Prof. Dr. Hugo Kubinyi
Donnersbergstrasse 9
D-67256 Weisenheim am Sand
Germany
FAX +49-6353-508233
E-mail kubinyi@t-online.de
URL www.kubinyi.de

November 28, 2005

The Program BILIN

BILIN is a program for IBM and IBM-compatible PCs, under MS-DOS (any version) for the calculation of linear multiple and some nonlinear regression analyses.
The **program does not run properly if called from the Windows surface** – go to the DOS prompt (= Command prompt in newer Windows versions) and start the Bilin program from there.
The program is available in five different languages, i.e. German, English, Spanish, Portuguese and Turkish. This text is available as a file **BILIN.DOC**, in MS-WORD format.
The program creates and reads files ***.REG** and stores results in a file **RESULT.TXT**.

Installation and Setup of the Program

Copy the files **BILIN.EXE**, **BILIN.TXT** and the demo files ***.REG** to a new directory, e.g. **C:\BILIN** . After having performed this step, the installation is complete.

Run the program only from the **DOS prompt (!)**. Within the program, it makes no difference, whether small or capital letters are typed. The program converts all small letters (ASCII codes 097 - 122) to capital letters.

Selection of the Language of the Program

All command and statement lines of the menus of the program are stored in the file **BILIN.TXT**. The letter in the fourth line of this file determines the language of the program:

- **D** rem: German
- **E** rem: English
- **S** rem: Spanish
- **P** rem: Portuguese.
- **T** rem: Turkish

Within the program, the language can be changed by typing either

/ [ENTER]

and then selecting the language by typing the corresponding letter **D, E, S, P,** or **T** and **[ENTER]**, or (better procedure) by directly typing **/D, /E, /S, /P** or **/T** and **[ENTER]**.

If you want to change the language permanently, edit the file **BILIN.TXT** and change the letter in the fourth line to **D, E, S, P** or **T** (take care to use a capital letter).

Start of the Program

After going to the root directory or to the subdirectory BILIN, type

bilin [ENTER].

The program is started and its first page appears at the screen. The first question is for a file name. This can be either the name of a file that already exists or a new file name. Take care to use only letters that are allowed in file names and do not use the extension **.REG**. If you can't remember the name of your file, just type *** [ENTER]** for a list of all files or, e.g., **h* [ENTER]** for all file names with h as the first letter.
After you have typed a file name, e.g.

hald [ENTER]

the next question is whether this file already exists or not. As **HALD.REG** is already present as a demo file, you may now just type

[ENTER].

It is a **special feature if the program** that in most cases the program chooses the most probable answer itself. Thus, if you only type **[ENTER]**, the program assumes that the file already exists. Alternatively you may type

j	[ENTER]	or	(rem German)
y	[ENTER]	or	(rem English and Turkish)
s	[ENTER]		(rem Spanish and Portuguese).

The data are loaded and the data table appears at the screen. If there are too many variables to be shown (e.g. more than 16 compounds = objects, or more than 9 variables = features), the table is shown in parts or the variable values are shown for each compound, separately.

The Data

Any kind of data are accepted by the program, as long as the number of variables and/or data sets is not too large for MS-DOS (in this case a corresponding message is given by the program) and as long as the values of the variables do not contain more than 4 significant decimal places (more than 4 will be truncated by the program) or if individual values are larger than 10^{10} (because of possible rounding errors in the matrix inversion). In those cases your variables should be adjusted, by multiplying them or dividing them by a constant value. The program can convert your data to logarithmic values if you choose the corresponding option (see below); however, **all** values, i.e. the y and all x values are then converted to logarithmic values.

Create a new file

If you have typed, e.g.

