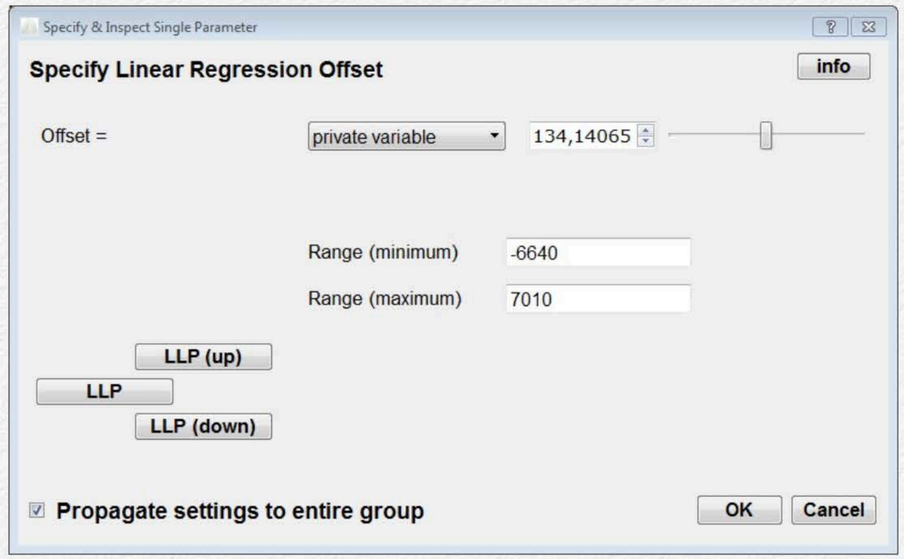

FIGURE 5.1.5 Dialog window for the offset

Clicking on the **offset** button opens the dialog box shown in Figure 5.1.5. It reveals the estimated offset and offers the option to request the LLP for this coefficient, which produces the graphical result of Figure 5.1.6, from which it can be deduced that

$\beta_0 = 134.14$, with 95% CI = [119.37,148.91].

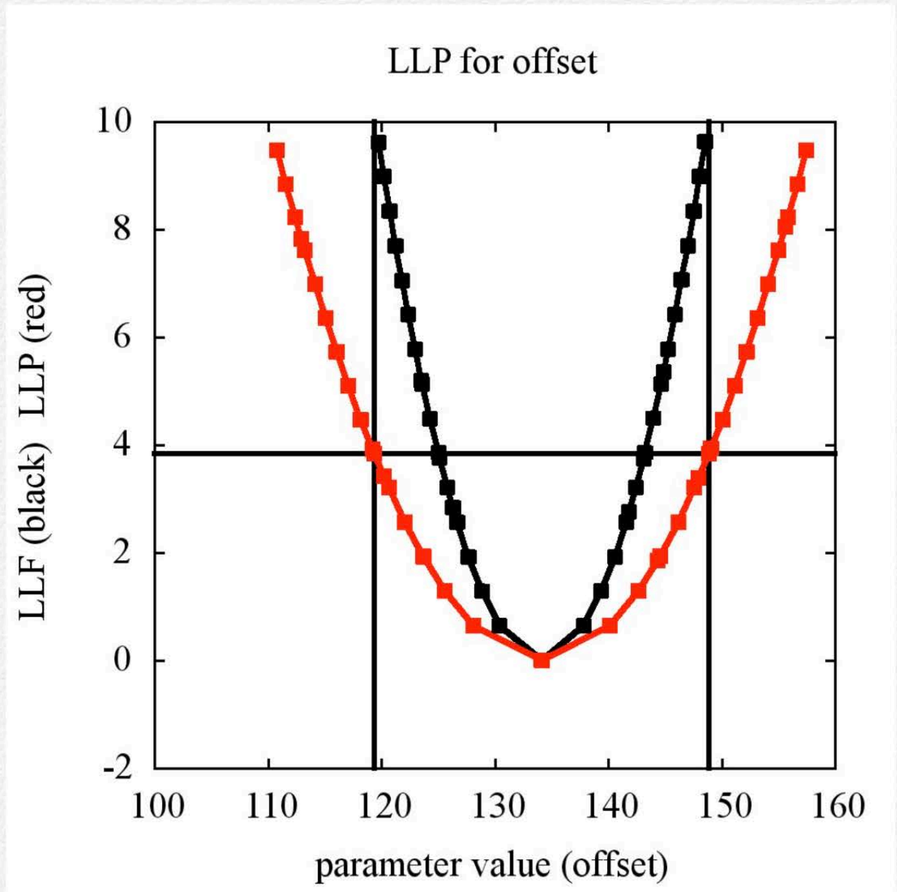

FIGURE 5.1.6 LLP (and LLF) for the offset in a simple regression (data: *Record1.csv*)

Clicking on the **gain** button opens the dialog box shown in Figure 5.1.7. It reveals the estimated gain, and offers the option to

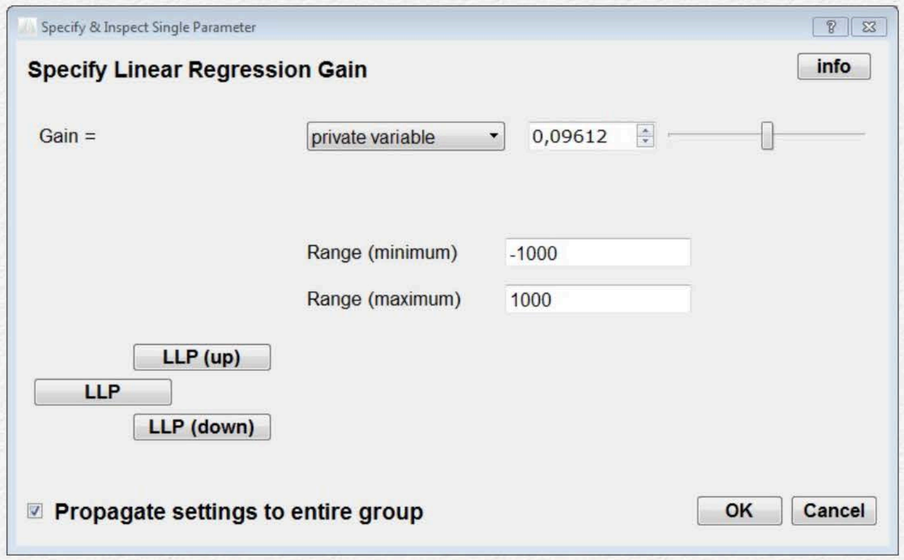

FIGURE 5.1.7 Dialog window for the gain

request the LLP for this coefficient, which produces the graphical result of Figure 5.1.8, from which it can be deduced that

$\beta_1 = 0.096$, with 95% CI = [0.077,0.115].

As zero is outside of this confidence interval, we can conclude that there is evidence for a linear relationship between the independent and the dependent variable.

Clicking on the **prob** button opens a dialog window with all the necessary functionality to examine the remaining parameter of the model, i.e., the standard deviation of the Gaussian distribu-

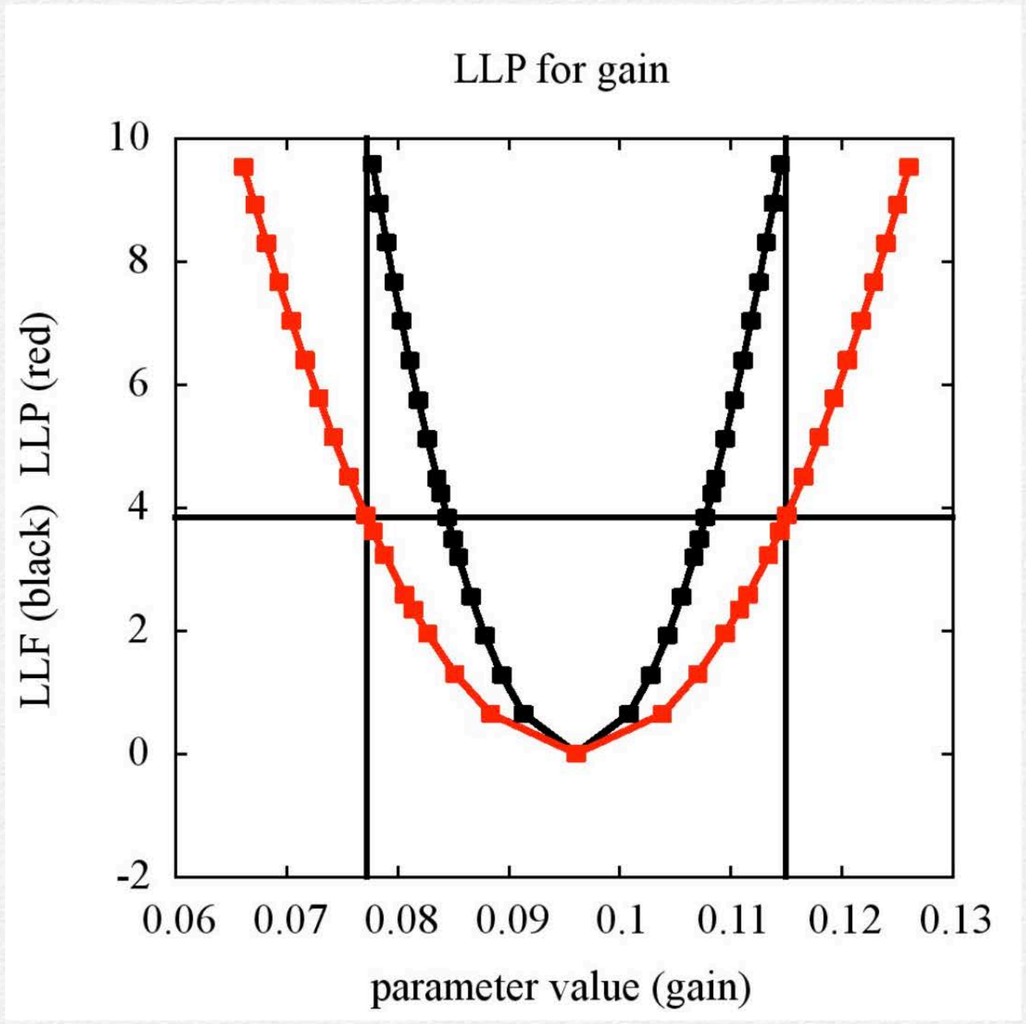

FIGURE 5.1.8 LLP (and LLF) for the gain in a simple regression (data: *Record1.csv*)

tions, and to trigger a LLP calculation which results in Figure 5.1.9. Using the LLP intersection method, we can easily establish that

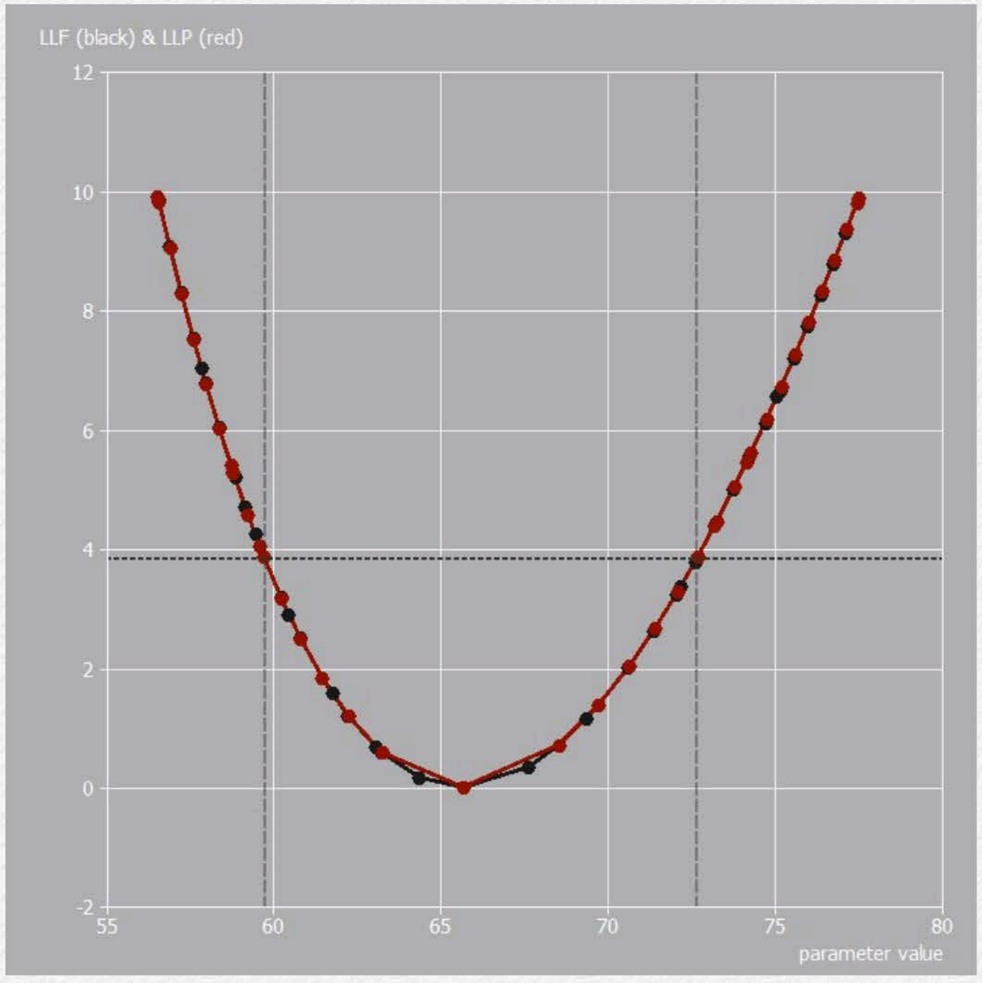

FIGURE 5.1.9 LLP (and LLF) for the standard deviation in a simple regression (data: *Record1.csv*)

$\sigma = 65.66$, with 95% CI = [59.72,72.66].

Traditional Regression Statistics

ILLMO also offers access to more traditional regression statistics by selecting the "*Simple Regression*" option next to the **traditional statistics/analyses button** in the **stats** dialog window (note: readers only interested in log-likelihood modeling can skip this section). The textual output is quite extensive and can be split into several parts.

The first part is the **ANOVA** table in Illustration 5.1.1 where the most relevant parts have been highlighted. The fraction of the variance in the observed data that is explained by the simple regression is $R^2 = 0.335$, which means that 66.5% of the variance in record sales cannot be accounted for by changes in advertisement costs. This relatively low correlation is reflected in the fact that the dots in the scatterplot in Figure 5.1.3 do not cluster very closely to the diagonal line. The **correlation coefficient**

$R = 0.578$, with 95% CI = [0.478,0.664]

is a frequently used measure for the **effect size**. As the CI for the correlation coefficient excludes zero, we can conclude that the regression coefficient is significantly different from zero.

The F-test reported in Illustration 5.1.1 provides the estimated probability that the gain β_1 is equal to zero, which is obviously very small.

```
Source       DOF     SS            MS
-----------------------------------------------------
residue      198     862264.168    4354.86954
model        1       433687.832    433687.83200
  R^2 = SS(model)/SS(corrected)
  R^2 = 0.334648, R = 0.578488
  R confidence interval: [ 0.478173, 0.663903]

  R_a^2 = 1 - (1-R^2)*dof(corrected)/dof(residue)
  R_a^2 = 0.331288, R_a = 0.575576 (adjusted)

  F(1,198) = 99.587 (p=2.94198e-19)
       > 3.887 (p=0.05) (accept model)
-----------------------------------------------------
corrected    199     1295952.000   6512.32161
average      1       7465248.000   7465248.00000
  R_c^2 = SS(average)/SS(total)
  R_c^2 = 0.852081, R_c = 0.923082
  F(1,199) = 1146.327 (p=1.59909e-84)
       > 3.887 (p=0.05) (accept average)
-----------------------------------------------------
total        200     8761200.000   3.88667
```

ILLUSTRATION 5.1.1 ANOVA table for a simple regression (data: *Record1.csv*)

```
Multi-model comparison:
 Model : P = 3, LLC = 2241.38, AIC = 2247.38
 Null  : P = 2, LLC = 2322.86, AIC = 2326.92
 AIC(null) - AIC(model) = 79.5487
 Null  : likelihood = 0.000000, weight = 0.000000
 Model : likelihood = 1.000000, weight = 1.000000
```

ILLUSTRATION 5.1.2 MMC table derived from the preceding ANOVA table (data: *Record1.csv*)

As discussed in earlier chapters, ANOVA tables can be converted into MMC tables, and the outcome in the current example is provided in Illustration 5.1.2. It leads to the conclusion that, in a comparison between the null model, with $\beta_1 = 0$, and the simple regression model, with $\beta_1 \neq 0$, all the weight goes to the latter model.

Traditional statistics can also produce 95% CIs for the regression coefficients, as shown in Illustration 5.1.3. Note that these CIs are almost identical to the ones obtained earlier using the more general LLP intersection method. The fact that the maximum-likelihood estimate $\sigma = 65.66$ is slightly smaller than the root mean square error of $rmse = 65.99$ is a well-known property of maximum-likelihood estimates.

As there is a correlation of -0.785 between the offset and the gain, the joint confidence ellipse for both coefficients is as

```
Regression results for Gaussian model:
 mean square error, mse = 4354.87
 root mean square error, rmse = 65.9914
 (ML estimate for standard deviation = 65.6606)

(offset - ML estimate = 134.14)
 offset - estimate = 134.14
 offset - standard error = 7.53657
 offset - CI(95pct) = [119.281,148.999]

(gain - ML estimate = 0.0961248)
 gain - estimate = 0.0961245
 gain - standard error = 0.00963237
 gain - CI(95pct) = [0.0771337,0.115115]

 (offset,gain) - correlation = -0.78527
 (offset,gain) - 95 pct confidence ellipse
  offset = 134.14 + 18.4468
  * [ 0.327666 * cos(t) + 0.944794 * sin(t) ]
   gain  = 0.0961245 + 0.0235766
  * [ 0.327666 * cos(t) - 0.944794 * sin(t) ]
   ( t = [0,2*pi] )
```

ILLUSTRATION 5.1.3 95% confidence intervals for the regressions coefficients obtained through traditional statistics (data: *Record1.csv*)

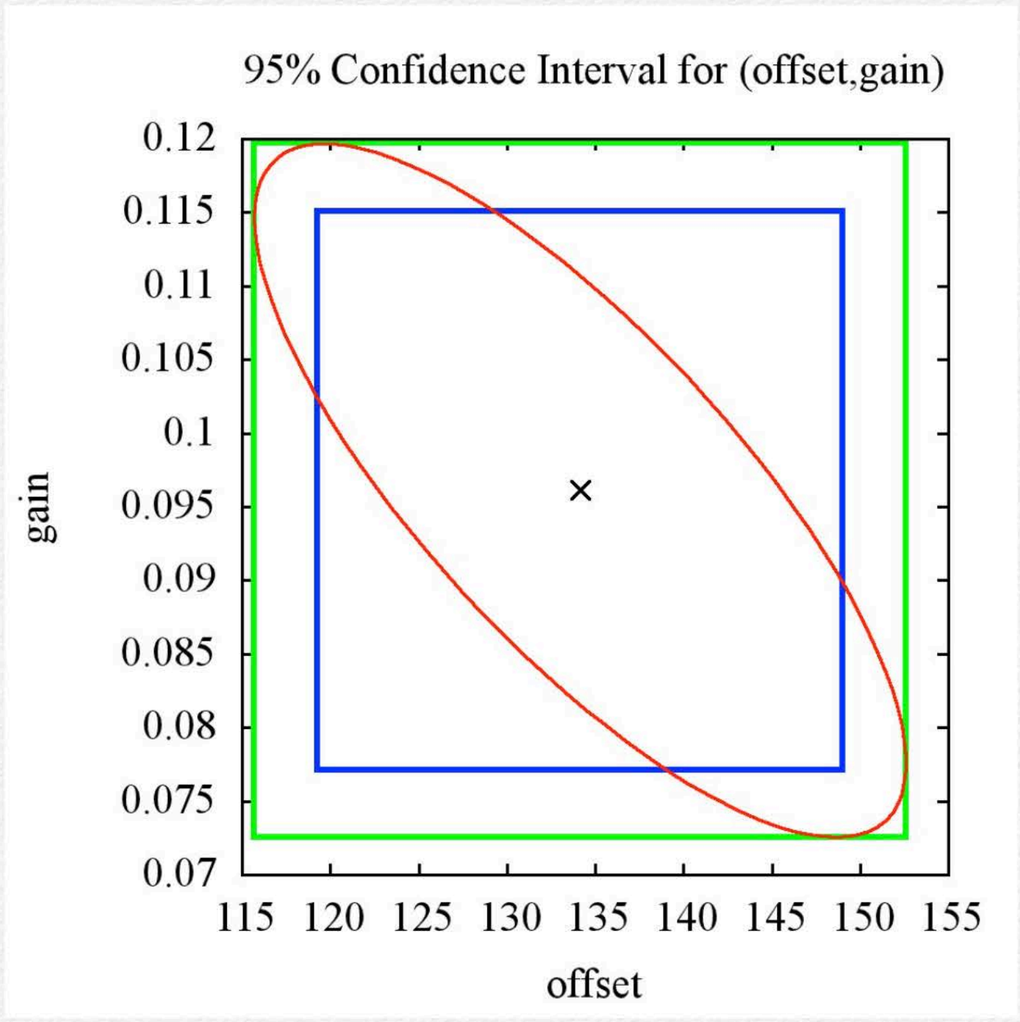

FIGURE 5.1.10 Confidence ellipse for the offset and gain in a simple regression (data: *Record1.csv*)

shown in Figure 5.1.10. This ellipse, with enclosing rectangle in green, expresses that the estimated gain will tend to be larger if the estimated offset is smaller, and vice versa. The blue rectangle indicates the confidence rectangle as derived from the separately estimated 95% CIs for the offset and gain.

The NULL model

Simple regression provides an alternative way to create a NULL model in which the averages are constant across conditions. By opening the **gain** dialog window and changing the setting from "*private variable*" to an "*adjusted value*" equal to zero, the regression coefficient is set to $\beta_1 = 0$ so that the predicted averages are all equal to $\hat{\mu}_k = \beta_0$, for $k = 1,\ldots,K$. Comparing this NULL model with the simple regression model of the previous section for our example data results in an increase in AIC of 79.43 so that the weight of the simple regression model is effectively equal to one, and the weight for the NULL model is equal to zero. Note that this increase in AIC observed using the general log-likelihood method is almost identical to the increase in AIC of 79.55 reported in Illustration 5.1.2 for the Gaussian model, as derived from the ANOVA table in Illustration 5.1.1.

Binary Logistic Regression

Traditional statistics distinguishes **logistic regression** as a separate method for performing confirmatory data analysis in case the dependent variable is discrete. Many descriptions of logistic regression even focus on the case where the dependent

variable is **binary**, i.e., when it can only assume two values (that we can code as 0 and 1).

