The R-Function `regr` for an Augmented Regression Analysis (Draft)

Werner A. Stahel, ETH Zurich

November 12, 2010

Abstract

The R function `regr` is a wrapper function that allows for fitting several different types of regression models by a unified call, and provides more informative numerical and graphical output than the traditional `summary` and `plot` methods. The package `regr0` contains the functions that go along with `regr` and a number of others. It is written to make data analysis more effective by providing user-oriented, flexible functions. It is available from `R-forge` and is still in development.

1 Introduction

Regression models are fitted in the statistical programming environment R by diverse functions, depending on the particular type of model. Outputs obtained by calling the function `summary` on the object produced by the fitting function look often quite similar. Graphical output for residual analysis is obtained by calling `plot`, but the result is not always informative.

The function `regr` allows for fitting various regression models with the same statement, provides a more informative numerical output and enhanced plots for residual analysis.

`regr` proceeds by checking arguments, then calling the suitable fitting method from standard R or other packages, collecting useful statistics from the resulting object and a call of `summary` on it and adding a few to generate an object of class `regr`.

In particular, the following models can be fitted by `regr`:

- ordinary linear models, using Least Squares or robust estimation, by calling `lm` or `rlm`,
- generalized linear models, by calling `glm`,
- multinomial response models, by calling `multinom` of package `nnet`,
- ordered response models, by calling `polr` of package `MARS`,
- models for survival data and Tobit regression, by calling `survreg` of package `survival`,
- multivariate ordinary linear models, by calling `lm`

This document presents the main features of the package `regr0` and explains the ideas behind them. It gives no details about the functions. They can be found in the help files.

The package is available from `R-forge`, e.g. by calling
`install.packages("regr0", repos="http://r-forge.r-project.org", lib=...)`.
The reason why it is not on CRAN and called `regr0` rather than `regr` is that the author is still developing additional features and does not yet want to guarantee upward compatibility. It also means that comments and suggestions are very welcome: `stahelstat.math.ethz.ch`

2 Numerical Output

The useful numerical output of fitting functions is usually obtained by calling `summary` on the object produced by the fitting method. This results, for most functions, in a table showing the estimated regression coefficients, their standard errors, the value of a test statistic (t or z or deviance) and, for the ordinary linear model, a p value for the tests for zero coefficients. It is followed by an overall summary, usually including a test for the hypothesis that all coefficients are zero, and a standard deviation of the error and coefficient of determination, if applicable.

If there are factors (qualitative explanatory variables) in the model, the coefficients are not always interpreted adequately, and the respective standard errors, t and p values are (almost) useless and often misinterpreted. On the other hand, the information whether a factor has a significant effect is not available from the summary but has to be obtained by calling `drop1` on the fit object. (The function `anova`, which seems to be suited according to its name, usually does not answer this question.)

This situation cannot be called user friendly. The function `regr` is meant to provide the results that are needed without having the user call different functions and select the output that is safe to be interpreted.

2.1 Standard output for continuous explanatory variables

Here is a result of printing a `regr` object.

```
Call:
regr(formula = log10(tremor) ~ location + log10(distance) + log10(charge),
    data = d.blast)
Fitting function  lm

Terms:
                  coef stcoef signif  R2.x df p.value
(Intercept)      2.964  0.000   13.6    NA  1       0
location            NA     NA   10.5 0.052  7       0
log10(distance) -1.518 -0.788  -12.0 0.277  1       0
log10(charge)    0.636  0.410    8.2 0.053  1       0

Coefficients for factors:
$location
    loc1     loc2     loc3     loc4     loc5     loc6     loc7     loc8
 0.00000  0.15306  0.13169 -0.16185 -0.03211  0.07161 -0.00889  0.00372

St.dev.error:  0.143   on 352 degrees of freedom
Multiple R^2:  0.795   Adjusted R-squared: 0.79
F-statistic:    152   on 9 and 352 d.f.,  p.value:    0
```