TEST1 [ENTER]

as the file name and if you have answered the question whether this file already exists, with

n [ENTER],

a new file will be created. You should take care not to duplicate file names because in such cases your old file will be destroyed, without any warning.
The menu asks for the input of the number of X variables, the number of objects, whether the variables must be converted to logarithms (take care to type **n [ENTER]**, if you dont want a conversion to logarithmic values or if you already have logarithmic values), the name of the Y variable (usually the biological activity), the names of the individual X variables and the names of the compounds (= objects). Here you may just type

[ENTER]

for the first compound name which attributes the "names" 001, 002, 003 ... nnn to the objects.
Then the data are requested. If you realize that you typed a wrong value, dont worry. It is possible to correct such values afterwards, either with the help of the program (in the case of only one or very few errors) or, more convenient, especially in the case of many errors, by editing the corresponding ***.REG** file and correcting the values with the help of the editor.
After you have finished by typing the last X_{ij} value, the program asks whether you want to create a file. Type

[ENTER]

and a file with the name selected by you will be created. The data are now saved in the file **TEST1.REG** and the data are displayed at the screen.
Only type **n [ENTER]** if you REALLY do not want to create a file. In this case the data will not be stored. After having finished the program, your data are lost.

Quadratic Terms, Interaction Terms, and Similarity Matrices

==> This option should be used only by experienced QSAR scientists; strange models may result by inappropriate application of this tool.

If only a small number of variables is included in the data set, this menu is displayed. **In most cases you should skip this option by just pressing [ENTER]**. However, you may automatically create quadratic terms X_1^2, X_2^2, ..., by selecting

q [ENTER],

interaction terms, e.g. $X_1 \cdot X_2$, $X_1 \cdot X_3$, ..., by selecting the option

i [ENTER],

and all these terms (i.e., quadratic and interaction terms) by selecting the option

a [ENTER]

and the calculation of an n x n $|X_{1i} - X_{1j}|$ similarity matrix by selecting the option

s [ENTER].

If any of these options is generated, the corresponding variables are calculated by the program, but they are not stored in a file; they are only available within the correponding run of the program. If the total number of variables, including the newly generated ones, is larger than n - 1, the program continues with the main menu, otherwise the regression equation including all the selected variables is calculated.
You should select variables with the help of option 5 (in the main menu, see below) or look for best models, using the options 2, 3, 4 or 6, or select the options ss or cv (in the main menu, see below), because it cannot be expected that all quadratic,

and/or interaction terms or all terms of the similarity matrix are significant at the 95% level (see below).

Correction of Data

After the presentation of the data in a table, the program asks whether you want to start the calculation. If you type

n [ENTER]

the menu asks you whether you want to correct values. Just press

[ENTER]

and follow the menu. After you have corrected your value, the program asks whether you want to save the corrected value. Just press **[ENTER]**, once again.
An alternative is to edit your ***.REG** file with the help of an editor and to do the corrections with the help of the editor commands. Afterwards, you must save your corrected file.

Elimination of Data Sets

Sometimes you may want to eliminate data sets only temporarily from your calculations, e.g., outliers or compounds structurally not related to the others. If you have answered the questions, whether you want to correct values, by

n [ENTER]

the program asks whether you want to eliminate data sets. Just press **[ENTER]** and follow the menu (compare option 7, see later). You may eliminate any number of objects but you should take care that the number of objects (= compounds) remains larger than the number of X variables, otherwise regression analysis will fail (for the case of more variables than objects, see later).
Within the program, the temporary elimination of data sets is possible by selecting option **7** from the main menu (see later).

Graphical Presentation of the Data

If the question for graphical presentation of the data is answered by

[ENTER],

a menu appears which allows to select any variable as the Y variable and any other variable as the X variable of a two-dimensional diagram, together with the linear regression of this pair of variables.

0 [ENTER] or just

[ENTER]

selects the Y variable, then any number (**1, 2, ... i**) and **[ENTER]** selects the corresponding X_1, X_2 ... X_i variable.

n [ENTER]

leaves the graphics and continues the program.
Within the graphics, the commands **n, nu, na,** and **su** are allowed.

[ENTER]

leads to a new menu for a graphical presentation and any other combination of variables can be selected.

n [ENTER]

leaves the graphics and continues the program by calculating the regression equation or by going to the main menu (only in certain cases, see below).

nu [ENTER]

automatically adds the numbers of the objects to the dots.

na [ENTER]

labels the dots with the names of the objects (= compounds). The labelling is done, one by one. Each time a new label appears, the key combination **na [ENTER]** must be repeated n times.

su [ENTER]

ask for a number or a name. If the **number of the object** and **[ENTER]** is typed, the corresponding dot starts blinking. The same happens, if the **name of the object** and **[ENTER]** is typed.