In order for our treatment to be in agreement with classical logistic regression, we should adopt a logistic distribution instead of the default Gaussian distribution. As the dependent variable is binary, the value of the cumulative distribution

$$P_k(x) = \frac{1}{1 + \exp\left[-\frac{x - \mu_k}{s}\right]}$$

at the boundary value of $b = 0.5$ is the relevant model parameter, as $p_k(0) = P_k(b)$ and $p_k(1) = 1 - P_k(b)$ are the probabilities of observing 0 and 1, respectively, in condition k.

The standard deviation of the logistic distribution is equal to $\sigma = s \cdot \pi/\sqrt{3} = 1.814\, s$. In case the dependent variable is binary, then the scale factor s, and hence also the standard deviation σ, needs to to be fixed to avoid that the value $P_k(b)$ is influenced by more than a single parameter, i.e., next to the predicted average μ_k. Note that fixing the parameter s (to $s = 1$) is not required if the dependent variable is discrete but not binary, i.e., can assume more than 2 distinct values.

The **effect size** in case of a binary variable is often expressed in terms of the odds ratio. The ***odds*** in condition k is the ratio of the probability $p_k(1)$ of observing a one over the probability $p_k(0)$ of observing a zero. In case a logistic distribution is used as model, it can be expressed as

$$Odds_k = \frac{p_k(1)}{p_k(0)} = \exp\left[\frac{\mu_k - b}{s}\right],$$

Obviously, both outcomes (0 and 1) are equally probable and the odds is equal to 1 if the average coincides with the boundary between both categories, i.e., $\mu_k = b$. If $\mu_k > b$, then $p_k(1) > p_k(0)$, while if $\mu_k < b$, then $p_k(1) < p_k(0)$.

The actual effect size is the ***odds ratio*** between two conditions k and m. In case a logistic distribution is adopted, it can be related to the difference in their averages, i.e.,

$$\frac{Odds_k}{Odds_m} = \exp\left[\frac{\mu_k - \mu_m}{s}\right].$$

This latter expression becomes

$$\frac{Odds_k}{Odds_m} = \exp\left[\frac{\beta_1}{s}(v_k - v_m)\right]$$

in case these averages are obtained through simple regression against the independent variable $\{v_k; k = 1,\ldots,K\}$. It implies that the odds ratio is equal to $\exp(\beta_1/s)$ when the independent variable changes by one unit, which is why traditional statistical programs often report $\exp(\beta_1/s)$ next to the gain β_1 itself as a relevant outcome of a logistic regression.

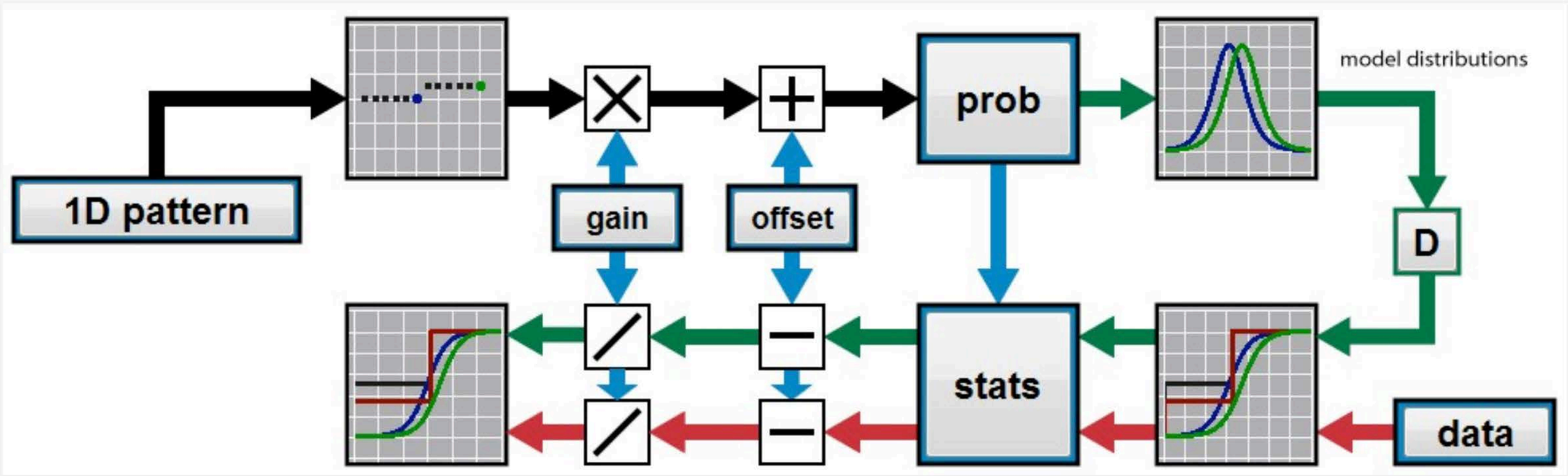

FIGURE 5.1.11 Simple (logistic) regression in case of binary data

We take an example from the chapter on logistic regression in the book(s) by Field as illustration. More specifically, the file *eel.csv* contains 4 columns and 113 entries. The first column entitled "cured" is the dependent variable, and indicates whether (1) or not (0) a patient has recovered from a specific affliction. The second and third columns are the independent variables. Column 2 contains a binary variable entitled "Intervention" that indicates whether (1) or not (0) the subject has undergone treatment, while the third column entitled "Duration" indicates the number of days since the first occurrence of the affliction. The fourth column entitled "Index" contains an index that can be used to distinguish the $K = 12$ different conditions, which correspond to all unique combinations of Intervention and Duration.

The file can be read using the *File* menu option "*Open CSV file (with attributes, field separator:semicolon)*" and should be interpreted as NOT containing data for distinct conditions in separate columns.

Simple regression against “Intervention”

A first option for simple regression is to use “Intervention” as the independent variable in the linear regression. This can be accomplished by using “Index” as the “*column with number of (first) condition*” and “Intervention” as the “*column(s) with independent variables*” in the CSV dialog box. The dependent variable “Cured” should obviously be treated as discrete in case of logistic regression, and the logistic distribution with default value $s = 1$ should be used to create distributions that match the cumulative histograms at the threshold value of $b = 0.5$.. The resulting model is shown in Figure 5.1.11.

The estimated regression coefficients in our example case are the offset

$\beta_0 = 0.173$, with 95% CI = [-0.429,0.770]

and the gain

$\beta_1 = 1.380$, with 95% CI = [0.518,2.255],

from which

$\exp(\beta_1/s) = 3.975$, with 95% CI = [1.679,9.535],

can easily be derived. The treatment hence has a significant effect on the chance to be cured as zero is outside of the CI for the gain β_1, or equivalently, one is outside of the CI for $\exp(\beta_1/s)$.

Simple regression against “Duration”

A second option for simple regression is to use “Duration” as the independent variable in the linear regression. This can be accomplished by again distinguishing 12 stimulus conditions with “Index” as the “*column with number of (first) condition*” and by substituting “Duration” as the “*column(s) with independent variables*” in the CSV dialog box. The dependent variable “Cured” should again be treated as discrete, and the logistic distribution with default value $s = 1$ should be used to create distributions that match the cumulative histograms at the threshold value of $b = 0.5$.The resulting simple regression has a gain

$\beta_1 = 0.144$, with 95% CI = [-0.215,0.510],

so that this gain is not significantly different from zero, and “Duration” is not a useful linear predictor for the chance to be cured.

Simple Nonlinear Regression

In the examples of simple regression up to now we have assumed that one of the independent variables is used to supply the values $\{v_k; k = 1,\ldots,K\}$ that are in turn used to predict the averages in the different conditions. However, if the relationship between the independent and the dependent variable is very nonlinear, this may not be an adequate choice, and considering a slightly more complex model where transformed values

$\{T(v_k); k = 1,\ldots,K\}$ are used instead may be a better option.

We illustrate how ILLMO can be used to perform ***simple nonlinear regression*** on the data in the file *Fitts_Mouse.scal* which was already used in the previous chapter to illustrate the use of data transformations. This file contains 9 rows with measured selection times, each line corresponding to a combination of 3 possible target distances D and 3 possible target widths W. This data set was collected in order to perform a confirmatory analysis of Fitts' law, which states that the average selection time in condition k, corresponding to target distance D_k and width W_k, should be proportional to the independent variable

$$v_k = \log(1 + D_k/W_k),$$

for $k = 1,\ldots,K$ (with $K = 9$ in the example case). The independent variable $R_k = D_k/W_k$ is available in the file *Fitts_Mouse.stim*. Using the *File* option "*Open TEXT file containing a pattern (connect to current attribute)*" allows to perform two functions at the same time. First, creating a 1D pattern $\{R_k, k = 1,\ldots,K\}$ with the specified ratios. Second, to use this pattern as the predictor in a simple regression of the measured selection times.

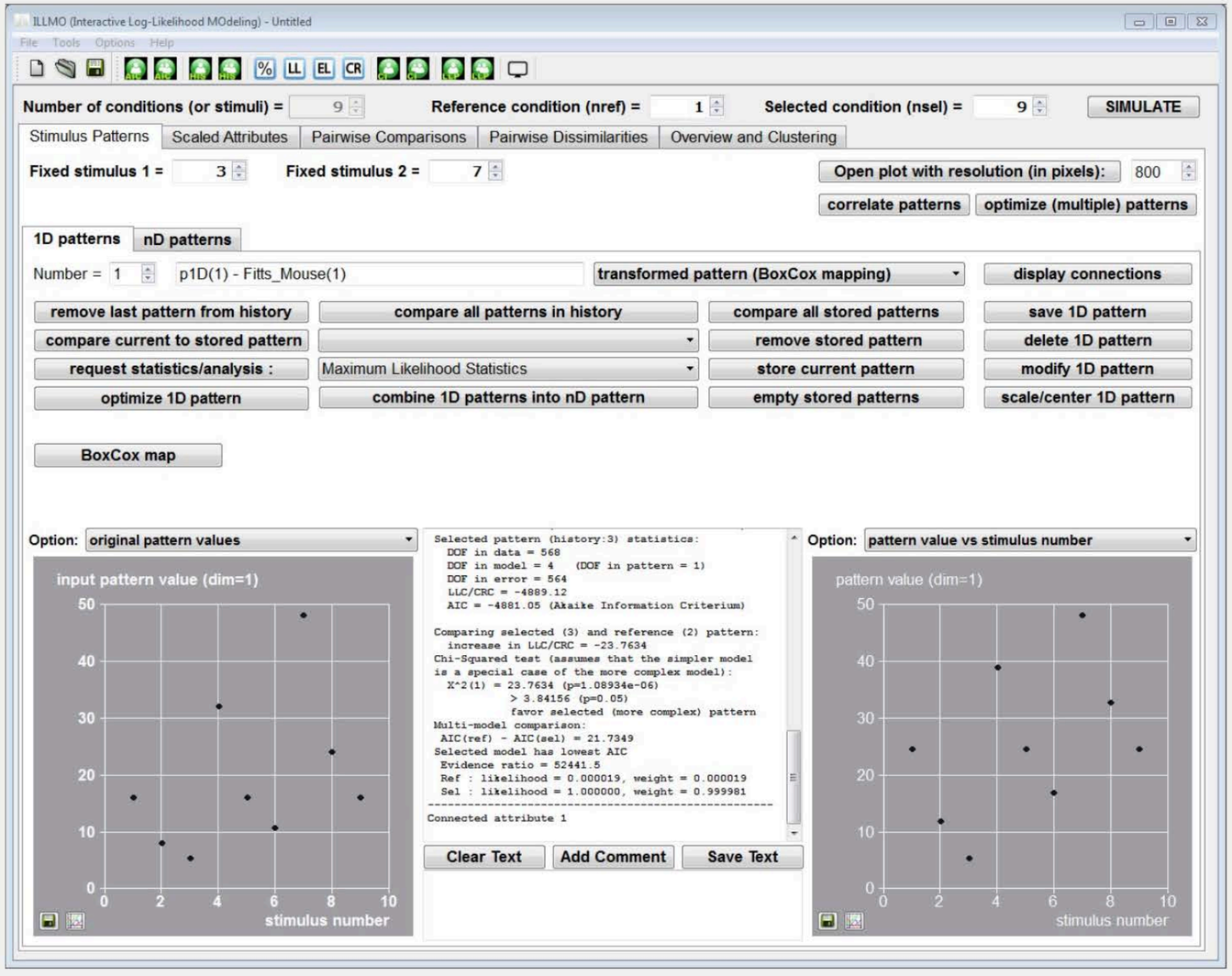

FIGURE 5.1.12 Input pattern used in the nonlinear simple regression of *Fitts_Mouse.scal*

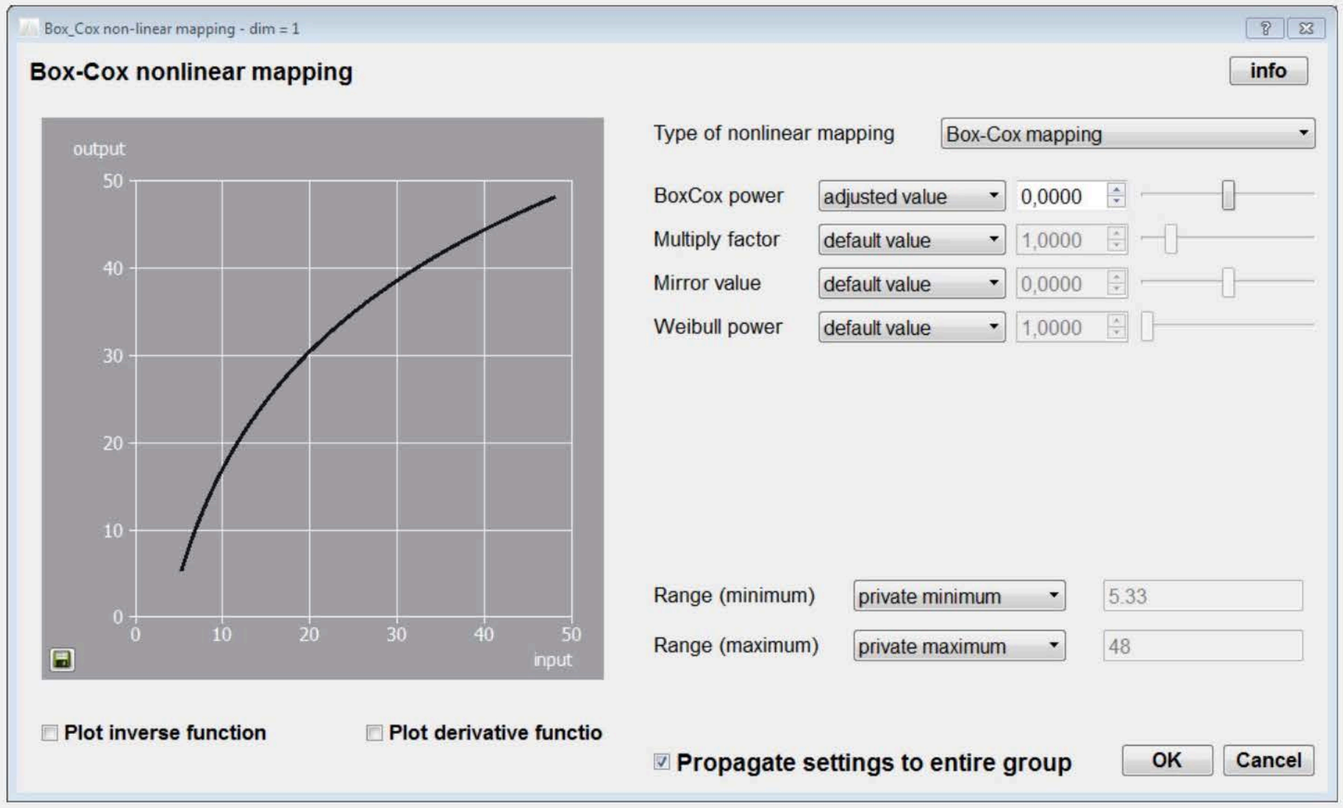

FIGURE 5.1.13 Dialog box for controlling the Box-Cox nonlinear mapping in a simple nonlinear regression

Moving to the Stimulus Patterns page of the ILLMO interface and selecting the 1D patterns subpage will show the pattern that has been loaded. Changing from "*fixed pattern (linear regression)*" to "*transformed pattern (BoxCox mapping)*" creates a **BoxCox map** button, as seen in Figure 5.1.12. Pressing this button opens de dialog window in Figure 5.1.13 that has already been adjusted to transform the ratios R_k into a pattern that agrees with Fitts' law (which is accomplished by adjusting the default value of the "*BoxCox power*" from one to zero). This decreases the LLC by 23.05, which is indeed a very substantial reduction.

The BoxCox power can however also be converted into a "*private variable*" and optimized. The optimal value of

$$q = 0.136, \text{ with 95\% CI = [-0.180,0.459],}$$

is not significantly different from zero, and hence doesn't constitute a substantial improvement over Fitts' law, as can also be verified from the corresponding MMC which assigns a weight of 0.658 to Fitts' model, compared to a weight of 0.342 for the alternative model with a variable power q.

Note that ILLMO supports different types of monotonically increasing transformations that can be used to transform independent variables, i.e., **Box-Cox power laws** and monotonically increasing (quadratic) **spline functions**, as well as one additional type that will be discussed in the next section.

Factor mappings

In the above discussion on simple regression it is either assumed that the independent variable has interval properties, i.e., can be used directly within the regression equation, or has ordinal properties, i.e., needs a monotonically increasing transformation before being entered into the regression equation. In the most general case, an independent variable however only has nominal properties, i.e., it distinguishes experimental conditions, but doesn't have any numerical significance. We use an example from the chapter on repeated-measures design in the book(s) by Field to illustrate how simple regression can be handled in such a case.

The file *Bushtucker.dat* contains the times it took eight participants to retch after eating the foods specified in columns 1 to 4. These four columns correspond to the four experimental conditions (1:stick_insect, 2:kangaroo_testicle, 3:fish_eye and 4:witchetty_grub). As this is a file with observations for different conditions in separate columns, it will be entered in ILLMO using the "*Open CSV file (with comparisons, field separator:tab)*" option from the *File* menu. In the dialog window for loading CSV files with measurements for different conditions specified in separate columns, columns 1 through 4 are selected as the column(s) with data values, and the option "*Pairs for all possible reference conditions*" is selected next to *Generate Pairs*.

The result is four Pairwise Comparisons, one for each of the possible reference conditions.

Each of the pairwise comparisons can be analyzed using a range of methods that we have already discussed before. For instance, for the first Pairwise Comparison, with reference condition 1, a one-way ANOVA produces the outcomes in Illustration 5.1.4, implying that there is no significant variation in the differences of the averages in conditions 2, 3 and 4 with reference condition 1, but that this (approximately constant) difference is significantly different from zero. This is confirmed by the corresponding MMC which assigns the majority of the weight (0.895) to a model where these differences are constant but distinct from zero.

We can alternatively use the LLP intersection method to establish the confidence intervals for the differences in averages between the reference condition and the other conditions. The resulting graph in Figure 5.1.14 is in agreement with the ANOVA analysis.

The two variations of Levene's test for constant variance, which are part of the outcome of the ANOVA analysis, produce slightly conflicting outcomes. The F-test that is based on absolute differences from the mean rejects, with a probability of p=0.034, the hypothesis that the variance is constant, while the F-test that is based on absolute differences from the median accepts, with a probability of p=0.079, the same hypothesis. Assuming con-

```
-------------------------------------------------------
Source        DOF    SS           MS
-------------------------------------------------------
 SS(within)    21    246.750      11.75000
 SS(between)    2     13.083       6.54167
   F(2,21) = 0.557 (p=0.581305)
         < 3.467 (p=0.05) (constant)
-------------------------------------------------------
 SS(corrected) 23    259.833      11.29710
 SS(average)    1    280.167     280.16667
   Average = -3.41667
   Rc^2 = SS(average)/SS(total)
   Rc^2 = 0.518827, Rc = 0.720297

   F(1,23) = 24.800 (p=4.90211e-05)
         > 4.281 (p=0.05) (accept average)
-------------------------------------------------------
 SS(total)     24    540.000      22.50000
-------------------------------------------------------
 Multi-model comparison:
  Variable : DF = 4, LLC = 24.2236, AIC = 34.3288
  Constant : DF = 2, LLC = 25.4635, AIC = 30.035
    Zero   : DF = 1, LLC = 43.0202, AIC = 45.202
  Variable : likelihood = 0.116841, weight = 0.104570
  Constant : likelihood = 1.000000, weight = 0.894975
```

ILLUSTRATION 5.1.4 One-way ANOVA for pairwise comparison data (data: *bushtucker.dat*)

stant variance, as is done in the ANOVA in Illustration 5.1.4, is therefore reasonable, although not perfect.

Empirical likelihood estimates that do not rely on the assumption of Gaussian distributions with constant variance may therefore be more reliable in this case. The outcome of the EL estimation is shown in Figure 5.1.15 and can be compared to the result in Figure 5.1.14. One can observe that the estimated confidence interval in condition 4 is indeed substantially larger than the estimated confidence intervals in conditions 2 and 3.