The "Terms:" table characterizes the effects of the individual terms in the model. For continuous explanatory variables (the last 2 lines in the example) it shows:

`coef`, the estimated value of the coefficient;

`stcoef`, the estimated standardized coefficient, defined as `coef` times the standard deviation of the explanatory variable, divided by the standard deviation of the response (if the response is continuous as assumed here), see below for its use;

`signif`, a significance measure that is > 1 for estimated coefficients differing significantly from 0, see below for its definition;

`R2.x`, the coefficient of determination for regressing the explanatory variable in question on the other terms in the model. This is one of the wellknown collinearity diagnostics.

`df`, degrees of freedom, always 1 for such variables;

`p.value`, the p value for testing if the coefficient could be zero.

2.2 Standardized Coefficients

The standardized coefficients are meant to allow for a comparison of the importance of explanatory variables that have different variances. Each of them shows the effect on the response of increasing "its" carrier $X^{(j)}$ by one standard deviation, as a multiple of the response's standard deviation. This is often a more meaningful comparison of the relevance of the input variables.

Note, however, that increasing one $X^{(j)}$ without also changing others may not be possible in a given application, and therefore, interpretation of coefficients can always be tricky. Furthermore, for binary input variables, increasing the variable by one standard deviation is impossible, since an increase can only occur from 0 to 1, and therefore, the standardized coeffient is somewhat counter-intuitive in this case.

2.3 Significance

The usual `summary` output of fitting functions includes the t values of the coefficients as a column in the coefficients table. They are simply the ratios of the two preceding columns. Nevertheless, they provide a kind of strength of the significance of the coefficients. The p value may also serve as such a measure, but it is less intuitive as it turns tiny for important variables, making comparisons somewhat more difficult than t values. The significance of t values depends on the degrees of freedom, but informed users will know that critical values are around 2, and they will therefore informally compare t values to 2. Based on these considerations, we introduce a new measure of significance here.

The new significance measure is defined as

$$\texttt{signif} = \texttt{t value} \,/\, \texttt{critical value}$$

where `critical value` is the critical value $q_{df}$of the t distribution and depends on the degrees of freedom. The definition is applicable for continuous explanatory variables as well as for binary factors. For other factors, we will extend this definition below.

2.4 Confidence Intervals.

The standard errors provided by the usual `summary` tables allow for calculating confidence intervals for continuous explanatory variables, by the formula $\texttt{coef} \pm q_{df} \cdot \texttt{std.error}$. The formula based on `signif` is

$$\texttt{coef} \cdot (1 \pm 1/\texttt{signif})$$

Numerically, this is slightly more complicated, but there is an additional interpretation of `signif` in terms of the confidence interval: If the input variable were scaled such that the confidence interval had half width 1, then the estimate would be `signif` units away from zero.

2.5 Factors

For factors with more than two levels, (`location` in the example), there are several coefficients to be estimated. Their values depend on the scheme for generating the dummy variables characterizing the factor, which is determined by the `contrasts` option (or argument) in use. Whereas the usual option `contrasts="contr.treatment"` gives coefficients with a clear interpretation (difference of

level k to level 1 for $k > 1$), other contrasts may be mixed up with them or be very difficult to interpret. In the `regr` output, the estimated coefficients for all levels of the factor – independently of the coding except for an additive constant – are shown after the Terms table, but tests for them are not provided.

Note that for factors with only two levels, the problem does not arise, since the single coefficient can be interpreted in the straightforward manner as for continuous explanatory variables. `regr` therefore treats binary factors in the same way as continuous explanatory variables.

The test performed for factors with more than two levels, which is shown in the `Terms` table by the `p.value` entry, is the F test for the whole factor (hypothesis: all coefficients are 0). It is obtained by calling `drop1`. The significance measure is defined as

$$\texttt{signif} = \sqrt{\texttt{F value} \,/\, q_{df1,df2}}$$

where $q_{df1,df2}$ is the critical value of the F distribution. It reduces to the former one for binary factors.