The Regression Equation

After you have left the graphics mode or if you have selected **n [ENTER]** to skip this option, a regression equation, including all variables is calculated (only if k, the number of variables is smaller than n - 1; n = number of data sets). The regression coefficients of all selected terms and their 95% confidence intervals are given (not standard deviations). A warning is given if any of these coefficients if not significant at the 95% confidence level. In addition, n, r (the correlation coefficient), s (the standard deviation of the y deviations) and F (the Fisher significance value) are given.
For running a cross-validation, see below.

Table and Graphics of y Deviations

After the calculation of a regression model, a table of observed and calculated y values and (y_{obs} - y_{calc}) values can be selected, by typing

t [ENTER].

If the option

g [ENTER]

is selected, a graphics of the y deviations is shown. The scaling corresponds to the single and double standard deviations, s and 2s (about 67% and 95% of the deviations, respectively). This graphics is produced only for examples where less than 73 objects (= compounds) are included in the calculation.

Cross-Validation

Cross-validation is a method to check the validity of regression models, e.g. by eliminating every object once and only once (= leave-one-out cross-validation, LOO) and deriving the models only from the objects remaining in the analysis. Thus there are n (n = number of objects) runs, leading to a prediction of each object outside the analysis. The squared error of predictions PRESS is used to calulate Q^2 and s_{PRESS} values.
Cross-validation can be activated after the calculation of any regression model (i.e., the first calculation, including all variables, or after selecting option 5 of the main menu) by typing

cv [ENTER].

The regression model remains at the screen and after finishing the calculation, Q^2 and s_{PRESS} values are added to this result. Please consider that the time needed to perform cross-validation is n-times larger than the time needed for the calculation of the regression model. Thus, for large data sets, the cv option may take a long time. So far, no cross-validation is implemented for bilinear models.

General Commands

At any time during execution of the program, the commands

e [ENTER] or
p [ENTER] or
$/, /D, /E, /S, /P or **/T [ENTER]**

may be chosen; **e** and **[ENTER]** closes all files, interrupts and ends the program, **p [ENTER]** writes the content of the screen to the file **RESULT.TXT**. If this file is not already open, it is opened. An already existing file **RESULT.TXT** (from a prior analysis) is destroyed at this time (see below, option 0 in the main menu). If this file is already open and contains information, the current screen information is added to the content of this file.
This latter option offers the opportunity to write all intermediate results to the file **RESULT.TXT**.

/, /D, /E, /S, /P or **/T [ENTER]**

allow to change the language of the program (see above).

The Main Menu

After the completion of a calculation, **[ENTER]** leads to the main menu of the program. Within this menu, the options

1 Intercorrelation Matrix
2 All One-Variable Regression Models
3 All Two-Variable Regression Models
4 All Three-Variable Regression Models
5 Selection of Variables
6 Systematic Search, up to 4 Variables
7 Original Table / Elimination of Data
8 Parabolic Model
9 Bilinear Model
0 New Start of the Program
e End of the Program

are available. In addition, at this stage, the options

ss, ss1, cv, or cv1

can be selected (systematic search and systematic search, including cross-validation, see below).
Any number between **0** and **9** and **[ENTER]** leads to the corresponding option, which all are described below.

Option 0: New Start of the Program

0 [ENTER] or just [ENTER]

leads to a new start of the program. All data of the previous run are lost, with the exception of the file **RESULT.TXT**. However, you should consider that also the file **RESULT.TXT** of the previous run is destroyed, if you select

p [ENTER]

the next time. Thus, option

e [ENTER]

should be used instead of option **0**, if you want to edit or to print the file **RESULT.TXT** after an analysis. Afterwards you start the program **BILIN**, by typing

bilin [ENTER].