The previous analysis can only be performed for one comparison at a time. We now show how the analysis of all four comparisons (one for each reference condition) can be performed simultaneously as a simple nonlinear regression, in combination with factor mapping. We start by creating a 1D pattern, for instance equal to (1,2,3,4). One way of accomplishing this is to click on the **enter 1D pattern** button on the 1D patterns subpage of the Stimulus Patterns page, which opens a dialog box with a spreadsheet into which the pattern values can be entered. As the pattern should not be interpreted as interval, we select the "*transformed pattern (factor mapping)*" as option on the 1D patterns page. Clicking on the **Factor map** button reveals the dialog window in Figure 5.1.16. In a factor mapping, the two extreme values are fixed (to 1 and 4 in our case), while all other values (in the example case, only the middle values of

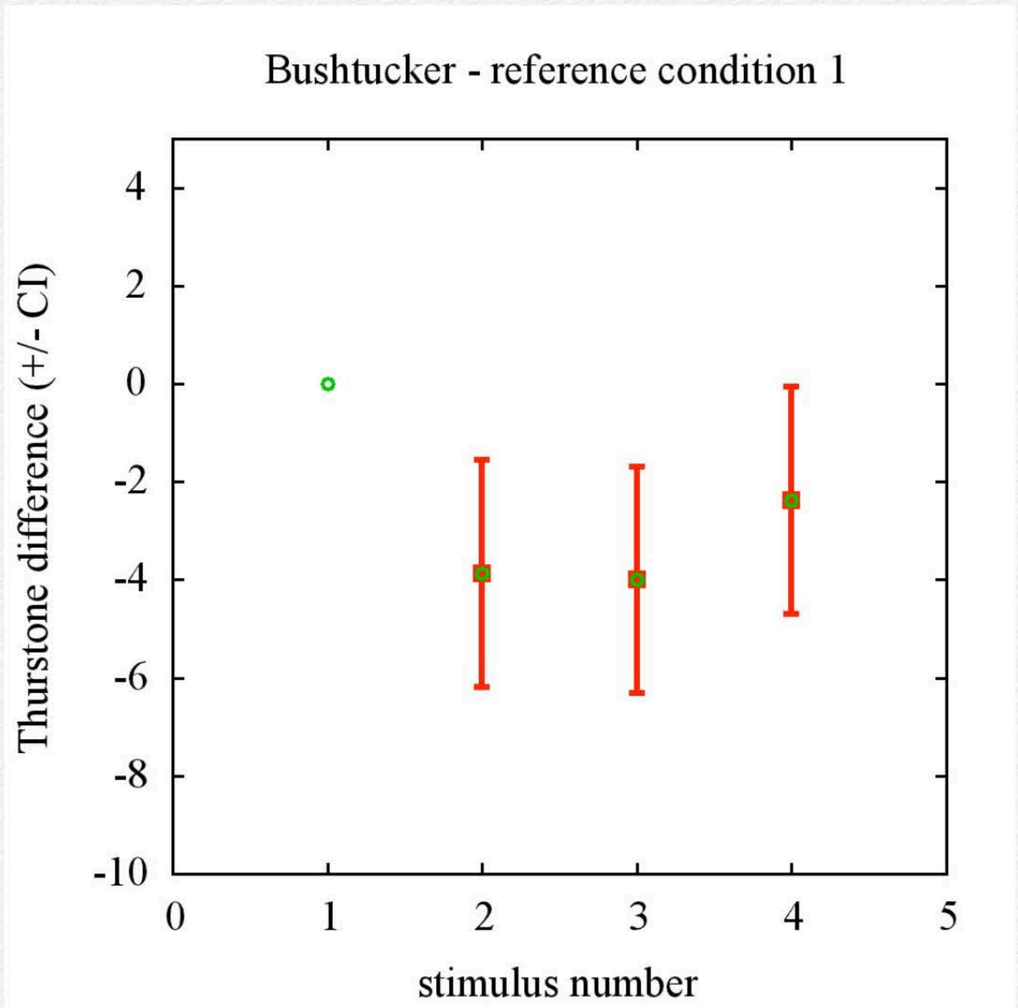

FIGURE 5.1.14 Differences in averages with LLP confidence intervals (data: *bushtucker.dat*)

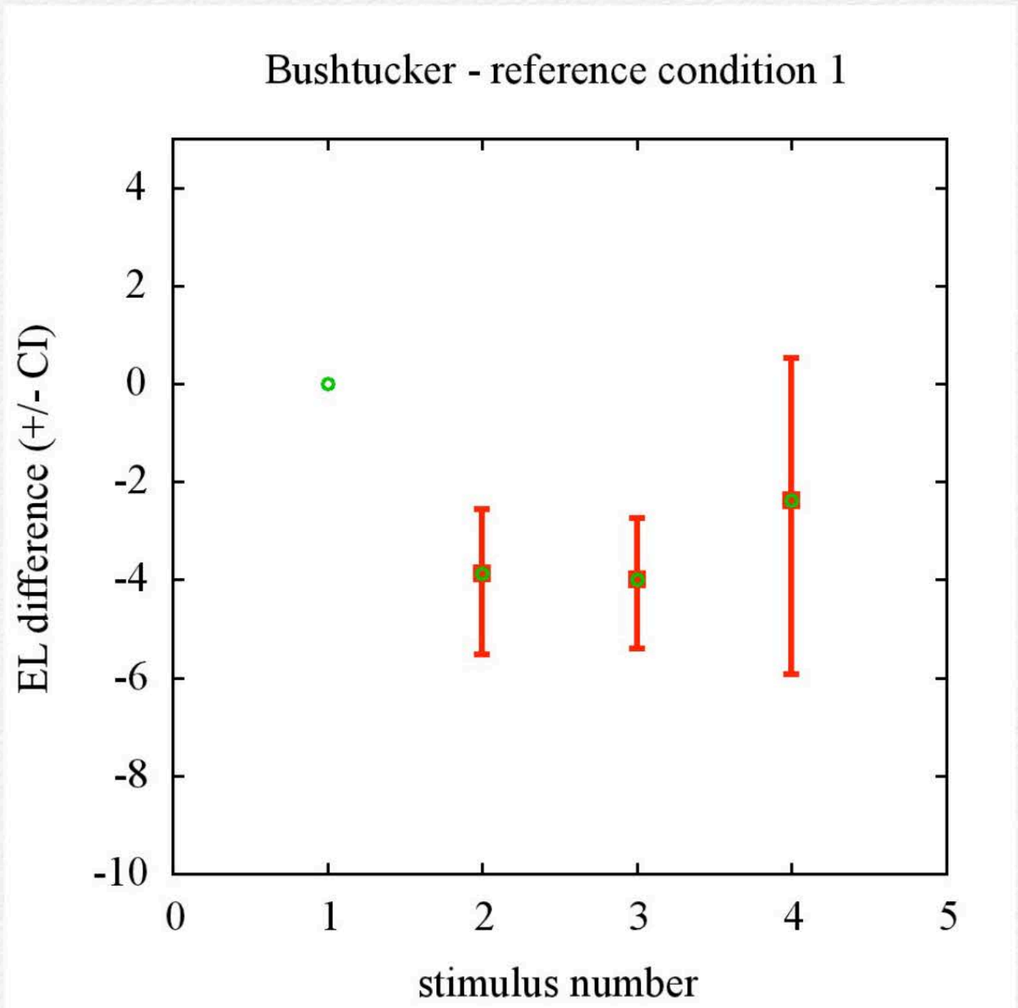

FIGURE 5.1.15 Differences in averages with EL confidence intervals (data: *bushtucker.dat*)

2 and 3) can be assigned any value, implying that these transformed values become parameters in the statistical model.

In order to use this pattern as predictor for **all** Pairwise Comparisons, we need to open another dialog box by clicking on the **private pattern** button on the Model(Confirmatory) subpage of the

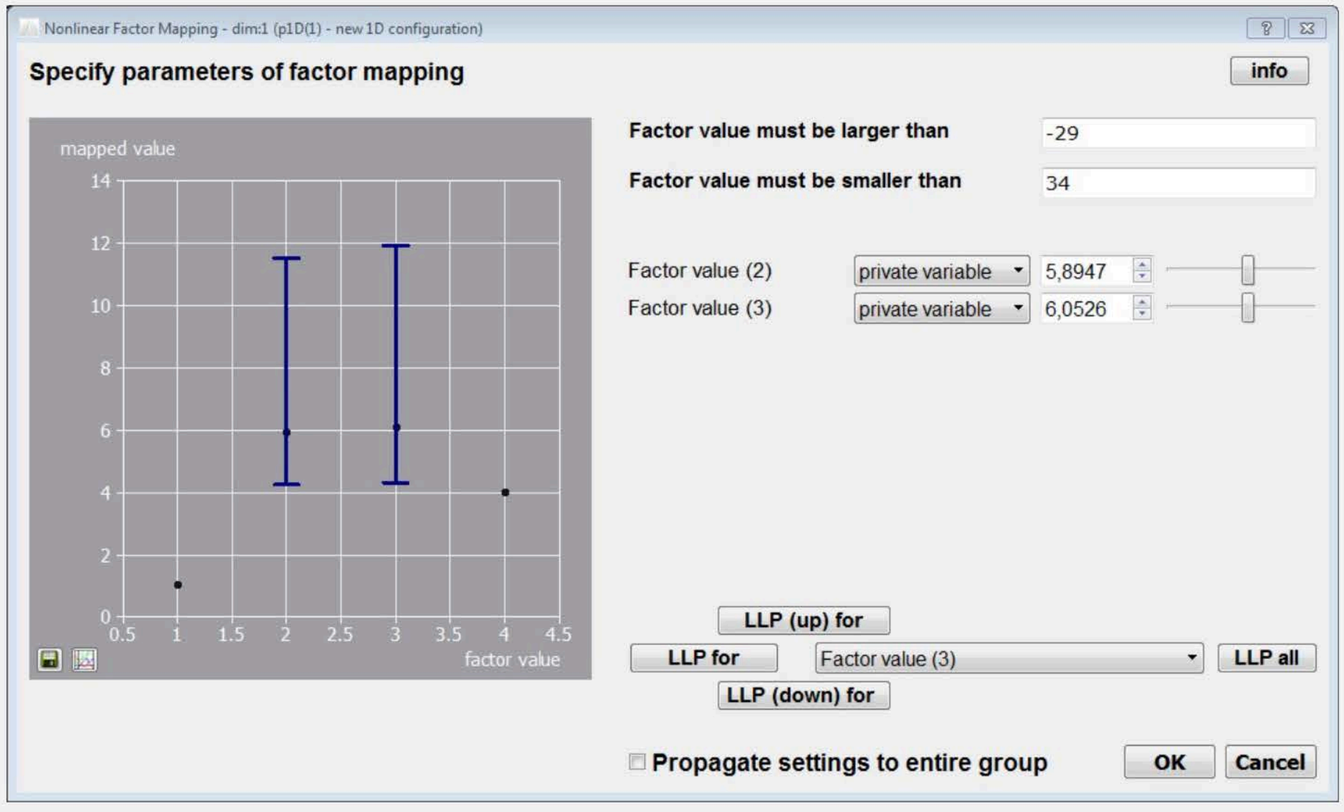

FIGURE 5.1.16 Dialog window for specifying a factor map on a 1D pattern used in simple regression (data: *bushtucker.dat*)

Pairwise Comparisons page. We select *"connect to 1D pattern"* for the pattern that we have just created and select the check box in front of *"Propagate settings to entire group"* in order to accomplish the same change for all four Pairwise Comparisons. After clicking the OK button, the name of the button has changed from **private pattern** to **1D pattern** in the main ILLMO interface to reflect the new setting. In order to verify that all settings have been accomplished successfully, one can check the table on the Overview and Clustering page in the ILLMO interface, see Figure 5.1.17. Note that all four comparisons do indeed share the same "Pattern Name".

Like for any other parameter, LLPs can be constructed for the factor values, from which a confidence interval can be deduced. In the example case the optimal transformed values are

$T(2) = 5.895$, with 95% CI=[4.22,11.49],

and

$T(3) = 6.053$, with 95% CI=[4.29,11.90],

These CIs are rendered on the left in Figure 5.1.15. We can conclude from these CIs that the time to retch is significantly different in conditions 2 and 3 than in condition 4 in this case where all four pairwise comparisons are analyzed simultaneously.

The linear regression coefficient in case of the first pairwise comparisons (where condition 1 is the reference) is

$\beta_1 = -0.79$, with 95% CI = [-1.09,-0.49].

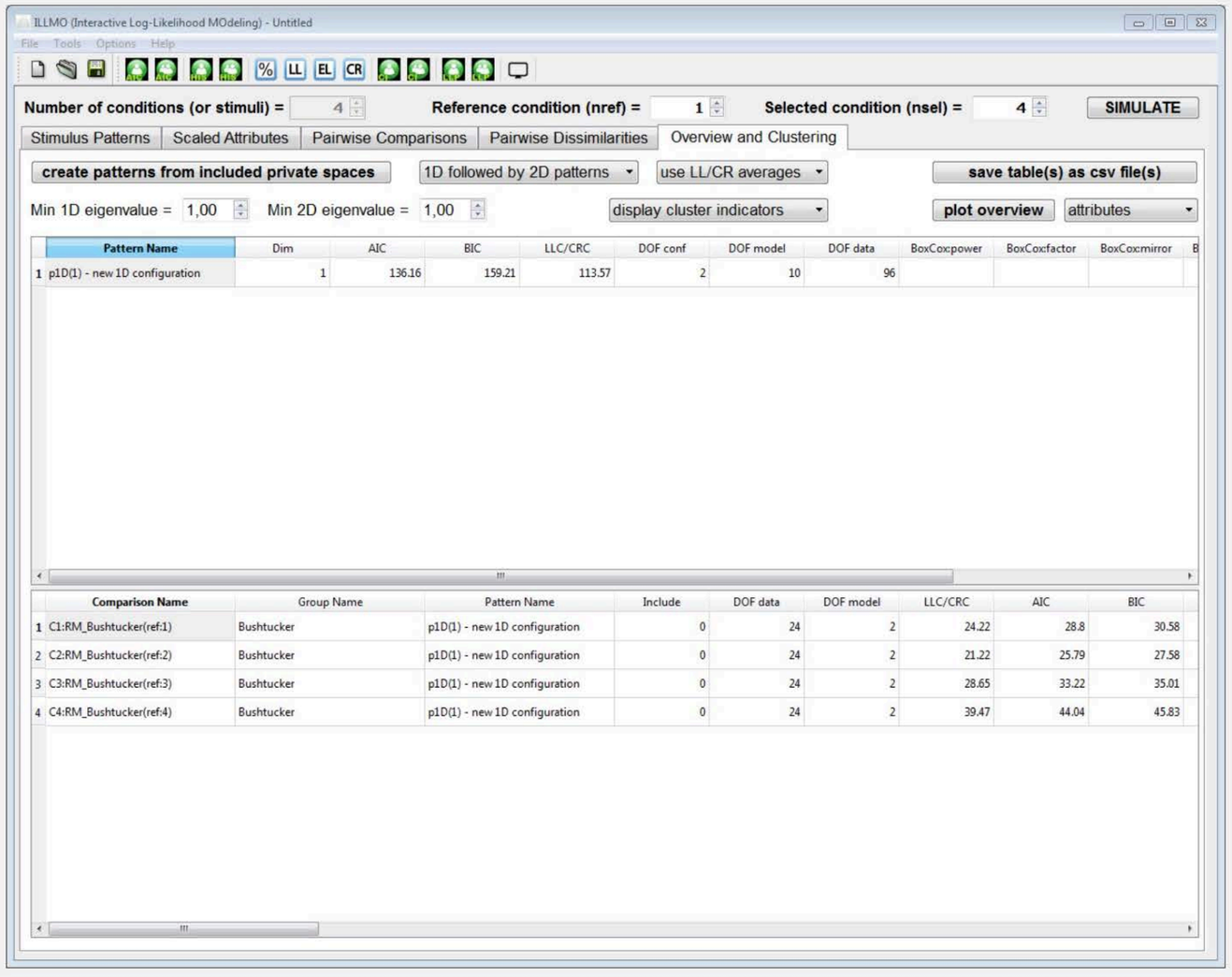

FIGURE 5.1.17 Overview page, showing parameters of the available pattern and the four pairwise comparisons (data: *bushtucker.dat*)

This value is significantly different from zero, confirming that the pattern in Figure 5.1.15 contributes significantly to the prediction of the dependent variable.

Generalized Linear Models

The method of generalized linear models (GLM) is an extension of linear regression that has been primarily introduced in traditional statistics to extend regression to distributions that are not Gaussian. As ILLMO already supports most popular parametric distributions, as documented in more detail in chapter 7, this argument is not very relevant in the case of ILLMO. GLM however also provides an interesting alternative approach towards modeling distributions with a variance that depends on the experimental condition (next to transforming the input data, as discussed in chapter 4, section 3, on data transformations). Not only could this approach be included quite easily into ILLMO, but an extension of the existing GLM method could also be accomplished in a rather straightforward way.

In order to accomplish the required extension of linear regression towards non-Gaussian distributions, GLM includes a non-linear function in between the linear predictions from independent variables and the averages of the parametric distributions that are used to model the observed histograms. This

means that the actual averages $\{\mu_k; k = 1,\ldots,K\}$ in the K experimental conditions are approximated by model averages

$$\left\{\hat{\mu}_k = M\left(x_k\right) = M\left(\beta_0 + \sum_{l=1}^{L} \beta_l \cdot v_{kl}\right); k = 1,\ldots,K\right\}$$

The nonlinear function $\hat{\mu} = M(x)$ is called the ***mean function***.
The inverse function $x = M^{-1}(\hat{\mu})$ is called the ***link function***.
The simple regression introduced at the start of this chapter corresponds to a mean function $M(x) = x$ that equals the identity mapping.

Distributions that belongs to the very general class of distributions called the overdispersed exponential family have a so-called ***canonical mean/link function***. This mean function is the identity function in case of a Gaussian distribution, but is a nonlinear function in case of other distributions. For example, in case of a Poisson distribution, the canonical mean function is an exponential function, i.e., $\hat{\mu} = \exp(x)$. In case of a gamma distribution, it is a power function with a power of -1. In cases where the average is a probability in the range [0,1], cumulative distribution functions, such as the logistic distribution defined in chapter 7, are used as mean functions. They map the linear predictions from the independent variables, i.e., $\{x_k, k = 1,\ldots,K\}$, which are usually defined for arbitrary real values, to the required range of [0,1].

An interesting insight from GLM is that, for the overdispersed exponential family of distributions, the variances $\{\sigma_k^2; k = 1,\ldots,K\}$ in the K experimental conditions are proportional to the derivatives of the mean function, i.e., they can be modeled by

$$\left\{\hat{\sigma}_k^2 = \sigma^2 \cdot M'\left(x_k\right) = \sigma^2 \cdot M'\left(\beta_0 + \sum_{l=1}^{L} \beta_l \cdot v_{kl}\right); k = 1,\ldots,K\right\},$$

where σ controls the overall level of the standard deviations. GLM hence points out that there may exist a relationship, expressed through the mean function and its derivative, between the averages in different conditions and the variances (or standard deviations) in these conditions. For example, in case of the exponential mean function, where $M(x) = M'(x) = \exp(x)$, we would obtain that the modeled variance $\hat{\sigma}_k^2$ is proportional to the modeled mean $\hat{\mu}_k$, for all conditions $k = 1,\ldots,K$.

ILLMO adopts and extends the use of non-linear mean/link functions. It not only supports the fixed link/mean functions that are typically adopted in traditional GLM, but also allows the inclusion of nonlinear **Box-Cox** or **spline** functions with parameters that can be optimized to improve the fit between observed histograms and modeled distributions. The default choice in ILLMO is to include Box-Cox transformations, but by checking the option "*use spline instead of Box-Cox mapping on averages*" in the Options menu, spline functions can be used instead. Some

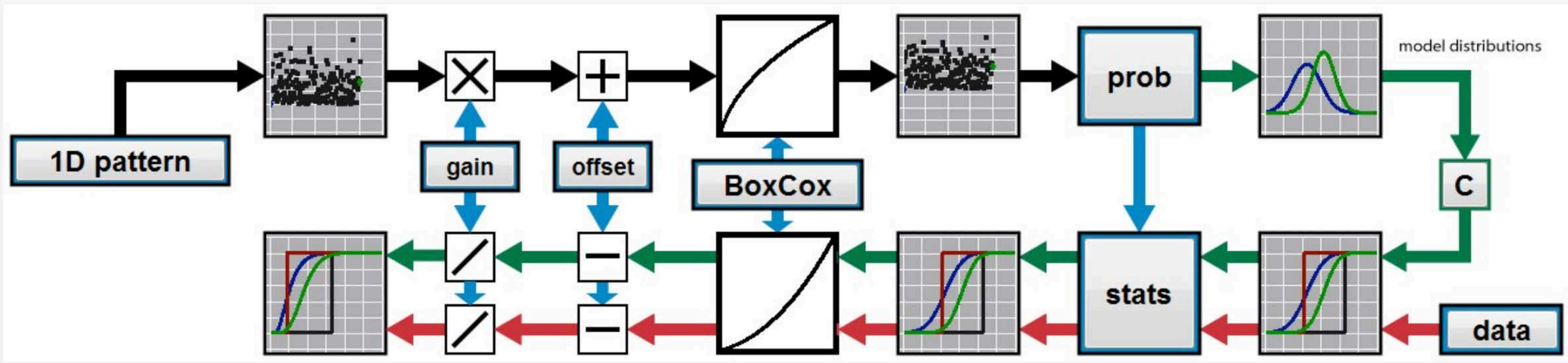

FIGURE 5.1.18 Generalized linear modeling (GLM) through inclusion of a non-linear Box-Cox function after the linear prediction from the independent variable; the Box-Cox function not only determines the averages, but also the standard deviations, of the model distributions.

fixed functions proposed by GLM, such as the exponential function, can also be selected. The canonical functions that are special cases of Box-Cox transformations, such as the power transformation with power equal to -1, are however not included separately.

We illustrate GLM using the example data in *Record1.csv.* The scatterplot in Figure 5.1.4 already gives the impression that the variance in the residue values decreases with increasing values of the independent variable, an effect which may be modeled with the help of GLM.

Changing the simple regression model in Figure 5.1.2 into the GLM in Figure 5.1.18 requires two separate actions. First, a BoxCox transformation with a variable power q (and a fixed factor $f = 1$) is included, creating a model in which only the averages are non-linearly related to the linear prediction. Second, GLM is activated by selecting the option *"Generalized linear model (derived from model transformation)"* next to the *"Options for standard deviation"* in the **prob** dialog box.