The collinearity measure `R2.x` for factors is a formal generalization of `R2.x` for terms with one degree of freedom, determined by applying the relationship with the "variance inflation factor", `R2.x`$= 1/(1 - \text{vif})$ to the generalized vif. [More explanation planned.]

2.6 Model summary

The last paragraph of the output gives the summary statistics. For ordinary linear models, the estimated standard deviation or the error term is given first. (It is labelled "Standard error of residual" in the `lm` output, which we would label a misnomer.) The `Multiple R^2` is given next, together with its "adjusted" version, followed by the overall F test for the model.

For generalized linear models, the deviance test for the model is given. If applicable, a test for overdispersion based on residual deviance is also added.

2.7 Model Comparisons

When model development is part of the statistical analysis, it is useful to compare the terms that occur in different models under consideration. There is a function called `modelTable` collects coefficients, p values, and other useful information, and a `format` and `print` method for showing the information in a useful way.

2.8 Model Selection

The well-known functions `drop1` and `add1` are adapted by providing additional methods. Whereas the terms (`scope`) for which `drop1` tests if they can be dropped from the model has a nice default in the standard R, such a default is not provided for `add1`. This latter function is very useful to check whether squared continuous variables or interactions between variables should be included into the model. Therefore, our version of `add1` provides this default `scope`.

Since `drop1` and `add1` methods are available, `step` can be used to select a model automatically. ... AIC ...

3 Residual Analysis

The residual plots that are produced by plotting an `lm` or `glm` object are:

- The Tukey-Anscombe plot showing residuals against fitted values, which is certainly the most important single diagnostic tool for assessing model assumptions;

- The normal quantile-quantile plot, which makes sense only for ordinary linear models and in some cases for generalized linear models. It is not essential since skewness and outliers can also be seen in the Tukey-Anscombe plot if they are relevant;
- The Scale plot, which shows square-root transformed absolute values of residuals against fitted values and helps to spot unequal variances (if these depend on the expected value of the response).
- The leverage plot, displaying residuals against leverage values. This plot is useful to detect influential observations.

3.1 What `plot.regr` does

The plotting method for `regr` objects has many addtional features, to be described in more detail in the following subsections.

- The plots are augmented by a smooth fitted to them (which is also available in the classical R functions), and simulated smooths are added in order to assess an informal "significance" of curvature in the smooth of the data. In addition, a reference line is given, which helps finding an appropriate modification of the model if significant curvature is found.
- The set of plots that is shown by default is adjusted to the type of model. A normal QQ-plot of residuals is only shown if appropriate. On the other hand, residuals are plotted against the index (sequence number) of the observation, since this may show trends or correlations of errors in time, if the sequence reflects time. If weights are used, the residuals divided by the weight are also plotted against the weights, which helps to see if the weighting rule was adequate.
- Most importantly, residuals are plotted against explanatory variables. This can also be achieved by calling `termplot` for other fit objects, but experience shows that this is often neglected by average users.
- Finally, plotting methods are defined for models for which no useful methods are available in the basic R packages, notably ordered response regression and censored responses.

The number of plots produced by calling `plot` on a `regr` object may be elevated. The argument `plotselect` allows for requiring or deselecting any of them. The arguments to `plot.regr` are numerous and allow for varying many features, including those for which the default behavior is discussed below.

The plotting pages are marked in the right lower corner by the date and a project title and step label, if available from `options`. (This `stamp` can be suppressed by setting `options(stamp=FALSE)`.)

The plots can be generated by executing the examples on `help('plot.regr')`. A single plot is shown here for easy reference in the following descriptions.