Option 1: Intercorrelation Matrix

The intercorrelation matrix of X variables can be selected by going to the main menu and typing

1 [ENTER].

The r^2 values (not r values; r = correlation coefficient) of all X_iX_j correlations are shown in a table. If the number of X variables is larger than 10, the program requests an r value limit. After typing this value and pressing [ENTER], only those r^2 values are given which are larger than the limit.

Option 2: All One-Variable Regression Models

All one-variable regression models are calculated if from the main menu option

2 [ENTER]

is selected. r, s, and F values of the corresponding regression equations are given. If more than 20 variables are included in the data set, the program ask for a r value limit. After typing this value and pressing **[ENTER]**, only those models and r values are given which are larger than the limit.

Option 3: All Two-Variable Regression Models

All two-variable regression models are calculated if from the main menu option

3 [ENTER]

is selected. r, s, and F values of the corresponding regression equations are given. If too many variables are included in the data set, the program ask for a r value limit. After typing this value and pressing **[ENTER]**, only those r values are given which are larger than the limit. The results are sorted according to the fit of the models.

The program asks you whether you want to select variables from a certain number of models. If you want to continue with a file which contains only the most relevant variables (e.g. selected variables from the Selwood data set, see below), you may type now 20 or a smaller number. Then the program requests a new file-name (without the REG extension) and stores the newly created data set under this name.

Option 4: All Three-Variable Regression Models

All three-variable regression models are calculated if from the main menu option

4 [ENTER]

is selected. r, s, and F values of the corresponding regression equations are given. If too many variables are included in the data set, the program ask for a r value limit.

After typing this value and pressing **[ENTER]**, only those r values are given which are larger than the limit. The results are sorted according to the fit of the models.

Option 5: Selection of Variables

If option

5 [ENTER]

is selected from the main menu, any variables can be selected to calculate a regression model. A menu appears, which asks how many X variables shall be included in the equation. If this value is given, e.g. 2, the program asks for the code numbers of these variables, e.g. 3 for X_3 and 4 for X_4.
If, instead,

-1 [ENTER]

is the answer to this question, regression models without a constant term are calculated (such a model may be appropriate, e.g. in the calculation of substance concentrations from multicomponent spectra) However, in most other cases such models are inappropriate. The question how many X variables shall be included is repeated.
A repetition of **-1 [ENTER]** returns to regression models that include a constant term.

0 [ENTER]

as the answer to the question, how many X variables shall be included, allows to select another variable, i.e. one of the X variables, as the Y variable of the regression equation. The program asks for the code number of this variable (0 selects the Y variable, 1, 2 .. i selects the X_1, X_2, ... X_i variable). The total number of X variables and the code numbers of these variables are then selected as above.
If you select the original number of (all) variables, the program will show the equation including all variables (as long as the number of objects is larger than the number of variables + 1).

Option 6: Systematic Search for all Regression Models with up to 4 Variables

6 [ENTER]

is an option which searches all possible one-, two-, three- and four-variable models. In the case of many variables this may take a long time. Nevertheless, for a data set with about 30 objects and 30 variables, it will take only a few seconds at a Pentium PC. The 317 682 (!) different models with up to four variables of the Selwood data set, with its 31 objects and 53 variables, are calculated and sorted in less than 75 seconds on a Pentium III, 500 MHz PC, i.e. with a speed of about 4 500 matrix inversions per second! As result, the "best" equations are presented at the screen. You may select variables from these models, as described in option 3.

Option 7: Original Table / Elimination of Data

7 [ENTER]

repeats the data table. If the question for elimination of data sets is answered by

[ENTER],

data sets can be temporarily excluded from the calculation (see above).
Follow the menu. The program first asks how many compounds shall be eliminated and then which ones (you have to type each single number). There is a shortcut for

the elimination of compounds: If you want to eliminate several objects in a row, e.g. compounds 12-21 from the Selwood data set (see below), first type 10 (beause these are 10 objects) and then 12- (which tells the computer that you want to eliminate object 12 and the following one, in total 10 objects.
If you have given a number, how many objects (= compounds) to eliminate and type then “r” or “R” (for “random”) instead of a number of the first data set, the program will randomly eliminate as many compounds as you have requested; e.g. typing “10” for the number of compounds to be eliminated, followed by [ENTER], and then “r”, followed by [ENTER], will randomly eliminate 10 data sets.
If the question for elimination of data sets is answered by

n [ENTER],

all objects remain in the data set; objects eliminated at an earlier stage are re-introduced into the calculation.