In ILLMO, the values of the square roots of the derivatives to the mean function across the different conditions are averaged to one. This accomplishes that the standard deviation in the

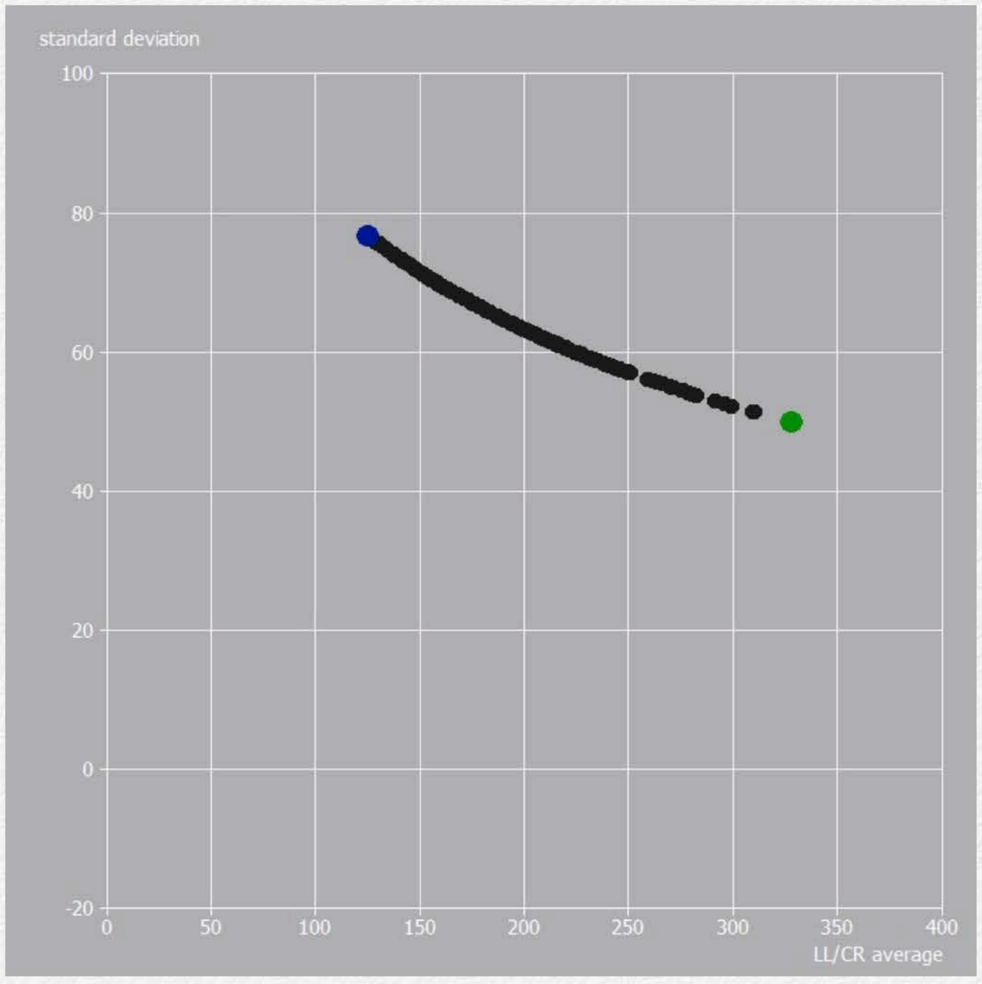

FIGURE 5.1.19 Standard deviation as a function of the average in the GLM (data: *Records1.csv*)

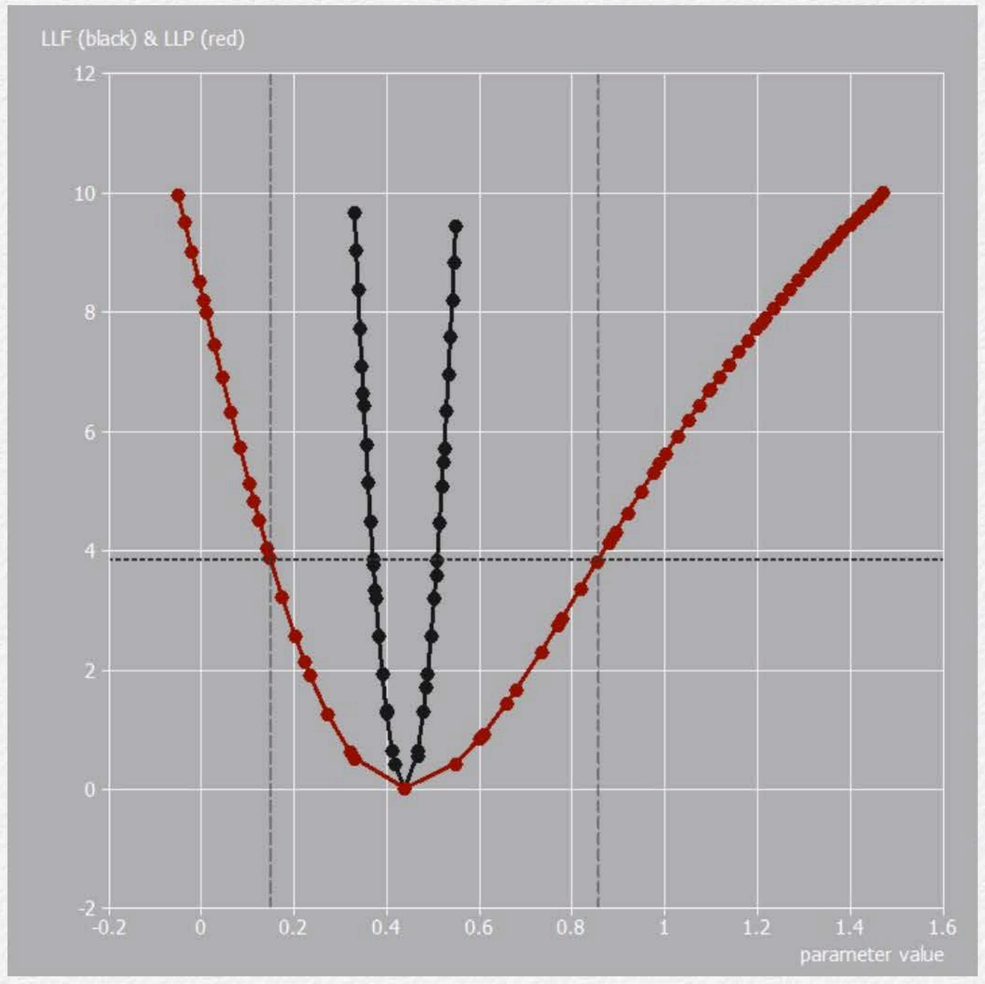

FIGURE 5.1.20 LLP for the power of the Box-Cox function in the GLM (data: *Records1.csv*)

prob dialog box still controls the overall level, while the derivatives to the mean function only control the variation in the standard deviation across conditions.

A MMC between all three models, shown in Figure 5.1.21 shows that, while a non-linear mapping for the predicted averages by itself does not result in an improved model, using such a non-linear mapping to model both averages and standard de-

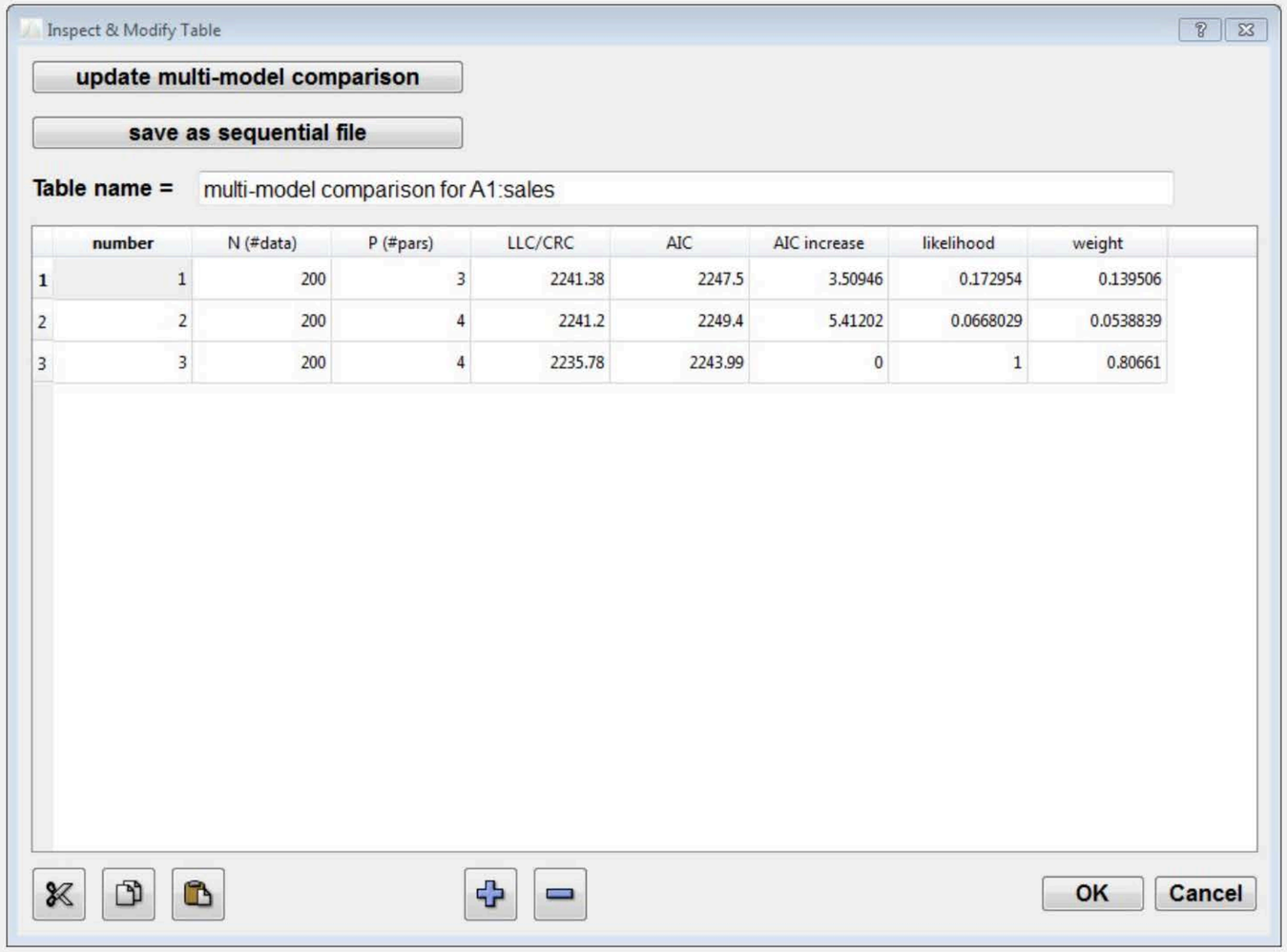

FIGURE 5.1.21 MMC between 3 regression models: 1) simple regression, 2) including Box-Cox mapping for model averages, and 3) GLM where both averages and standard deviations are determined by the Box-Cox mapping (data: *Records1.csv*)

viations of the distributions does provide the best model. More specifically, the likelihood ratio is 5.57 in favor of the GLM over the simple regression. The resulting variation in the standard deviation is plotted in Figure 5.1.19, using plot option: *"LL/CR: average versus standard deviation"*.

Like any model parameter, the LLP of the power in the Box-Cox transformation can be determined, see Figure 5.1.20, to establish the estimate

$q = 0.439$, with 95% CI=[0.149,0.857],

which allows to conclude that both a linear and a logarithmic relationship can be excluded with a confidence of 95%. The factor in the Box-Cox transformation is equal to its default value of $f = 1$.

Section 2

Multiple Regression

Regression With Multiple Independent Variables

1. **Multiple regression is a fairly straightforward extension of simple regression**
2. **ILLMO offers access to most of the traditional multiple regression statistics, but promotes the more general LLP intersection method**
3. **ILLMO implements extensions to traditional multiple regression statistics in the form of multi-model comparisons and model averaging**
4. **New methods for handling confounding variables, related to existing methods such as ANCOVA and multi-level modeling, are proposed**

Multiple Regression

As already announced in the previous section, an obvious extension of simple regression is multiple regression, where the averages of the dependent variable(s) in the distinct conditions are related, in a linear way, to several independent variables, instead of just one. In order to make this situation concrete, we again resort to an example form the book(s) by Field.

The file *Record2.csv* is an extension of the file *Record1.csv that was* used as an example in case of simple regression, as it contains four columns with 200 observations and the first two columns in both files are identical. The first column entitled “adverts” contains the advertising budget (in thousand of pounds) for a record, which will be considered as an independent variable, while the second column entitled “sales” contains the record sales (in thousands) as dependent variable. Columns 3 and 4 are two additional independent variables named “airplay” (the number of times the record is played on the radio in the week preceding release) and “attract” (a subjective score of the attractiveness of the band members).

The file can be read using the *File* menu option “*Open CSV file (with attributes, field separator:semicolon)*” and should be interpreted as NOT containing measurements for different conditions in separate columns. In the CSV dialog box, either one, two or three independent variables can be selected as predic-

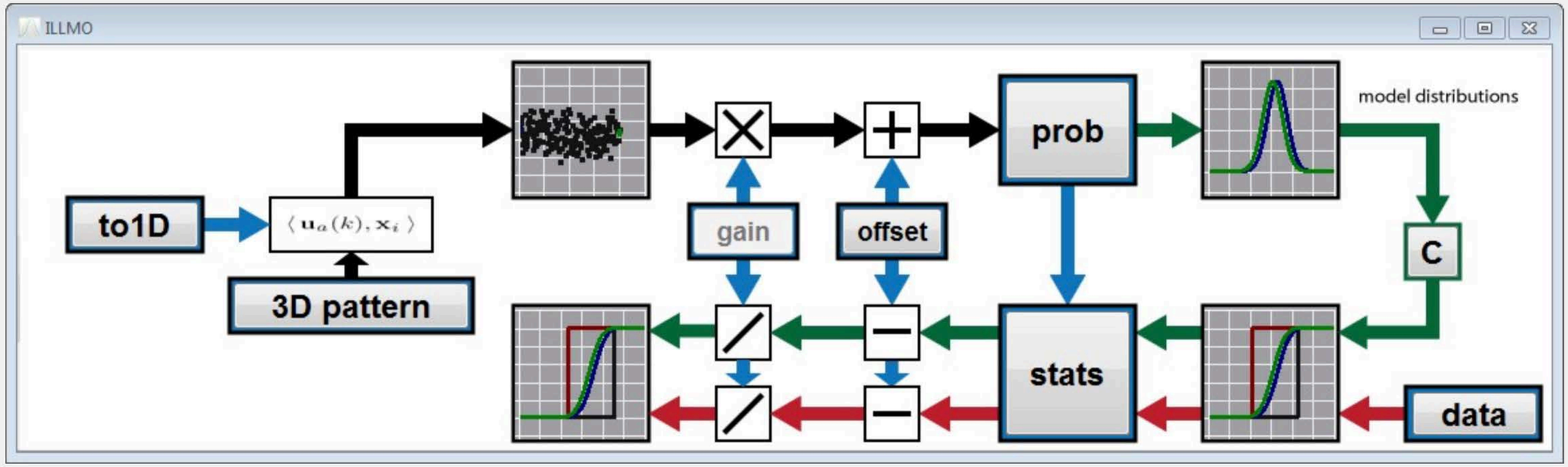

FIGURE 5.2.1 Multiple regression model as visualized by ILLMO (data: *Record2.dat*)

tors in the linear regression. In the current example we select all three independent variables. Upon exit of the CSV dialog, these three independent variables are stored as both a single 3D pattern and three separate 1D patterns, as can easily be verified by inspecting both the 1D patterns and the nD patterns subpages in the Stimulus Patterns page of the ILLMO interface.

The complete regression model, which can be exported from ILLMO by pressing on the "camera" button on the right in the toolbar, or by selecting the corresponding entry in the pop-up menu that appears when right-clicking in ILLMO, is shown in Figure 5.2.1. The **3D pattern** operator provides the independent variables "adverts" , "airplay" and "attract", so $L = 3$ is the number of predictors. The independent variables for condition k, i.e., $\{v_{kl}, l = 1,\ldots,L\}$, are compressed into a single number

$$v_k = \sum_{l=1}^{L} \beta_l \cdot v_{kl}$$

under the control of the operator **to1D** which provides the regression coefficients $\beta_l, l = 1,\ldots,L$**.** The small graph in the upper left shows the resulting 1D pattern $\{v_k; k = 1,\ldots,K\}$, while the **off-**

set button allows to inspect the regression coefficient that was previously termed β_0. The predicted averages

$$\hat{\mu}_k = \beta_0 + \sum_{l=1}^{L} \beta_l \cdot v_{kl},$$

for $k = 1,\ldots,K$, are used by the **prob** operator to generate Gaussian distributions with a constant standard deviation σ that is an additional parameter of the model.

Note that the **gain** button has been deactivated in this case of multiple regression, as the regression coefficients can be inspected and modified using the dialog window in Figure 5.2.2 that appears in response to clicking the **to1D** button. This dialog box offers several alternative ways for mapping from multiple dimensions to a single dimension (click **info** in case you want to get a complete overview of the available options), but the default option entitled *"orthogonal projection on vector (multiple regression)"* corresponds to multiple regression. Other options will be discussed in the next chapter.

It is worthwhile to note that the pattern $\{v_k; k = 1,\ldots,K\}$ that is displayed in the graph in Figure 5.2.2 can be saved as a new 1D pattern by ticking the checkbox in front of "*Store as 1D pattern*" before pressing the **OK** button.

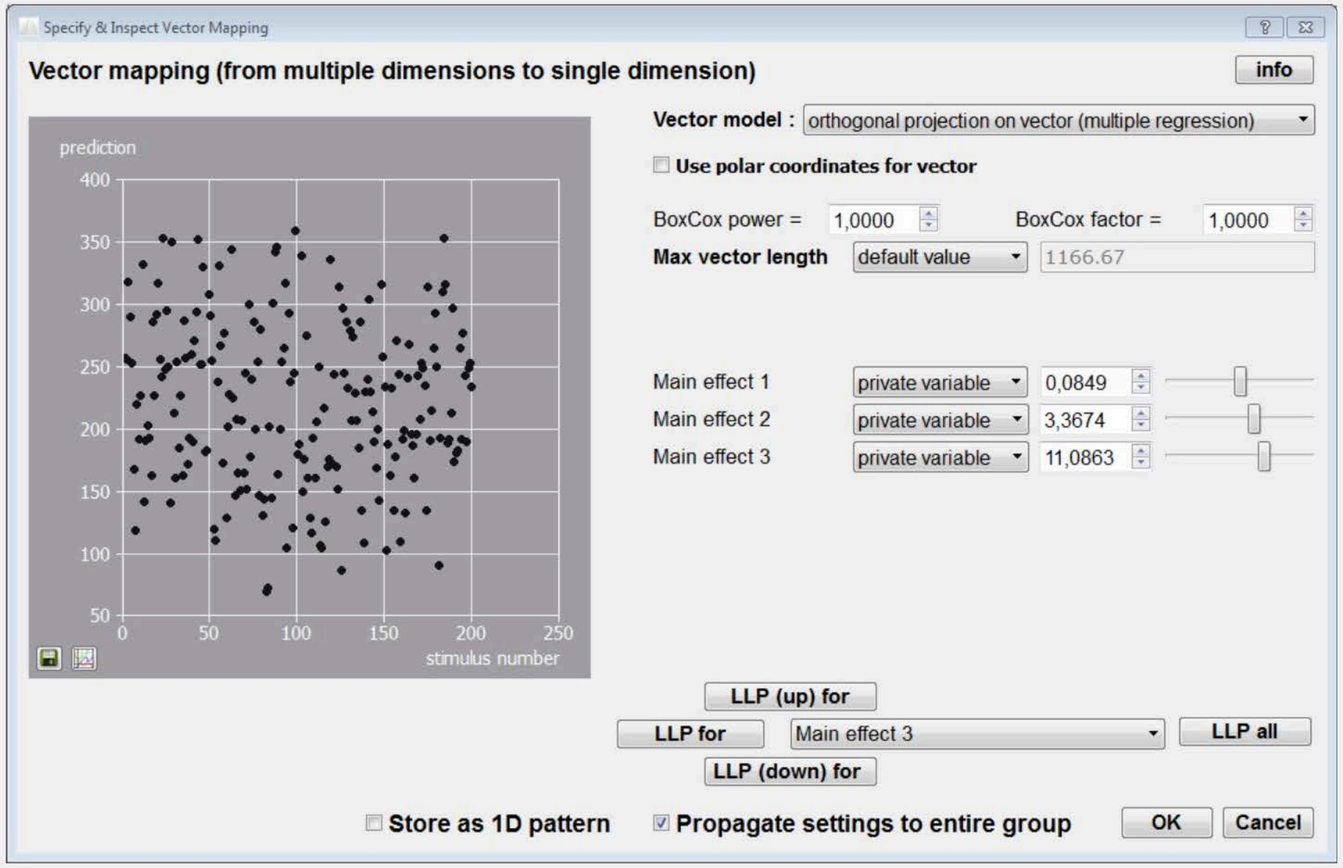

FIGURE 5.2.2 Dialog box **to1D** for controlling multiple regression coefficients (data: *Record2.csv*)

The dialog box in Figure 5.2.2 displays the values of the three separate regression coefficients, and offers the option to calculate LLPs for each of these coefficients. The resulting values of

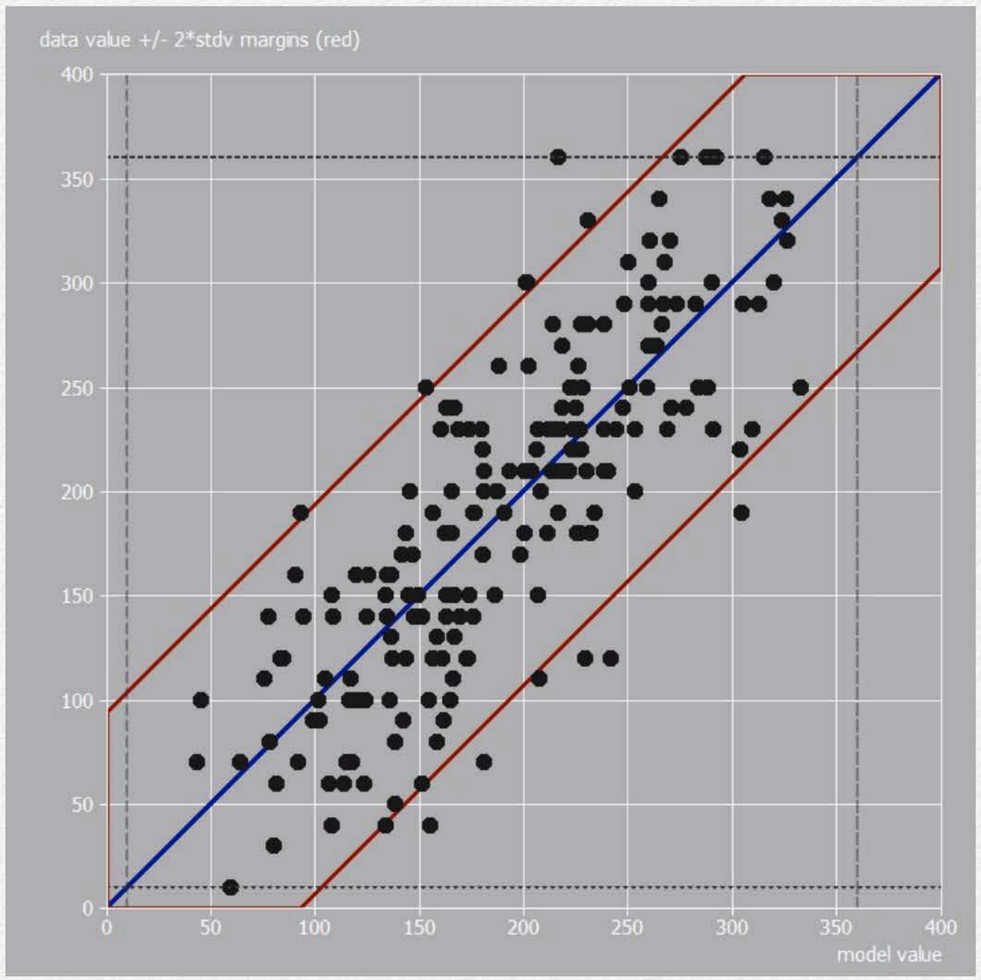

FIGURE 5.2.3 Scatterplot of data value (sales) versus the model value (multiple regression on independent variables) (data: *Record2.csv*)

$\beta_1 = 0.085$, with 95% CI = [0.071,0.098],

$\beta_2 = 3.367$, with 95% CI = [2.826,3.910],

$\beta_3 = 11.086$, with 95% CI = [6.333,15.839],

indicate that all regression coefficients are significantly different from zero.

The scatterplot in Figure 5.2.3, obtained by selecting the plot option "*Model vs data*", shows the correspondence between the model values, as derived from the independent variables and the data values observed as dependent variable. Most dots are in between the two red lines that signal a deviation that is maximally $2 \cdot \sigma$, where

$\sigma = 46.61$, with 95% CI = [42.39,51.58],

is the estimated standard deviation of the Gaussian distributions, indicating a good correspondence between the observed data and the linear predictions.

Note that, unlike in Figure 5.1.3, the prediction error seems to be approximately constant across the entire range of model values, so that GLM is unlikely to provide an improved model (we leave it up to the reader to verify this impression).

Traditional Regression Statistics

ILLMO contains a fully functional module for doing multiple regression in the traditional way, inspired by the book *Applied Regression Analysis* (Draper and Smith, 1998). This module is available through either the "*Multiple Regression (main results)*"

option or the "*Multiple Regression (detailed results)*" option associated with the **traditional statistics/analyses** button in the **stats** dialog box. Note that the module assumes continuous data with Gaussian distributions of constant variance across conditions.

It is beyond the scope of this book to explain in detail all the textual and graphical output that the multiple regression module can produce, but of course the user is encouraged to try out the features being offered and to compare the obtained results with those of more traditional statistical packages (obviously, in many cases, the detailed outcomes should be identical).

Several popular linear regression outcomes, such as the ***prediction error*** (or residue), ***influence statistics*** (of Cook and Atkinson), ***leverages*** and ***predictions,*** with or without confidence intervals, are also rendered in a separate display window whenever multiple regression is performed, in order to supply graphical, next to textual output. The output files generated by the module are stored into a subdirectory of the current directory; the name of this subdirectory reflects the current data and time.

Some outputs from the multiple regression module are worth mentioning for our example case. For instance, the reported regression coefficients of

$\beta_1 = 0.0849$, with 95% CI = [0.0712,0.0985],

$\beta_2 = 3.367$, with 95% CI = [2.820,3.915],

$\beta_3 = 11.086$, with 95% CI = [6.280,15.893],

and the estimated standard error of $s = 47.09$, are extremely close to the estimates that were obtained previously by means of the much more general LLP intersection method that is the default approach in ILLMO.

The fraction of the variance in the observed data that is explained by the multiple regression is reported to be $R^2 = 0.665$, which is about double of what was obtained previously with the simple regression of "sales" against "adverts". The traditional measure of **effect size** is the correlation coefficient

$R = 0.815$, with 95% CI = [0.763,0.857].

Partial Correlations

A popular measure for assessing the individual contribution of an independent variable on the prediction of the dependent variable is the **partial correlation**. The partial correlation for a single independent variable is obtained by performing a simple regression between this independent variable and the residue obtained from the multiple regression of the remaining independent variables with the dependent variable. The partial correlation hence expresses how well an independent variable is able to predict the variation in the dependent variable that cannot be

```
#
# PARTIAL CORRELATIONS
# (after removing effect of other variables)
#
  eta(Y,X1) = 0.658847 (squared=0.434079)
    confidence interval [ 0.571754, 0.731256 ]
  eta(Y,X2) = 0.654613 (squared=0.428518)
    confidence interval [ 0.566721, 0.727773 ]
  eta(Y,X3) = 0.308938 (squared=0.0954426)
    confidence interval [ 0.176843, 0.430100 ]
```

ILLUSTRATION 5.2.1 Partial correlations reported as part of the multiple regression of *Record2.csv*

accounted for by the other independent variables. The obtained partial correlations in our example case are reported in Illustration 5.2.1. Note that the contribution of the third independent variable is substantially smaller than the contributions of the other two independent variables.

An equivalent of partial correlations can also be created by the more general modeling approach used in ILLMO. We illustrate this for the case of the third predictor. We start by performing a multiple regression between the original data and the first two predictors. This can be accomplished by clicking on the **3D pattern** button on the Model(Confirmatory) subpage of the Scaled

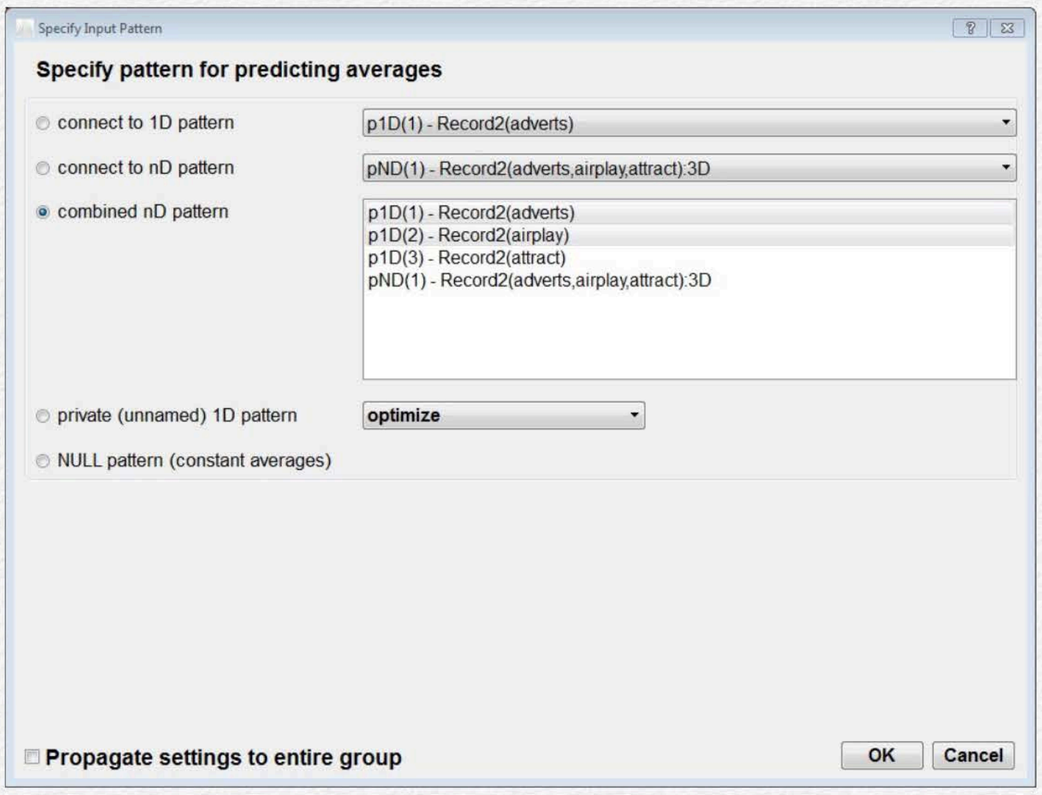

FIGURE 5.2.4 Specifying a new 2D pattern for multiple regression by combining two existing 1D patterns.

Attributes page of the ILLMO interface, which opens the dialog box in Figure 5.2.4. The figure shows how a combined 2D pattern has been specified by selecting two out of the three independent variables, i.e., "adverts" and "airplay". Upon clicking the **OK** button, the specified 2D pattern is used to perform multiple regression, which is for instance evidenced by the fact that the button label changes from **3D pattern** to **2D pattern**.

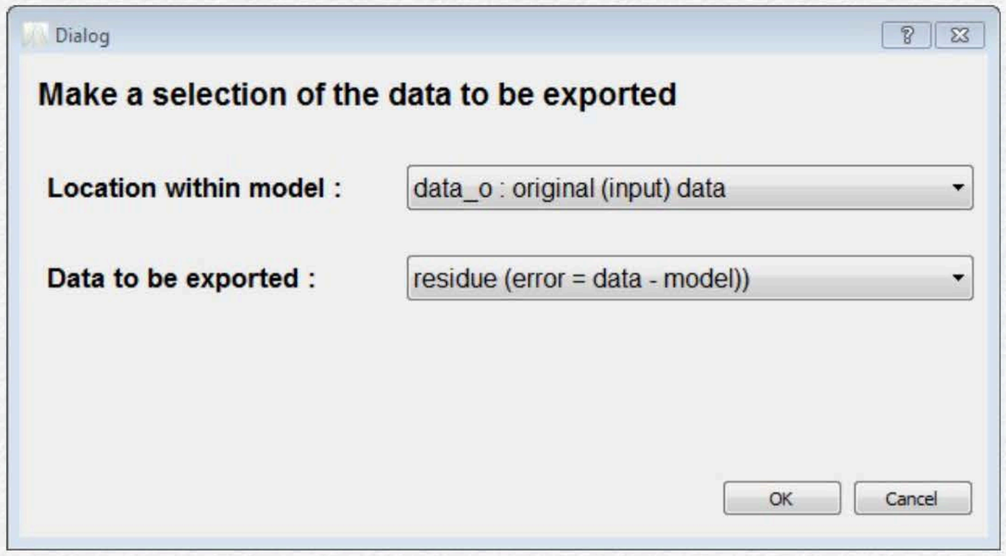

FIGURE 5.2.5 Dialog window for specifying which data to export

We can now export the residue data, i.e., the differences between the original data and the predicted averages, using the button **create copy of data** in the **data** dialog window. This will produce the dialog window shown in Figure 5.2.5 in which the selection behind "*Data to be exported:*" can be changed from the default option "*data*" into "*residue (error=data-model)*". This will create a new Scaled Attribute (under Number=2) on which we can perform simple regression against the third predictor named "attract". This simple regression has an effect size of

$R = 0.303$, with 95% CI = [0.171,0.425],

for the (partial) correlation that is indeed very close to the value reported in Illustration 5.2.1.

Multi-Model Regression

ILLMO also offers an interesting alternative for what is traditionally called ***hierarchical regression***. In hierarchical regression, different alternative regression models are compared. As traditional methods can only compare two models at a time, a hierarchy needs to be established, as a model can only be compared pairwise to models that are lower or higher in the hierarchy. In hierarchical regression, one model is higher up in the hierarchy than another model if the first model contains all predictors that are present in the second model, plus some extra ones. MMC is more flexible as it can compare several models at the same time, even if such models are not hierarchically related.

ILLMO uses the **ANOVA** tables produced by traditional multiple regression analyses to create MMC tables. Specifically, by selecting the option "*Multi-Model Regression (drop variables)*" in the **stats** dialog window all (7) multiple regression models with zero, one and two independent variables dropped are compared, which produces the **multi-model regression** table in Illustration 5.2.2. Note that the model with all three independent variables included is by far the most likely one in the example case. The weights for each independent variable are obtained by summing the weights for those models that include that spe-

```
Multi-Model Comparison:
 X1 X2 X3 :      AIC    Likelihood  Weight
 x  x  x  :   2114.65   1.000000  0.999874  (1)
    x  x  :   2226.40   0.000000  0.000000  (2)
 x     x  :   2224.45   0.000000  0.000000  (3)
 x  x     :   2132.60   0.000126  0.000126  (4)
       x  :   2306.50   0.000000  0.000000  (5)
    x     :   2240.13   0.000000  0.000000  (6)
 x        :   2247.50   0.000000  0.000000  (7)

Total weight for each independent variable:
 X1 :  1.00000
 X2 :  1.00000
 X3 :  0.99987
```

ILLUSTRATION 5.2.2 MMC table produced for the multi-model regression of *Record2.csv*

cific independent variable. As all these weights are close to one, we again conclude that the model with all three variables included is indeed the most probable one. Note that ILLMO also produces a dialog window with this MMC table, similarly as in Figure 4.1.3, so that MMC comparisons can also be performed with some of the regression models removed.

Model Averaging

A second part of the output produced by the multi-model regression analysis is called ***model averaging*** and is reproduced in Illustration 5.2.3. The reported regression coefficients are the weighted sums of the coefficients obtained in the separate multiple regressions, as proposed in the book on *Model Selection and Multimodel Inference* (Burnham and Anderson, 2002). We refer to this book for a more extensive discussion on model averaging. In the example case, where almost all weight is attributed to the model with three predictors, there is hardly any difference with the coefficients reported earlier for this most likely regression model. This is not necessarily the case for all example data sets (i.e, when no single model has a weight close to one), as will be demonstrated in later examples.

Note that Illustration 5.2.3 shows only half of the output generated by ILLMO when performing model averaging. In their book, Burnham and Anderson suggest two alternative ways of averaging across models. In the case of Illustration 5.2.3, the weighting is applied across all (7) models. An alternative option is to average only across those models in which a specific independent variable occurs (which will be 4 out of 7 models in the example) when deriving the corresponding weighted regression coefficient for that variable. ILLMO offers the outcomes for both alternatives, so that the user himself can judge the impact of

```
Weighted regression coefficients & standard errors:
 (weighted across all models)
  C0 : -26.6044, se_C0 = 17.3564, ase_C0 = 17.4687
     -> CI = [-60.6059,7.39705] (standard error)
     -> CI = [-60.8258,7.61694] (adjusted standard error)
     -> raise LLP threshold by factor 1.00074 (to 95.0088 pct)
  C1 : 0.0848851, se_C1 = 0.00692309, ase_C1 = 0.00696789
     -> CI = [0.0713227,0.0984475] (standard error)
     -> CI = [0.0712349,0.0985353] (adjusted standard error)
     -> raise LLP threshold by factor 1.00002 (to 95.0005 pct)
  C2 : 3.36745, se_C2 = 0.277781, ase_C2 = 0.279579
     -> CI = [2.82328,3.91163] (standard error)
     -> CI = [2.81975,3.91515] (adjusted standard error)
     -> raise LLP threshold by factor 1.00008 (to 95.0012 pct)
 C3 : 11.0849, se_C3 = 2.43894, ase_C3 = 2.45471
     -> CI = [6.30702,15.8629] (standard error)
     -> CI = [6.27613,15.8937] (adjusted standard error)
     -> raise LLP threshold by factor 1.00089 (to 95.0105 pct)
```

ILLUSTRATION 5.2.3 Weighted regression coefficients and standard errors for *Record2.csv*

this choice. Very often, as in the example case of one clearly preferred model, the effect is actually quite small.

The cited book also proposes an alternative way for estimating the standard error of the predicted regression coefficients. Where the standard error is usually derived solely from the most-likely model, the ***adjusted standard error*** takes into account the likelihoods of all regression models. The non-zero likelihoods of alternative models are reflected in the adjusted standard error being larger than the standard error according to the most likely model. Of course, both estimates will be very close if the weight of the most-likely model is close to one, i.e., when there is little uncertainty about this model. The output in Illustration 5.2.3 shows that ILLMO reports two confidence intervals for each of the regression coefficients, one derived from the standard error and one derived from the adjusted standard error.

The cited book also proposes an interesting approach for how to take ***model selection uncertainty*** into account in the LLP intersection method that we have been advertising throughout this book. The authors suggest to adopt the model with the highest weight but to raise the LLP threshold for each regression coefficient by a factor that is derived from the multi-model regression analysis. This corresponds to setting a higher percentage of required confidence when performing the LLP intersection. For example, in case of the third regression coefficient

$\beta_3 = 11.086$, with 95.0105% CI=[6.331,15.842]

the CI is slightly larger than the one derived earlier when only considering the most likely regression model (with 3 predictors).

Logistic Regression

Within ILLMO, multiple regression can also be applied in case the observed data is binary. We again take an example from the chapter on **logistic regression** in the book(s) by Field. More specifically, the file *penalty.dat* contains 4 columns and 75 entries. The fourth column entitled “Scored” is the dependent variable, and indicates whether (1) or not (0) a player scores a penalty. The first and second columns are measures for how “Worried” or “Anxious” a player is about taking penalties, while the third column is the percentage of penalties scored previously. All three latter columns can be used as independent variables. The file can be read using the *File* menu option “*Open CSV file (with attributes, field separator:tab)*” and should be interpreted as NOT containing measurements for different conditions in separate columns. In order to obtain logistic regression, the probability distribution should be changed from Gaussian to logistic in the **prob** dialog window and the standard deviation should be fixed.

The dialog window **to1D** allows to inspect the regression coefficients and to determine their CIs using the LLP intersection method. The resulting values of

$\beta_1 = -0.0363$ (with 95% CI=[-0.0639,-0.0154])

$\beta_2 = 0.0398$ (with 95% CI=[-0.0297,0.1164])

$\beta_3 = 0.0292$ (with 95% CI=[-0.00577,0.0689])

indicate that the second and third regression coefficients are not significantly different from zero.

Removing the third predictor produces a multiple regression model that is almost equally likely, i.e., an MMC assigns a weight of 0.553 to the model with 3 predictors, and a weight of 0.447 to the model with 2 predictors (i.e., an likelihood ratio of 1.237). The regression coefficients for the model with 2 predictors are

$\beta_1 = -0.0320$, with 95% CI = [-0.0574,-0.0127],

$\beta_2 = -0.0168$, with 95% CI = [-0.0303,-0.0062],

which are both significantly different from zero. The fact that the second regression coefficients changes drastically by the removal of the third predictor indicates a substantial correlation between the second and third independent variables.

Moving to the Stimulus Patterns page of the ILLMO interface and pressing the **correlate patterns** button will reveal the dialog window in Figure 5.2.6. The result of performing regression between the second independent variable (Anxious) and the third one (Previous) is shown both graphically and textually. The two variables are indeed highly correlated (R=0.993), which explains why multiple regression with both predictors included shows little advantage over including only one predictor

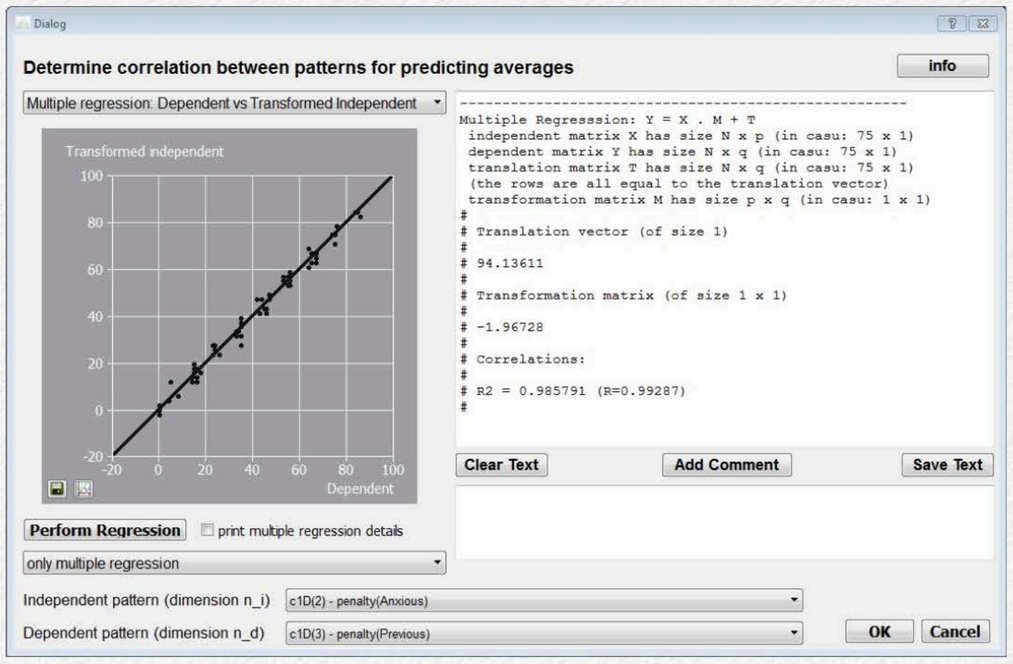

FIGURE 5.2.6 Dialog window showing correlations between the patterns for Anxious and Previous (data: *penalty.dat*)

(and can even become numerically unstable if the correlation approaches one).

A multi-model regression with all predictors produces the outcome in Illustration 5.2.4, attributing maximum weight to a model with the first (Worried) and third (Previous) independent variables as predictors. A MMC between this model with 2 predictors and the model with 3 predictors results in weights of 0.621 and 0.379 for the former and latter model, respectively.

Multi-Model Comparison:

X1 X2 X3 :	AIC	Likelihood	Weight	
x x x :	58.21	0.431728	0.221987	(1)
x x :	71.58	0.000541	0.000278	(2)
x x :	56.53	1.000000	0.514183	(3)
x x :	57.88	0.508316	0.261367	(4)
x :	69.38	0.001622	0.000834	(5)
x :	70.61	0.000876	0.000451	(6)
x :	69.23	0.001751	0.000900	(7)

Total weight for each independent variable:

X1 : 0.99844
X2 : 0.48408
X3 : 0.73728

ILLUSTRATION 5.2.4 MMC table produced for the multi-model regression of *penalty.dat*

According to a maximum-likelihood approach, the regression coefficients for a model with only the first and third predictor included are

$\beta_1 = -0.0327$, with 95% CI = [-0.0587,-0.0132],

$\beta_3 = 0.0092$, with 95% CI = [0.0037,0.0163],

which are indeed both significantly different from zero.

```
Multi-Model Comparison:
 X1 X2 :      AIC    Likelihood  Weight
 x  x  :     56.53   1.000000  0.996638  (1)
    x  :     69.38   0.001622  0.001617  (2)
 x     :     69.23   0.001751  0.001745  (3)
Total weight for each independent variable:
 X1 :  0.99838
 X2 :  0.99825

Weighted regression coefficients & standard errors:
 (weighted across all models)
 C0 : 0.584854, se_C0 = 0.177503, ase_C0 = 0.180586
    -> CI = [0.237124,0.932584] (standard error)
    -> CI = [0.231084,0.938623] (adjusted standard error)
    -> raise LLP threshold by factor 1.01731 (to 95.1944 pct)
 C1 : -0.0258736, se_C1 = 0.00651481, ase_C1 = 0.00662807
    -> CI = [-0.0386362,-0.013111] (standard error)
    -> CI = [-0.0388581,-0.0128891] (adjusted standard error)
    -> raise LLP threshold by factor 1.0157 (to 95.1768 pct)
 C2 : 0.00884124, se_C2 = 0.00223954, ase_C2 = 0.00227846
    -> CI = [0.00445396,0.0132285] (standard error)
    -> CI = [0.00437769,0.0133048] (adjusted standard error)
    -> raise LLP threshold by factor 1.01601 (to 95.1802 pct)
```

ILLUSTRATION 5.2.5 Weighted regression coefficients and standard errors for *penalty.dat*

Illustration 5.2.5 shows the corresponding outcome of the multi-model regression analysis if only the first and third predictor are included. Note that the regression coefficients are different from those reported in the previous paragraph. This is hardly surprising given the differences in model assumptions. The maximum-likelihood model in ILLMO assumes discrete (binary) data and logistic distributions, and uses iterative log-likelihood estimation of the parameters, while the multi-model regression analysis assumes continuous data with Gaussian distributions and uses analytic expressions for estimating the parameters and their confidence intervals.