Option 8: Parabolic Model

8 [ENTER]

leads to a parabolic model. If more than one X variable is in the data set, a menu asks for the code number of the variable which shall be considered with a linear and a parabolic term in the regression model. Then the program asks how many additional variables shall be included as additional linear terms (if several or all variables shall be considered as linear and parabolic terms, the option **q** at the begin of the program must be selected, see above). This number is given and then the program asks for the code numbers of the corresponding X variables (see option **5**).

Option 9: Bilinear Model

9 [ENTER]

leads to a menu which calculates a bilinear equation, e.g.

log 1/C = a log P - b log(ßP + 1) + ... + c

Only one bilinear term can be included in the model, but several other variables can be added to the regression model. In this case proceed as described under option **8** (see above). **No cross-validation** is available for the bilinear model.

Systematic Search for Regression Models

Sometimes, especially if larger numbers of X variables are considered, the search for the "best" regression model "by hand" may be very time-consuming. In such cases, the option

ss [ENTER]

can be chosen, directly at the start of an analysis, instead of the file name (in this case the question for the file name is repeated by the program and only then the file name is typed), or at the main menu, instead of typing 1, 2, ... e (see above).
After answering the usual questions, whether to start the calculation and whether to go to the graphics option, the program asks for the maximum number of X variables in the regression models. This can be either the total number of X variables included in the data set or a smaller number. If a number of variables is chosen that is larger than n - 2 (n = number of objects), the program automatically corrects this value. Next the program asks for a maximum F value. If you type a value, e.g.

100 [ENTER],

only regression models with F values > 100 will be stored. Otherwise all models are stored. In choosing an F value you should take care not to select a value which is too large (in the worst case larger than the F value of the "best" model) or, in the case of many models, a too small value which will store too many models.
In the case of many variables, too many different models will lead to much too long calculation times (for the Selwood data set with its 53 X variables, there are 23,426 different three-variable regression models, 22,957,480 six-variable models and, in total, $7 \cdot 10^{15}$ different models, with up to 29 variables). Therefore the program asks whether the calculation shall really be started. The estimated calculation time refers to a 486 PC, 33 MHz (longer or shorter calculation times will be needed if your PC is slower or faster).
If only models with a certain number of variables shall be calculated, the option

ss1 [ENTER]

is chosen instead of the option **ss**. Afterwards the number of variables is selected, e.g., 3 for all three-variable models.
All results are stored in a file **RESULT.LST**. After calcatulion of the last model, the program automatically sorts the best results, writes these results to the file **RESULT.TXT** (which is newly created at this occasion, thus all prior information in this file is lost) and ends its operation.

Systematic Search, Including Cross-Validation

Cross-validation can be activated either after the calculation of a regression equation (e.g. options **5**, **6** or **8** in the main menu; for option **9**, the bilinear model, the cross-validation option is not yet available), instead of choosing the table of deviations or the graphics of devations. Cross-validation can also be activated in the systematic search procedure, by typing

cv [ENTER]

instead of the option **ss** (see above, chapter systematic search), either at the begin of the program or in selecting this option at the main menu, instead of **1, 2, ... 9, 0, e**. Then the usual questions and finally the question for the maximum number of variables is answered. If only models with a certain number of variables shall be calulated, the option

cv1 [PRESS]

is selected instead of the option **cv**.

More Variables than Objects

Sometimes a data set may contain more X variables than objects (e.g. the file SELWOOD.REG, 31 objects and 53 X variables). In such cases, regression analysis can be applied only to a subset of the X variables. After the question, whether you want to start the calculation, the program goes directly to the main menu. You may choose any option. Option **5** gives you the opportunity to select any combination of variables. Option **6** gives you the best 4-variable models.