Confounding Variables

A difficult issue in empirical research, and hence also in statistical analysis, is the occurrence of **confounding variables**. A confounding variable is an independent variable that is not of interest by itself but that may affect the relationship between the dependent variable and other independent variables that are of interest. The difficulty in establishing the latter relationship is caused by the fact that there not only exists a correlation between the confounding variable and the other independent variables, but also between the confounding variable and the dependent variable. The traditional statistical method **Analysis of Covariance (ANCOVA*)*** addresses such situations by combining regression against the confounding variables with ANOVA against the independent variables of interest.

We discuss the example from the chapter on ANCOVA in the book(s) of Field to demonstrate how we propose to handle confounding variables in ILLMO in an elegant way, and extend traditional ANCOVA in the process.

More specifically, the file *ViagraCovariant.csv* contains measurements for 30 individuals that are coded into 5 columns. Column 1 entitled “Dose” contains the number of the experimental condition: either 1 (no Viagra), 2 (low dose of Viagra) or 3 (high dose of Viagra). Column 2 entitled “Libido” contains the observed variable and is a measure for the libido of the participant, i.e., an integer value in the range [2,9]. Column 3 entitled “Partner_Libido” is a similar measure for the libido of the partner of the subject. The two last columns entitled “Low_Placebo” and “High_Placebo” provide two **dummy variables** that together provide a binary code for the diverse conditions: either (0,0) (placebo), (1,0) (low dose Viagra) or (0,1) (high dose Viagra).

A classical analysis using a one-way **ANOVA** of the observed data against the independent variable “Dose” results in an F-ratio $F(2,27) = 2.416$ (with p=0.108) so that there is insufficient evidence to conclude that “Dose” has a significant effect on “Libido”. The ANOVA can be mapped to a MMC which leads to a similar conclusion as it results in an likelihood ratio of 1.114 in favor of a “constant” (or NULL) model over a model in which the averages vary across conditions.

Interpreting the “Libido” as a discrete variable and performing Thurstone modeling on it leads to a similar conclusion. A MMC between a NULL model, with constant averages across “Dose” conditions, and a model with variable averages reports an evidence ratio of 1.092 in favor of a model with variable averages.

ANCOVA within ILLMO (Shared Regression Slope)

In order to perform regression analysis on the data, it needs to be entered into ILLMO, after specifying the number of conditions as 30, as a CSV file with “Libido” being the column with dependent values, and the last three columns as independent variables (note that the field separator is a semicolon). It is expected a priori that “Partner_Libido” could be a confounding variable.

A multiple regression of the observed variable “Libido” against the three independent variables (X1 = “Partner_Libido”, X2 = “Low_Placebo” and X3 = “High_Placebo”) reports the partial correlations in Illustration 5.2.6. The first partial correlation confirms that there exists a correlation between the dependent variable “Libido” and the independent variable “Partner_Libido”. The latter two partial correlations moreover indicate that the Viagra dose, coded in the latter two variables, does have an effect on the “Libido” if the effect of the confounding variable is removed. We can hence expect that the variable “Partner_Libido” also correlates with the Viagra dose. A multiple regression of

```
eta(Y,X1) = 0.400214 (squared=0.160171)
  confidence interval [ 0.024017, 0.677125 ]
eta(Y,X2) = 0.381184 (squared=0.145301)
  confidence interval [ 0.001562, 0.664777 ]
eta(Y,X3) = 0.477534 (squared=0.228039)
  confidence interval [ 0.119332, 0.725740 ]
```

ILLUSTRATION 5.2.6 Partial correlations reported as part of the multiple regression of *ViagraCovariant.csv*

"Partner_Libido" against the other two independent variables does indeed reveal a correlation of

$R = 0.358$, with 95% CI = [-0.010,0.640]

which borders on significance.

Step 1: Removing the Confounding Variable

In ANCOVA, the effect of the confounding variable "Partner_Libido" on the observed variable "Libido" is removed (or reduced) by performing a simple regression between both variables and subtracting the predicted averages, which are linearly related to the independent variable "Partner_Libido", from the original observations for the dependent variable "Libido". The linear regression coefficient in this simple regression is

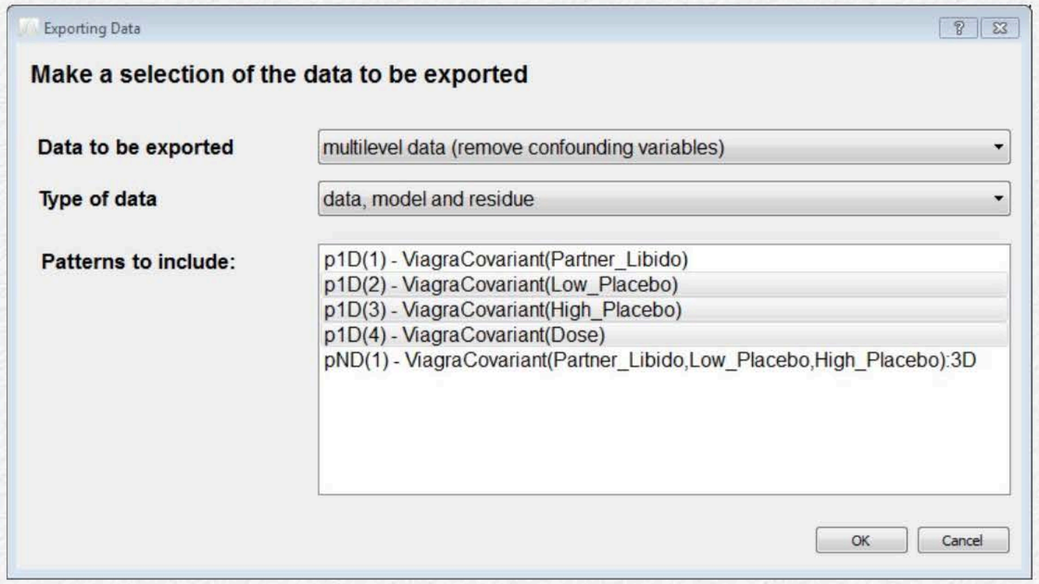

FIGURE 5.2.7 Dialog box for specifying options when exporting data to a CSV file (data: *ViagraCovariant.csv*)

$\beta_1 = 0.260$, with 95% CI = [-0.118,0.637],

while the correlation is

$R = 0.246$, with 95% CI = [-0.125,0.557].

We can subsequently use the option "*save current attribute as CSV file*" from the *File* menu to save key information from this simple regression. The dialog box displayed in Figure 5.2.7 appears after specifying an output file (which we made available under the name *ViagraCovariantSubtracted.csv*). The data to be exported is set to *"multilevel data (remove confounding vari-*

ables)", while the type of data to be exported can be either: "*residue (=data-model)*" or "*data, model and residue*", and we have selected the latter option in Figure 5.2.7. Note how all independent variables (Dose, Low_Placebo and High_Placebo), except the confounding variable itself, have also been selected to be included in the output file.

Step 2: Analyzing the Residue Values

We can perform either an exploratory or a confirmatory analysis of the independent variable entitled "residue_ML" in the file *ViagraCovariantSubtracted.csv.* By clicking on the "New project file" icon at the left of the ILLMO toolbar, we can open a second ILLMO window, in which we set the number of experimental conditions equal to 3 (which are the number of "Dose" conditions). Once the data has been entered into this second ILLMO window, the "*Degrees of freedom used in pre-processing the data*" should be set to 1 in the **data** dialog window to reflect that a simple regression (with 1 DOF) was used to pre-process the data.

A traditional exploratory analysis (such as performed within ANCOVA) would apply a one-way **ANOVA** of the residue data as a function of the independent variable "Dose". This results in an F-ratio of $F(2,26) = 3.692$ (with p=0.039), so that there is sufficient evidence to conclude that "Dose" has a (marginally) significant effect on "Libido". This ANOVA can be mapped to a MMC, where a comparison is made between a "variable" model, where the average "Libido" varies with "Dose", and a NULL model, where the average "Libido" is not affected by "Dose", which leads to a similar conclusion as it results in an likelihood ratio of 2.78 in favor of a "variable" model.

The problem with this traditional approach is however that a closer look at the histograms in Figure 5.2.12 shows that the assumption of Gaussian distributions with constant variance on which this ANOVA is based is clearly invalid. ILLMO can however also adopt a Gaussian model with condition-dependent variance (estimated at $\sigma_{placebo} = 1.26$, $\sigma_{low} = 0.96$ and $\sigma_{high} = 2.14$, respectively). The weights for a model with condition-dependent variance and a model with constant variance are 0.964 and 0.036, respectively, or a likelihood ratio of 26.57.

The LLP intersection method can subsequently be used to construct the CIs for the differences in averages between the distinct conditions, resulting in significant differences "low dose - placebo" of

$d_{low} = 1.736$, with 95% CI = [0.612,2.860],

and "high dose - placebo" of

$d_{high} = 1.999$, with 95% CI = [0.508,3.490],

and an insignificant difference "high dose - low dose" of

$d_{high-low} = 0.263$, with 95% CI = [-1.147,1.673].

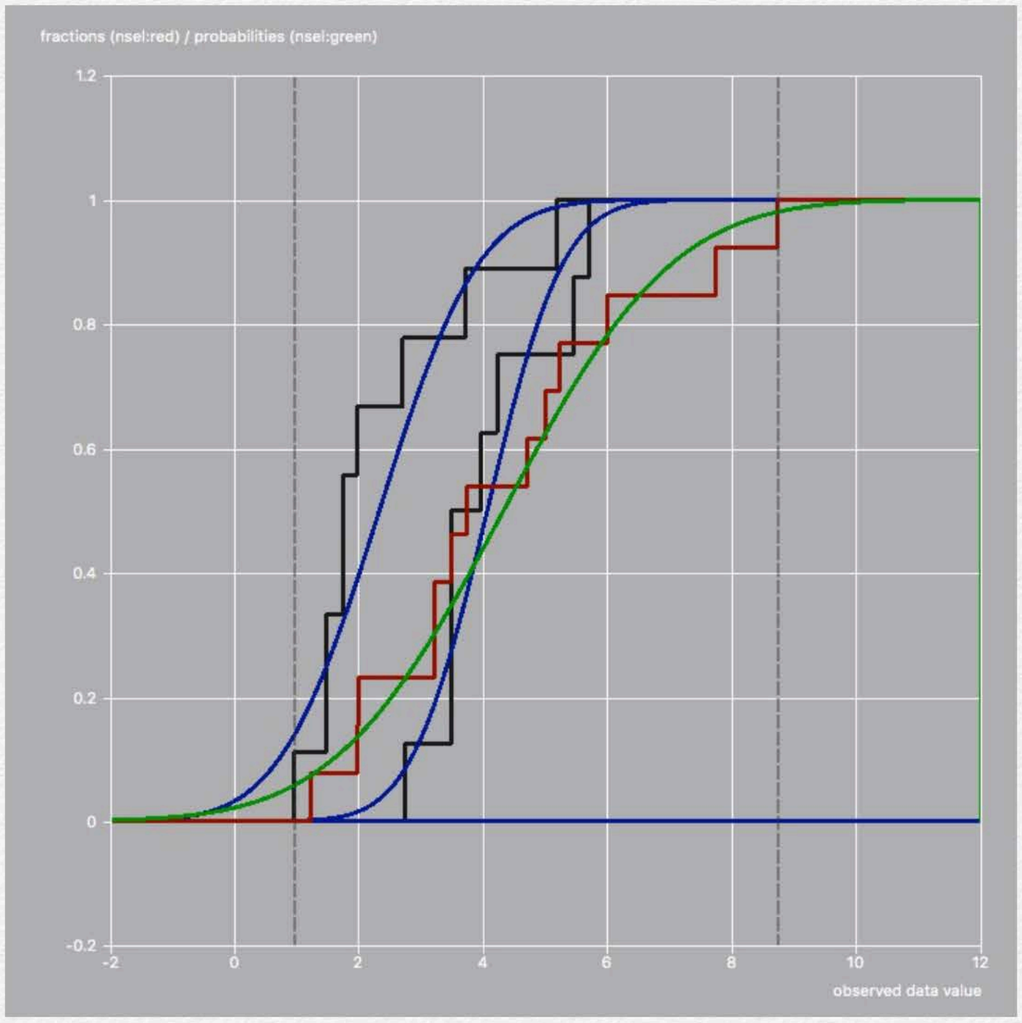

FIGURE 5.2.8 Histograms of the residue values produced by subtracting the effect of a confounding factor (data: *ViagraCovariantSubtracted.csv*)

These differences between the reference condition (placebo:1) and the other conditions (low dose:2 and high dose:3) are illustrated, together with their CIs, in Figure 5.2.9. We could also have applied a **Bonferonni correction** by adopting a 98.33% confidence to account for the 3 unplanned pairwise comparisons that can be performed.

At the end of his chapter on ANCOVA, Field discusses shortly "what to do when assumptions are violated in ANCOVA" and concludes that "ANCOVA is the first example of a test that does not have an SPSS-friendly non-parametric test". By adopting the procedure that we have described in this section, i.e., converting the analysis of the observed data into the analysis of the residue values after regression against the confounding variable(s), the complete arsenal of statistical methods becomes automatically available, however. This means that this observed disadvantage is no longer valid in case of the more general log-likelihood modeling offered by ILLMO. For example, the non-parametric method of **empirical likelihood** can be used to estimate the differences in averages and their confidence intervals, as shown in Figure 5.2.10. Note that the results in Figures 5.2.9 and 5.2.10 are almost identical, however, in this example.

A confirmatory analysis could consist of a multiple regression of the two dummy variables "Low_Placebo" and "High_Placebo" against the residue data. If we would assume Gaussian distributions of constant variance, then the effect size could be expressed by means of the correlation coefficient

$R = 0.47$, with 95% CI = [0.118,0.717],

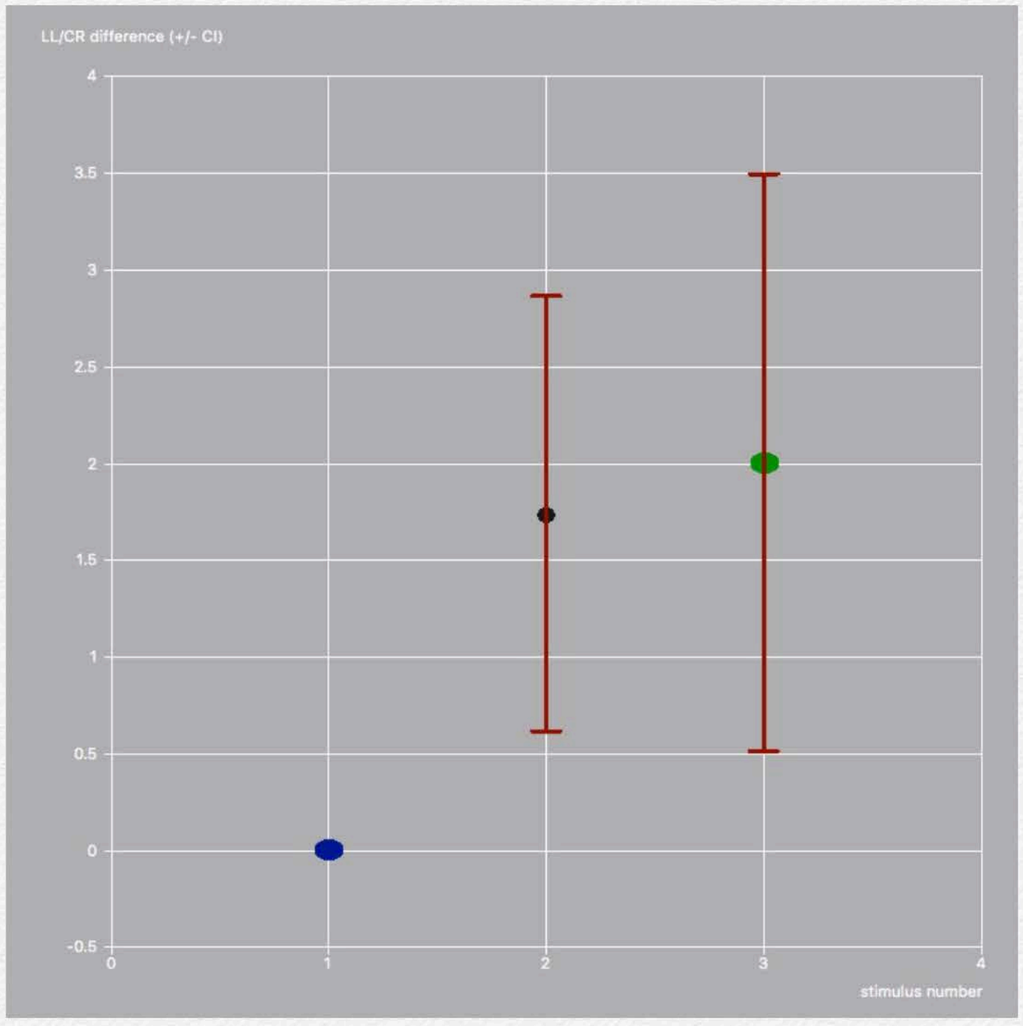

FIGURE 5.2.9 Differences in "Libido" between conditions after a simple correction for the covariant "Libido_Partner", according to the LLP intersection method (data: *ViagraCovariantSubtracted.csv*)

(or $R^2 = 0.221$, where Field reports an effect size of $R^2 = 0.227$).

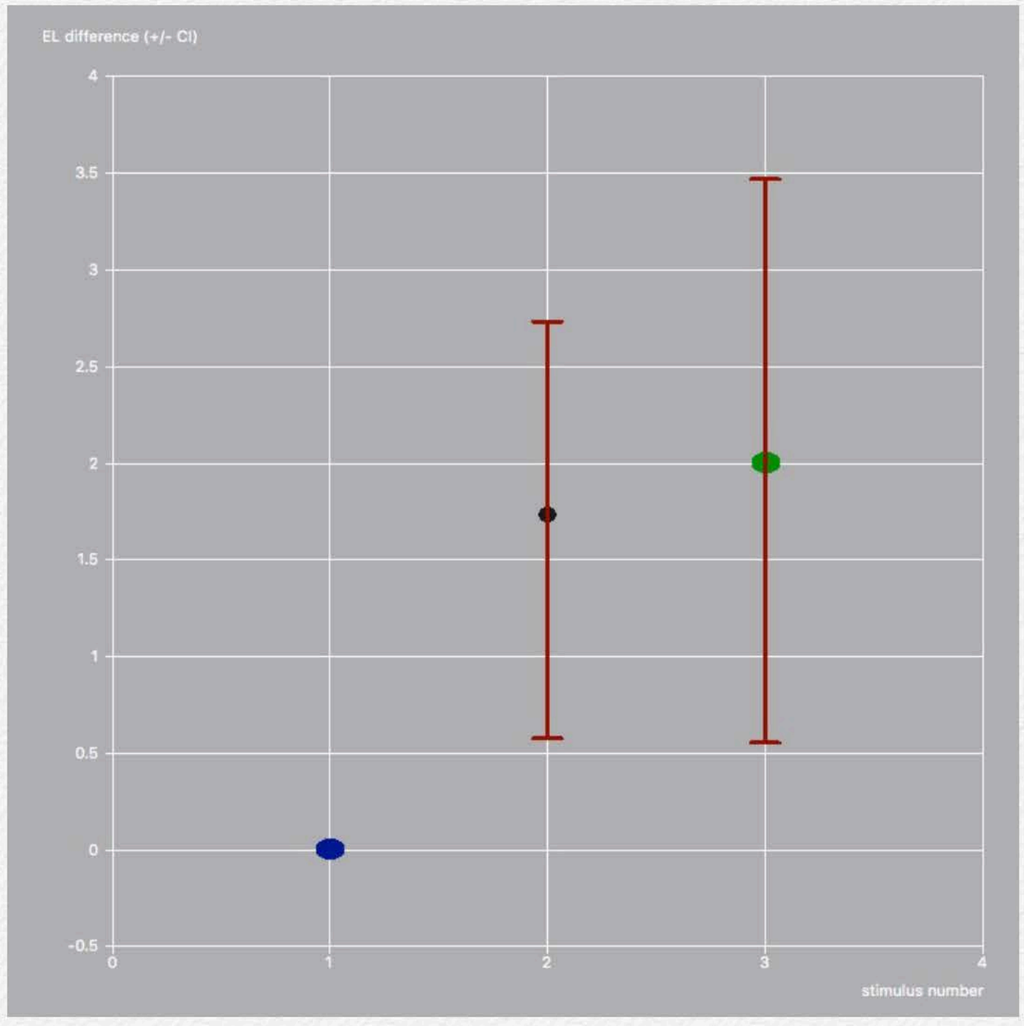

FIGURE 5.2.10 EL differences in "Libido" between conditions after a simple correction for the covariant "Libido_Partner" (data: *ViagraCovariantSubtracted.csv*)

We have however already established that this is not an adequate model. If we adopt the more appropriate model with

condition-dependent variance instead, we obtain the regression coefficients

$\beta_1 = 1.736$, with 95% CI = [0.612,2.860],

$\beta_2 = 1.999$, with 95% CI = [0.508,3.490],

which leads to the conclusion that both regression coefficients are indeed significantly different from zero. Note that, because of the coding by means of dummy variables, these linear regression coefficients are identical to the “low_dose - placebo” and “high-dose - placebo” differences found previously in the exploratory data analysis.

Beware of condition-dependent variance

Note that a condition-dependent variance not only poses a problem for traditional methods, such as one-way ANOVA and multiple regression, but also for the extended methods of multi-model regression and model averaging that we have introduced before. This is a consequence of the fact that the latter methods combine the information contained in the **ANOVA** tables that are produced upon performing multiple regressions of the same dependent variable with different combinations of independent variables. We illustrate some of the apparent contradictions that can arise by applying such techniques, without due consideration of the underlying conditions, in a case for which they are not really suitable.

A **multi-model regression** analysis of our example data would result in a weight of 0.625 for a regression model with both (dummy) variables, and weights of 0.291 and 0.083 for simple regression with the second and first variable only, respectively. This is hence an example where the weights for both independent variables

$w_{low} = 0.709$

$w_{high} = 0.917$

are substantially smaller than one, and where the ability to perform model averaging could be assumed to be of added value.

More specifically, if we perform a multiple regression with both predictors, we obtain the following regression coefficients

$\beta_1 = 1.736$, with 95% CI = [-0.030,3.501],

$\beta_2 = 1.999$, with 95% CI = [0.424,3.574],

which leads to the conclusion that only the second coefficient is significantly different from zero.

When performing model averaging instead, we take into account the likelihoods of different alternative regression models. This leads to a different conclusion, i.e, that neither of the regression coefficients is significantly different from zero, as the reported coefficients are

$\beta_1 = 1.132$, with 95% CI = [-1.010,3.273],

$\beta_2 = 1.595$, with 95% CI = [-0.241,3.431]

in case coefficient weighting is performed over all (3) models, or

$\beta_1 = 1.597$, with 95% CI = [-0.650,3.844],

$\beta_2 = 1.739$, with 95% CI = [-0.099,3.578]

in case weighting for a coefficient is only performed of those (2) models in which the corresponding independent variable actually occurs.

According to (Burnham and Anderson, 2002), the **LLP intersection method** could also be used as an alternative to model averaging, provided that the raised confidence values proposed by the multi-model regression analysis are adopted instead of the default value of 95%. In the example case this leads to yet another conclusion as the estimated regression coefficients

$\beta_1 = 1.736$, with 98.936% CI = [0.146,3.326],

$\beta_2 = 1.999$, with 97.986% CI = [0.708,3.290],

are both significantly different from zero.

It should be realized that both the model averaging and the correction to the LLP intersection method have been derived by Burnham and Anderson under the assumption that the histograms in different conditions can be modeled by Gaussian distributions with constant variance. If this assumption is not valid, as in our example, contradictory results may indeed arise.

Homogeneity of Regression Slopes

There is another issue with covariant analysis that is covered in depth in the book(s) by Field, i.e., that it should only be applied in case the effect of the confounding variable is similar in all conditions that are being compared. More specifically, the assumption of ***homogeneity of regression slopes*** needs to be verified by performing separate simple regressions between the observed data “Libido” and “Libido_Partner” in each of the three “Dose” conditions. Figure 5.2.11 shows how to make the appropriate selections in the **CSV** dialog boxes in order to easily implement such an analysis in ILLMO. There will be three sets of “Scaled Attributes” after the CSV file has been processed, one corresponding to each possible value of “Dose”, and a simple regression will automatically be performed for all three of them. Note that, while the total number of specified conditions is 30, each of the “Scaled Attributes” being created contains only a subset of these conditions,

Opening the **offset** and **gain** dialog boxes allows to subsequently request calculation of the CI for both regression coefficients in each of the three “Dose” conditions. In order to obtain an overview of the obtained results one can proceed to the

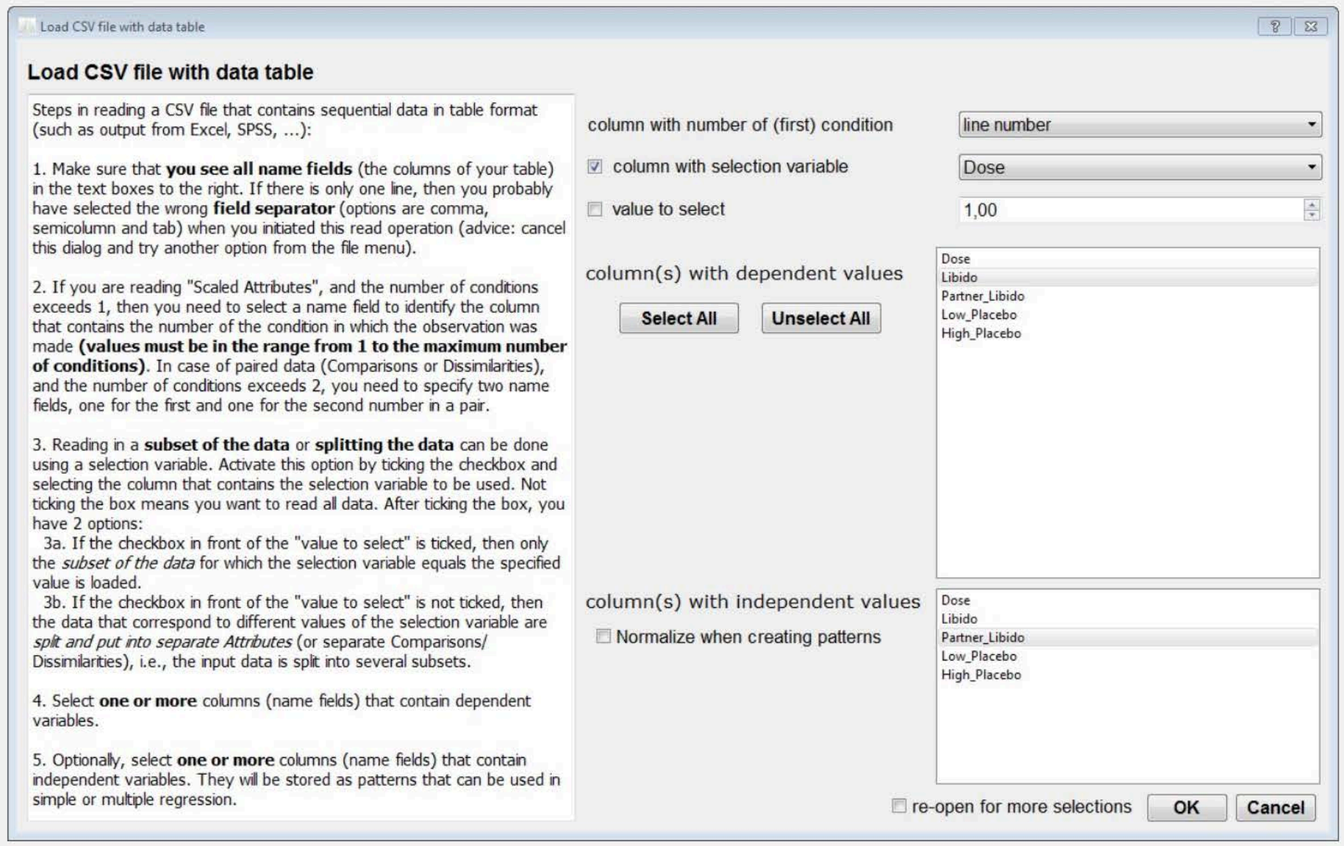

FIGURE 5.2.11 Separating the observed data "Libido" in three subsets according to "Dose" and specifying "Partner_Libido" as the independent variable to be used in simple regression (data: *ViagraCovariant.csv*)

Overview and Clustering page in the ILLMO interface. This page contains a large table with all model parameters, including the gain and offset for each of the three separate simple regressions. One option is to save this table as a CSV file, so that the data can be imported into other programs. However, more appropriate for the current discussion is the **plot overview** button that creates a visualization of several model parameters, including the gain and offset, across different attribute numbers.

Figure 5.2.12 and Figure 5.2.13 show the resulting graphs of the offset and gain in the simple regression of "Libido" against "Libido_Partner" for different values of the "Dose". Note that the gain is significantly lower in the "High Dose" condition (3) than in the other two conditions. This implies that the previous covariant analysis is actually questionable, as was already mentioned by Field.

The difference between the high-dose condition and the other conditions is also reflected in the observed correlation coefficients

$R_{placebo} = 0.883$, with 95% CI = [0.529,0.975]

$R_{low-dose} = 0.935$, with 95% CI = [0.848,0.995]

$R_{high-dose} = 0.169$, with 95% CI = [-0.421,0.658]

meaning that there is no significant correlation between the confounding variable "Libido_Partner" and the dependent variable "Libido" in the high-dose condition. This is also reflected in Fig-

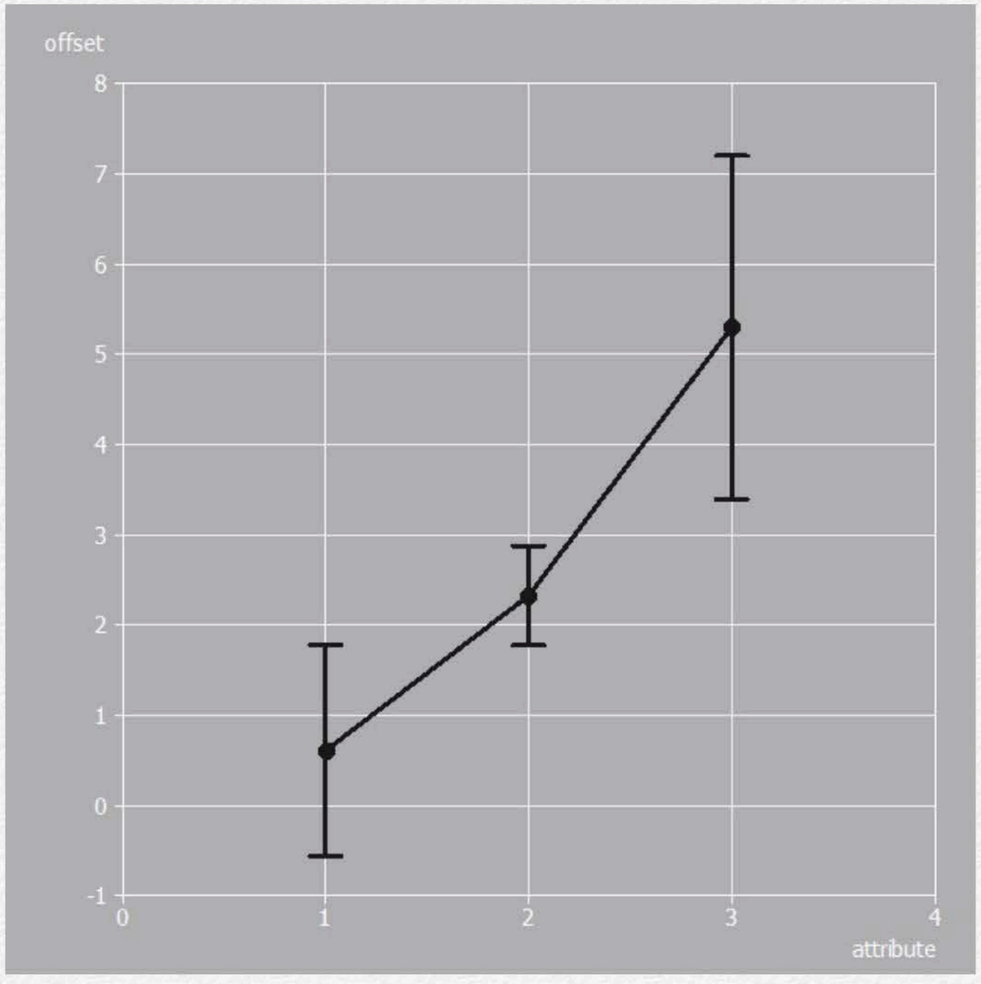

FIGURE 5.2.12 The offset in the simple regression of "Libido" against "Libido_Partner" for three different "Dose" conditions (data: *ViagraCovariant.csv*)

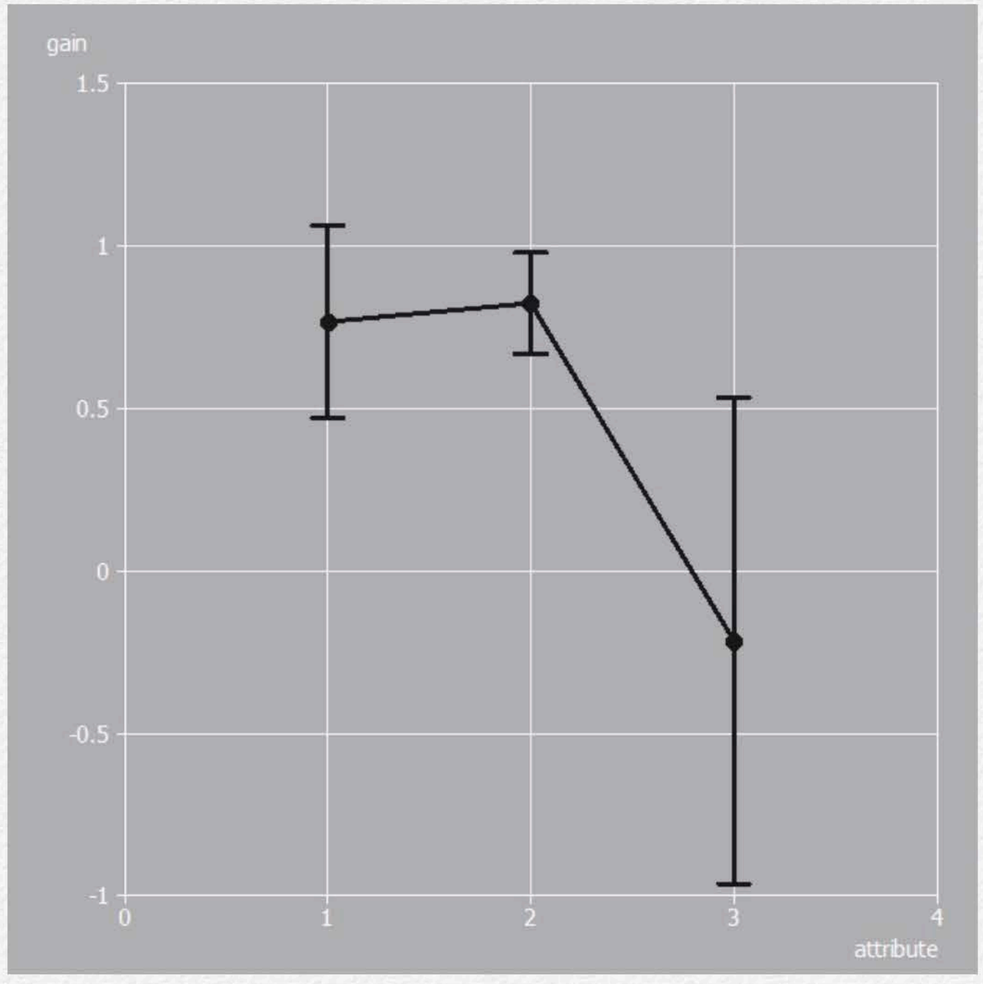

FIGURE 5.2.13 The gain in the simple regression of "Libido" against "Libido_Partner" for three different "Dose" conditions (data: *ViagraCovariant.csv*)

ure 5.2.11, where it can be observed that the gain is not significantly different from zero in this case. An alternative approach towards detecting such inhomogeneity in regression slopes will be discussed in the next section on factorial analysis.

Multi-level Modeling (Variable Regression Slopes)

A way to resolve confounding variables in case the assumption of homogeneity of regression slopes is not valid is to introduce a variation to the procedure that we have proposed earlier.

Step 1: Removing the Confounding Variable

More specifically, we propose to use the option "*save current attribute group as CSV file*" from the *File* menu to save key information from the three simple regressions that we have just introduced. After specifying an output file, which is made available under the name *ViagraCovariantMultilevel.csv*, the dialog box shown earlier in Figure 5.2.7 is displayed. Just as before, the data to be exported is set to *"multilevel data (remove confounding variables)"*, the type of data is chosen equal to "*data, model and residue*", while the independent variables (Low_Placebo and High_Placebo) are also included in the output file.

The key difference with the earlier approach is that a simple regression with a different regression slope is used for each of the "Dose" conditions when calculating the residue values. The first column in the file *ViagraCovariantMultilevel.csv* is entitled "number" and indicates the attribute number, which is equal to the "Dose" number, so that this independent variable need not be included separately.

Step 2: Analyzing the Residue Values

Just as before, the resulting independent variable "residue_ML" can be analyzed as a function of the 3 conditions indicated by "number" in either an exploratory or a confirmatory way. Once the data has been entered into a new ILLMO window in which the number of experimental conditions is set to 3, the "Degrees of freedom used in pre-processing the data" should be set to 3 in the **data** dialog window to reflect that three regression equations (with 1 DOF each) were used to pre-process the data.

A traditional exploratory analysis would consist of a one-way **ANOVA** of the residue data against the independent variable "number" (equal to "Dose"), which results in an F-ratio of $F(2,24) = 25.275$ (with p=0.000001) that is substantially larger than observed before in the covariant analysis, so that the evidence to conclude that "Dose" does indeed have a significant effect on "Libido" has become much more convincing than before. This ANOVA can be mapped to a MMC, where a comparison is made between a "variable" model, where the average "Libido" varies with the "Dose", and a NULL model, where the average "Libido" is not affected by the "Dose", which assigns a weight of 0.999997 to a "variable" model.

The problem with this approach is however that a closer look at the histograms in Figure 5.2.14 shows an even more pronounced variation in the slopes of the cumulative histograms than in the previous analysis. ILLMO resolves this by adopting

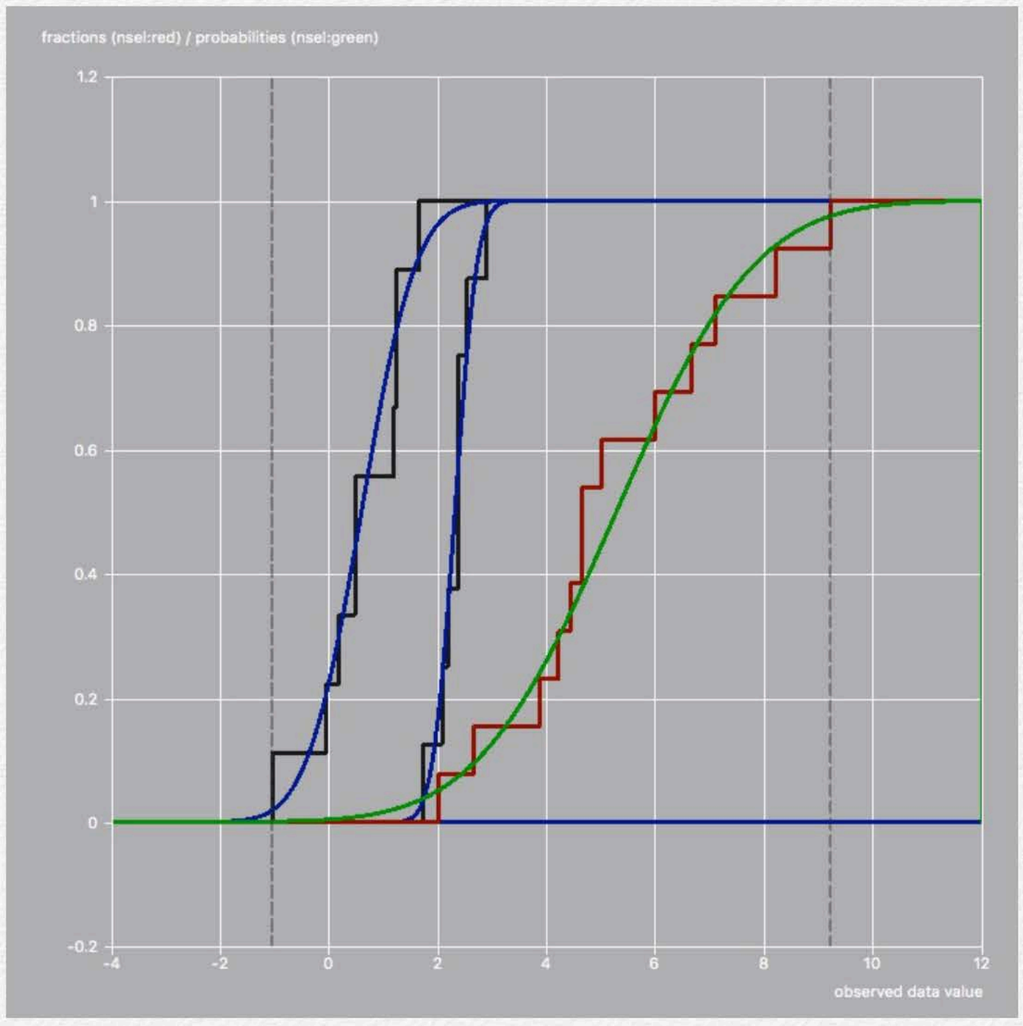

FIGURE 5.2.14 Histograms of the residue values produced by subtracting the effect of a confounding factor (data: *ViagraCovariantMultilevel.csv*)

Gaussian distributions with condition-dependent variance, which are estimated at $\sigma_{placebo} = 0.79$, $\sigma_{low} = 0.32$ and $\sigma_{high} = 2.00$, respectively.

The LLP intersection method can be used to construct the CIs for the differences between conditions, resulting in significant differences “low dose - placebo” of

$d_{low} = 1.717$, with 95% CI = [1.105,2.330],

and “high dose - placebo” of

$d_{high} = 4.690$, with 95% CI = [3.416,5.963],

but also a significant difference “high dose - low dose” of

$d_{high-low} = 2.972$, with 95% CI = [1.780,4.165].

The differences between the reference condition (placebo:1) and the other conditions (low dose:2 and high dose:3) are illustrated, together with their CIs, in Figure 5.2.15. The difference with the result obtained through a covariant analysis assuming homogeneity of regression slope (see Figure 5.2.9) is obvious. More specifically, from Figure 5.2.15 we can conclude that the low and high dose of Viagra have a different effect on the dependent measure “Libido”, something that cannot be concluded from the equivalent rendering in Figure 5.2.9.