Errors, e.g. from Intercorrelated Variables

The program may stop with an error message. In this case the run should be interrupted by typing **e** and **[ENTER]**, or in the worst case

[CTRL] + [BREAK] or **[CRTL] + [ALT] + [DELETE]**.

Try a new start of the program. If the error occurs again, check your data. Most often errors result from **singluar matrices**, if two or several variables are inter-correlated. In these cases you may eliminate one of these variables from the original data set or just continue with the calculations. The program will always come up with an error message, if you try to combine such variables in one regression model but you can, nevertheless, continue with other options.
If you need further help, send me your data set, either by e-mail or on a diskette and I will try to solve your problem.

Demo Files

The demo files **HALD.REG, TIMM1.REG, DHFR-INH.REG** and **SELWOOD.REG** are at your diskette.
For the **HALD** file [1], try to find the "best" equation. You should end up with the variable combination X_1 and X_2.
For the **TIMM1** data [2], calculate the parabolic and the bilinear model, using X_1 as the nonlinear term and X_2 as additional variable.
DHFR-INH [3] is a data set for Free-Wilson analysis. Find the "best" model by using the **ss** option (results are not shown at the screen but are contained in the file **RESULT.TXT**, that is automatically generated).
The **SELWOOD** data set [4] has been described in detail. Search for the best three-variable model. If you have a very fast computer (e.g. a 486 or a Pentium), you may use the **ss** option or even the **cv** option (select F > 20) to find the best 5-variable models (over night!).

Have fun with my program!

References

a) QSAR, Hansch and Free-Wilson Analyses

Ramsden, C. A., Ed., *Quantitative Drug Design*, Vol. 4 of *Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*, Hansch, C., Sammes, P. G., and Taylor, J. B., Eds., Pergamon Press, Oxford, 1990.

Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches (Methods and Principles in Medicinal Chemistry,* Volume 1, Mannhold, R., Krogsgaard-Larsen, P., and Timmerman, H., Eds.), VCH, Weinheim, 1993.

b) Data Sets

[1] Draper, N. R., and Smith, H., *Applied Regression Analysis*, 2nd ed., Wiley, New York, 1981.

[2] a) Timmermans, P. B. M. W. M., de Jonge, A., van Meel, J. C. A., Slothorst-Grisdijk, F. P., Lam, E., and van Zwieten, P. A., *J. Med. Chem.* **24**, 502 - 507 (1981).
b) Timmermans, P. B. M. W. M., de Jonge, A., Thoolen, M. J. M. C., Wilffert, B., Batink, H., and van Zwieten, P. A., *J. Med. Chem.* **27**, 495 - 503 (1984).
c) Kubinyi, H., *Chemie in unserer Zeit* **20**, 191 - 202 (1986).
d) Kubinyi, H., *Physical Property Prediction in Organic Chemistry*, Jochum, C., Hicks, M. G., and Sunkel, J., eds., Springer-Verlag Berlin, 1988, p. 235 – 247.

[3] Hansch, C., Silipo, C., and Steller, E. E., *J. Pharm. Sci.* **64**, 1186 - 1191 (1975).

[4] a) Selwood, D. L., Livingstone, D. J., Comley, J. C. W., O'Dowd, A. B., Hudson, A. T., Jackson, P., Jandu, K. S., Rose, V. S., and Stables, J. N., *J. Med. Chem*. **33**, 136 - 142 (1990).
b) Wikel, J., and Dow, E., *Bioorg. Med. Chem. Lett*. **3**, 645 - 651 (1993).
c) McFarland, J. W., and Gans, D. J., *Quant. Struct.-Act. Relat*. **13**, 11 - 17 (1994).
d) Rogers, D., and Hopfinger, A. J., *J. Chem. Inf. Comput. Sci*. **34**, 854 - 866 (1994).
e) Kubinyi, H., *Quant. Struct.-Act. Relat*. **13**, 285 - 294 (1994); Kubinyi, H., *Quant. Struct.-Act. Relat*. **13**, 393 - 401 (1994).