A confirmatory analysis can be performed in a similar way as before by executing a multiple regression of the two dummy vari-

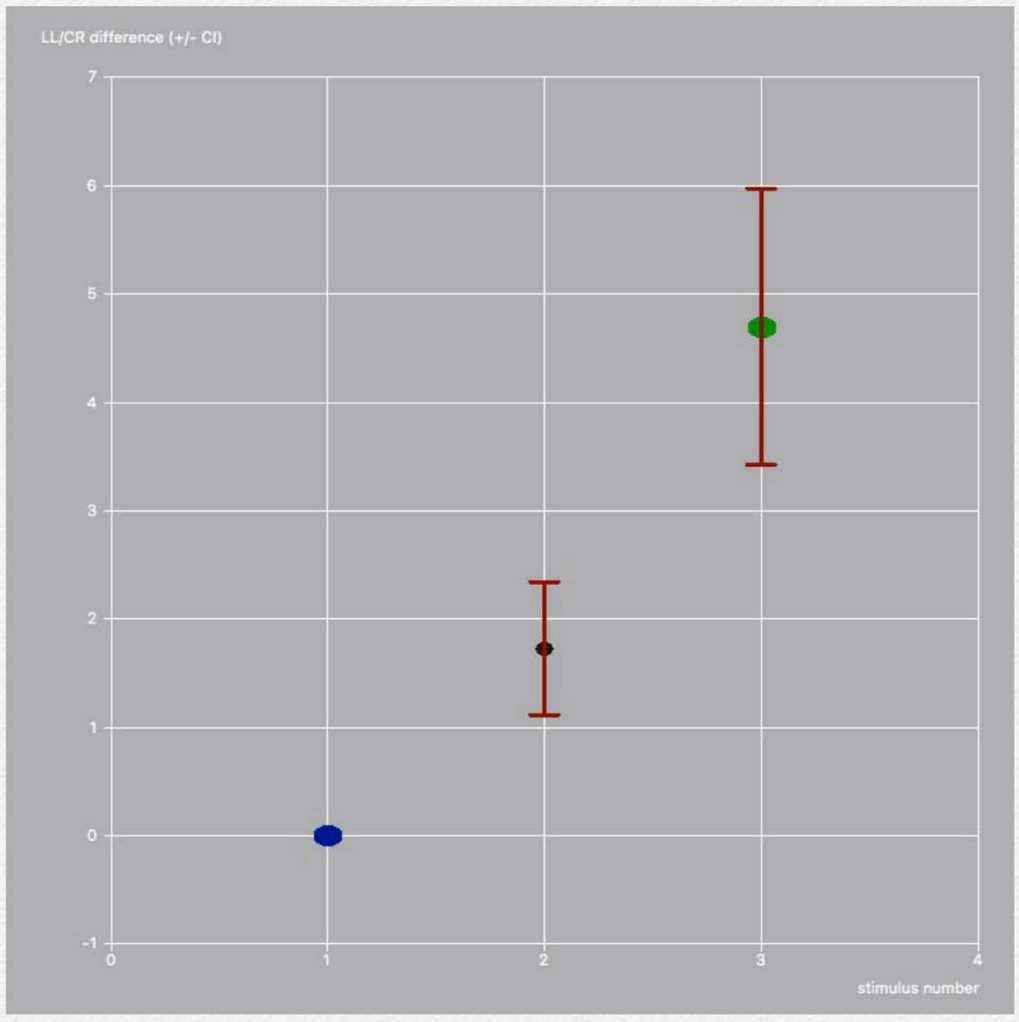

FIGURE 5.2.15 Differences in "Libido" between conditions after a multilevel correction for the covariant "Libido_Partner" (data: *ViagraCovariantMultilevel.csv*)

ables "Low_Placebo" and "High_Placebo" against the residue data, resulting in the following estimates

$\beta_1 = 1.717$, with 95% CI = [1.105,2.330],

$\beta_2 = 4.690$, with 95% CI = [3.416,5.963],

which leads to the conclusion that both regression coefficients are significantly different from zero. A consequence of using dummy variables as independent variables is again that these coefficients are equal to the differences between the averages in the low and the high dose condition, on the one hand, and the average in the placebo condition.

The approach that we have just described is an example of **Multilevel Modeling**, a statistical analysis method that Field admits he finds difficult to understand and apply in existing software programs such as SPSS. We will describe the method in more detail and provide additional examples in chapter 6.

Empirical Likelihood Regression

The discussion on **empirical likelihood (EL)** in chapter 3 has clarified how many exploratory statistical analyses, such as finding a confidence value for a distribution parameter such as the average, can also be performed in a non-parametric way. Empirical likelihood (EL) methods were introduced as an alternative to parametric log-likelihood (LL) methods and deemed especially useful in case none of the available parametric distributions, such as the Gaussian distribution, can accomplish a close fit to the observed histograms.

EL uses the ***probabilities of the observed values*** as parameters, and optimizes the LLC while imposing a number of boundary conditions on these probabilities, such as specifying the required average and standard deviation. While this approach appears to be more flexible than parametric approaches at first sight, we will show that it also has a downside that can render it less suited for confirmatory statistics. More specifically, we will show that the implicit EL assumption, i.e., that values that do not occur in the actual observations are so unlikely that they can be assigned a probability of zero, can create critical problems. Especially when the number of observations is limited, not all likely values may actually also have been observed.

In this section we discuss if and how EL can be adapted for confirmatory statistics such as regression. We first discuss the original proposal from the book on empirical likelihood (Woods, 2001), and subsequently introduce an alternative approach that is a more straightforward extension of the exploratory EL method that was introduced in chapter 3.

It is important that the reader understands that the material covered in this section is somewhat speculative, and is likely to undergo changes in the future. The method that we propose has only been tried on a limited number of examples, and not all detailed problems have yet been resolved in a satisfactory way. We nevertheless decided to include this material, both for the sake of completeness and because the method certainly has merit and potential.

Non-robust empirical likelihood regression

The book on EL (Woods, 2001) describes one possible approach towards EL regression, which we summarize here. It is assumed that there are K distinct conditions, and that there are N_k observed values x_{ki}, for $i = 1,\ldots,N_k$, in condition k, for $k = 1,\ldots,K$. EL regression adopts (positive) weights w_{ki} for the observed residue values $e_{ki} = x_{ki} - \mu_k$. In case of regression, the k-th average

$$\mu_k = \beta_0 + \sum_{l=1}^{L} \beta_l \cdot v_{kl}$$

is expressed in terms of the known independent variables with values $\{v_{kl}, l = 1,\ldots,L\}$. The regression coefficients β_l, for $l = 0,1,\ldots,L$, are the parameters to be optimized (instead of the averages themselves in case of exploratory statistics).

EL regression proposes to minimize the lack-of-fit criterium

$$ELC = -2 \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log(N w_{ki}), \text{ with } N = \sum_{k=1}^{K} N_k,$$

under the following conditions for the (positive-valued) weights

$$\sum_{k=1}^{K}\sum_{i=1}^{N_k} w_{ki} = 1$$

$$\sum_{i=1}^{N_k} w_{ki} \cdot (x_{ki} - \mu_k) = 0, \text{ for } k = 1,\ldots,K$$

The model probabilities are related to the weights through

$$p_{ki} = \frac{w_{ki}}{\sum_{j=1}^{N_k} w_{kj}}, \text{ for } i = 1,\ldots,N_k,$$

for condition $k = 1,\ldots,K$.

Obviously, a problem arises with the solution proposed by the EL regression approach if all weights $\{w_{ki}, i = 1,\ldots,N_k\}$ within one or more conditions k are equal to zero. Let us analyze the problem in somewhat more detail. Suppose that m_k and M_k are the minimum and maximum observed values in condition k, then the interval $[m_k, M_k]$ is called the ***convex hull*** for the average value μ_k. What this means is that there is obviously no weighted sum of values within this interval that can produce a result that is outside of this interval (as the weights are restricted to be probabilities that are positive numbers between 0 and 1 that sum up to 1). This means that the ELC is undefined and the probabilities $\{p_{ki}, i = 1,\ldots,N_k\}$ cannot be determined for any condition $k = 1,\ldots,K$ for which $\mu_k < m_k$ or $\mu_k > M_k$.

This requirement didn't create a problem in case of exploratory data analysis as the averages μ_k were the parameters to be optimized and they could a priori be restricted to the interval $\mu_k \in [m_k, M_k]$. In confirmatory analysis, the averages are instead derived from other parameters, i.e., the regression coefficients, and values outside of the convex hull can therefore occur. Whenever they do, the EL optimization in its original form would not be able to continue.

The approach taken in ILLMO is to adopt an optimization criterium that is equal to the ELC within the convex hull but that is also defined outside of it, in order to avoid that the optimization algorithms cannot continue when parameter values outside of the convex hull occur. As long as the optimum parameter values are contained within the convex hull, this should be a valid approach. We will spare the reader the details of the approach, partly because they are still under development. The example that follows will indeed demonstrate that we have not yet been able to resolve the problem in all cases.

In summary, as long as only parameter values within the convex hull need to be considered, the EL regression should work fine, and the results should be just as easy to interpret as before with parametric regression. More specifically, as long as the LLPs produced for regression coefficients are smooth curves that lead to well-defined confidence intervals, the results should be trustworthy. If irregularities (such as discontinuities)

occur in the LLP, however, they are very likely caused by a breach of the convex hull condition that is not adequately handled by the current algorithm implemented in ILLMO.

Despite the potential problems with it, ILLMO does support the EL regression method discussed in this section. More specifically, the method can be activated by choosing the option entitled *"Empirical likelihood (non-parametric) regression"* next to *"Probability Distributions"* in the **prob** dialog box. We illustrate the method with an example.

The data set that we introduce here was generated as part of an experiment into subjective image quality (Kayargadde & Martens, 1996). More specifically, the stimuli in the experiment were 16 variations of the same image. They were created by combining four levels of image processing (either no processing or blurring the image with a Gaussian filter with standard deviation equal to 2, 4 or 8 pixels) with four levels of additive noise (either no noise or additive noise with standard deviation equal to 7, 10 or 14 greyscale units for an 8-bit image with greyscale values in the range [0,255]). The file *mon5.stim* contains two independent variables for each of the $K = 16$ conditions; the first value on a line specifies the level of blur, while the second value specifies the level of noise. This file can be read using the "*Open TEXT file containing a stimulus pattern*" from the *File* menu, resulting in a two-dimensional pattern entitled "pND(1)-mon5(1):2D" and two one-dimensional patterns entitled "pND(1)-mon5(1:1)" and "pND(2)-mon5(1:2)". These patterns can be used to perform simple or multiple regression.

The 16 images were presented four times, in random order, to 7 subjects, resulting in a total of 28 ratings per image. After each presentation, subjects were requested to rate the three attributes blur, noisiness and quality on an integer scale from 0 to 10. The combined data for these attributes are contained in the file *mon5_bnq,scal*. This file can be read into ILLMO using the option "*Open TEXT file containing scaled attributes (.scal or .txt)*" in the *File* menu. All data can be interpreted as being discrete. After reading this file, ILLMO contains three Scaled Attributes, one for each of the attributes.

We will discuss the result of the EL regression in case the second attribute (noisiness) is regressed agains the level of applied noise (pattern "pND(2)-mon5(1:2)"), primarily because this example most clearly displays the difference between parametric (Gaussian) regression and EL regression.

In Figure 5.2.16, we show the cumulative histogram of the residue values, together with the Gaussian distribution that is used to model it, in the default parametric approach. Note that the distribution does not provide a very close fit to the histogram. In Figure 5.2.17, we show the corresponding result in case of EL regression. The irregular curve (in blue) in Figure 5.2.17 is the cumulative distribution of the weights w_{ki} in the EL regression. It agrees more closely with the observed histogram.

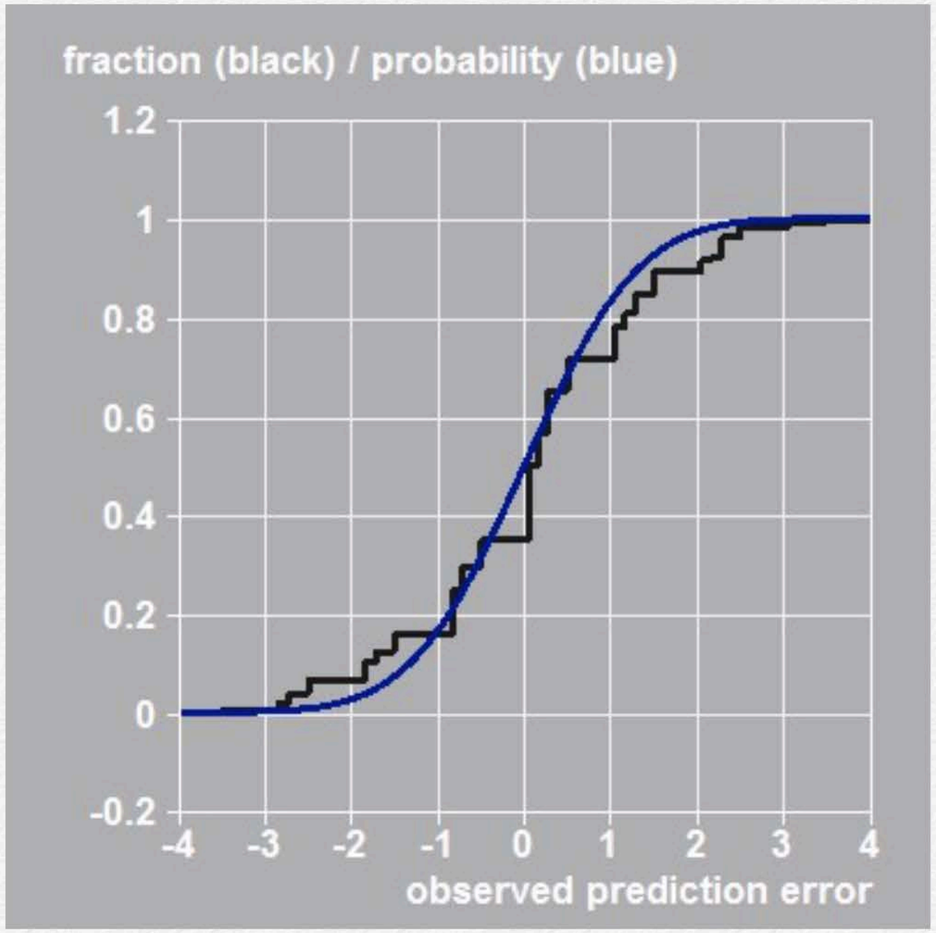

FIGURE 5.2.16 Gaussian regression: observed residues (black) and modeled distribution (blue)

The regression coefficients are estimated at

$\beta_0 = -0.046$, with 95% CI = [-0.252,0.159],

$\beta_1 = 0.555$, with 95% CI = [0.533,0.576],

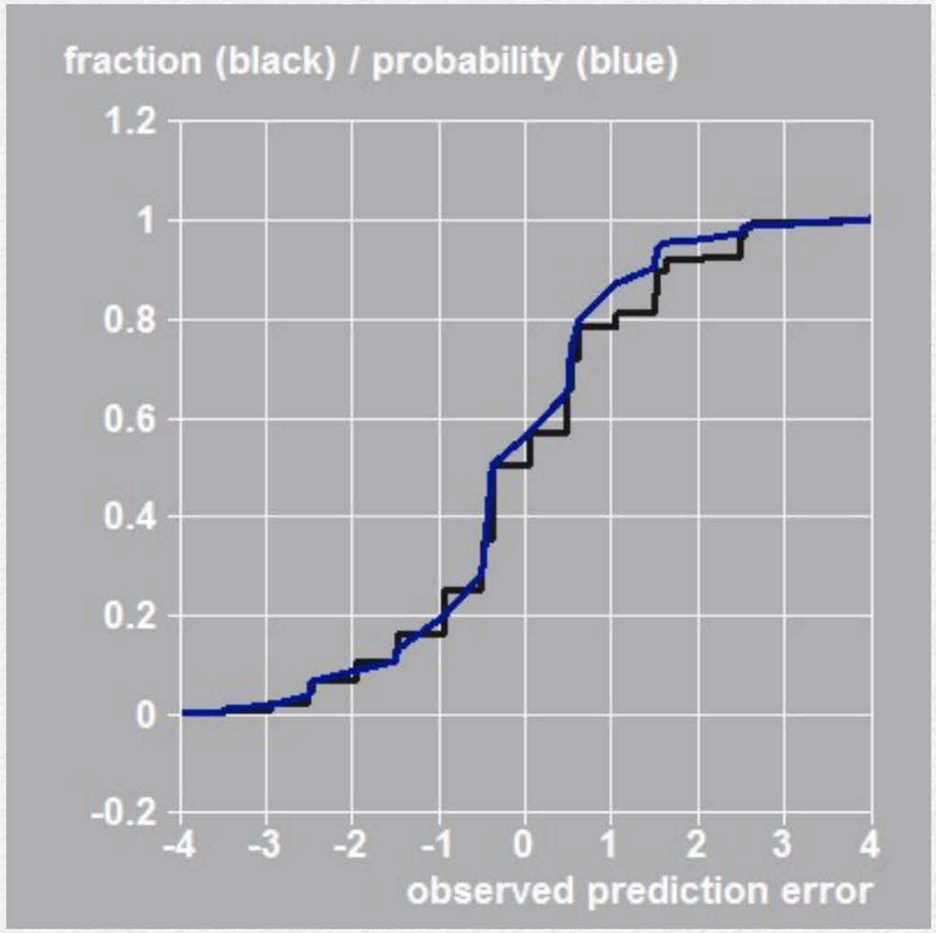

FIGURE 5.2.17 EL regression: observed residues (black) and modeled distribution (blue)

in case of the (default) Gaussian regression, and at

$\beta_{0,EL} = 0.380$, with 95% CI = [0.289,0.482],

$\beta_{1,EL} = 0.509$, with 95% CI = [0.497,0.522],

in case of the (default) Gaussian regression, so that the analysis has an effect on whether or not we conclude that the offset value β_0 is significantly different from zero.

Robust empirical likelihood regression

We propose an alternative method for EL regression that is more flexible than the one discussed in the previous section. We have seen earlier, in chapter 3, that EL minimizes the lack-of-fit ELC_k between the observed histogram in condition k and the modeled probabilities, assuming that the lower-order moments of the distribution, such as the average μ_k, standard deviation σ_k , skewness $\gamma_{1,k}$ and/or kurtosis $\gamma_{2,k}$, are specified. An additional advantage of EL is that this minimum $ELC_k(\mu_k, \sigma_k, \gamma_{1,k}, \gamma_{2,k})$ can also be calculated as a **robust statistic** in order to reduce the sensitivity to data values that are considered to be outliers, as they deviate by several standard deviations (2 or more) from the observed average value.

The number of moments that are specified can be any number between one and four. In order to simplify the notation, we will for instance use $ELC_k(\mu_k, \sigma_k, \cdot\,, \cdot\,)$ to indicate that boundary conditions are only imposed on the first two parameters, i.e., the average and the standard deviation, but not on the last two parameters, i.e., the skewness and the kurtosis.

In exploratory statistics, the overall log-likelihood criterium

$$ELC2 = \sum_{k=1}^{K} ELC_k(\mu_k, \sigma_k, \cdot\,, \cdot\,)$$

is for instance used in the LLP intersection method to determine confidence intervals for individual parameters, such as an average μ_k, or a combination of such parameters, such as a difference effect

$$JND_{kl} = (\mu_k - \mu_l)/\sqrt{\sigma_k^2 + \sigma_l^2}.$$

Conditions such as the requirement of **homoscedasticity** can easily be taken into account by assuming that all standard deviations are equal, i.e., $\sigma_k = \sigma, k = 1,\ldots,K$, hence reducing the number of parameters.

The extension to confirmatory statistics such as simple regression would seem to be fairly trivial, as the k-th average

$$\mu_k = \beta_0 + \sum_{l=1}^{L} \beta_l \cdot v_{kl}$$

is expressed in terms of the known independent variable values $\{v_{kl}, l = 1,\ldots,L\}$ and the regression coefficients, which are the parameters to be estimated.

Note that the ELC defined above cannot only be used to find the optimum regression coefficients themselves, but also to de-

termine confidence intervals for these coefficients by means of the LLP intersection method.

The procedure sketched above is however not full-proof, as ILLMO may fail to determine the optimized value $ELC_k(\mu_k, \sigma_k, \gamma_{1,k}, \gamma_{2,k})$ for some values of the parameters, i.e., when these parameters are outside of the ***"convex hull"*** that is established by the observed values. ILLMO tries to remedy such problems by adopting an optimization criterium that is equal to the ELC within the convex hull, but that is also defined outside of it, so that the optimization algorithms can continue to function when such parameter values occur.

Note that the "convex hull" problem is less likely to occur if a wide range of values is observed in all conditions. Small data sets with few observations in the different conditions can therefore be much more problematic, as the EL assumption that values that are not observed should be assigned a zero probability may simply not be realistic.

ILLMO supports 4 alternative EL regression methods, which can be selected by choosing an appropriate option next to *"Probability Distributions"* in the prob dialog box. The option entitled *"Empirical likelihood with averages as parameters"* behaves very similarly to the EL regression method proposed by Woods. It uses the criterium

$$ELC1 = \sum_{k=1}^{K} ELC_k(\mu_k, \cdot, \cdot, \cdot),$$

so that only the averages $\mu_k, k = 1,\ldots,K,$ are imposed as boundary conditions on the model distributions. This choice excludes the possibility to apply robust statistics. As long as the regressed averages satisfy $\mu_k \in [m_k, M_k]$, the optimization should work fine. For the example data set, the estimated regression coefficients, and their confidence intervals, turn out to be identical to the ones estimated earlier using the original EL regression method.

The option entitled *"Empirical likelihood with averages & variances as parameters"* instead uses the criterium

$$ELC2 = \sum_{k=1}^{K} ELC_k(\mu_k, \sigma_k, \cdot, \cdot),$$

so that both the averages and standard deviations are imposed as boundary conditions on the distributions. This means that robust statistics are an option in this case.

The variances can be either fixed or can be parameters that can be optimized, as specified in the *"Options for standard deviation"* in the **prob** dialog box. Note that the option *"Varies between conditions"* for the standard deviation should produce the same value for ELC2 as obtained for ELC1. The advantage of using ELC2 is that the values of these standard deviations

show up in the **prob** dialog box, and can be visualized using either the *"LL/CR: condition vs standard deviation"* or the *"LL/CR: average vs standard deviation*" plot options.

In case we adopt a constant standard deviation across all conditions, the estimated regression coefficients are equal to

$\beta_{0,EL2} = 0.611$, with 95% CI = [0.539,0.693],

$\beta_{1,EL2} = 0.490$, with 95% CI = [0.483,0.497],

while the standard deviation itself is estimated at

$\sigma_{EL2} = 0.985$, with 95% CI = [0.951,1.105].

This is significantly smaller than the estimate

$\sigma = 1.437$, with 95% CI = [1.333,1.554],

in case of a parametric (Gaussian) regression model. This confirms that the Gaussian regression model is indeed not very reliable for the example data.

The third option entitled *"Empirical likelihood with moments up to order 3 as parameters"* uses the criterium

$$ELC3 = \sum_{k=1}^{K} ELC_k(\mu_k, \sigma_k, \gamma_1, \cdot),$$

i.e., moments up to order 3 are imposed as boundary conditions on the distributions. When selecting this option, an additional parameter entitled *"Skewness"* appears in the **prob** dialog window. It can be either fixed, for instance to its default value of $\gamma_1 = 0$, or optimized as a parameter. Note that ILLMO adopts the same skewness for all conditions. This method can for instance be used to establish a confidence interval for the average skewness of the distributions in different conditions.

The fourth option entitled *"Empirical likelihood with moments up to order 4 as parameters"* uses the criterium

$$ELC4 = \sum_{k=1}^{K} ELC_k(\mu_k, \sigma_k, \gamma_1, \gamma_2),$$

i.e., moments up to order 4 are imposed as boundary conditions on the distributions. When selecting this option, two parameters entitled *"Skewness"* and *"Kurtosis"* appear in the **prob** dialog window. They can be either fixed, for instance to their default values of $\gamma_1 = 0$ and $\gamma_2 = 3$, or optimized as parameters. Note that ILLMO adopts the same skewness and kurtosis values for all conditions.

We have observed that EL regression become increasingly sensitive to the "convex hull" problem as the order of the moments involved increases. Moreover, increasing the order of the moments involved can significantly increase the computation time

required. For the example data, no reliable solution could for instance be obtained for ELC3 and ELC4.