3.2 Extreme Residuals

If there are one or a few outliers in any plot, the structure of the majority of the data becomes more difficult to grasp. Therefore, the `plot` method first flags outliers in the residuals. If there are any, the range of the vertical axis is split into an "ordinary plotting range", where the non-outliers are shown as usual, and an "outlier margin", where the outliers are shown on a highly nonlinear scale that maps the range from the limit of the ordinary range to infinity to this margin.

In order to ease identification, a suitable number of most extreme residuals are labeled with their `row.names` by default.

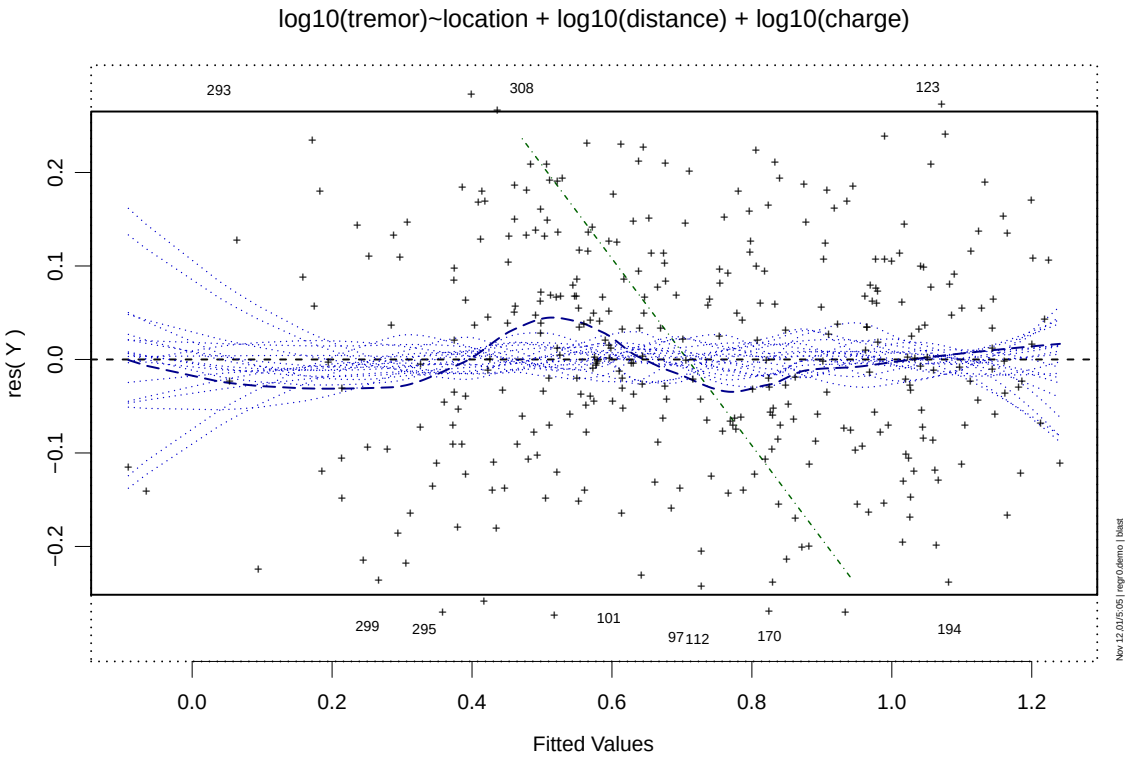

Figure 1: A Tukey-Anscombe plot.

3.3 Smooths.

Fitting a smooth (blue dashed line in the Figure) to a residual scatterplot helps to spot a potential nonlinearity of the relationship between the response and the variable shown in the plot – the fitted values or a single explanatory variable. The smooth used by default is `loess` with a span depending on the number of observations (currently $= 3n^{-0.3}$ for ordinary regression, and twice this number for generalized linear regression; do not ask why). Such a smooth is more essential for generalized linear models than for ordinary ones, since artefacts of the residuals may make it impossible to see curved relationships from a display of the residuals.

It is difficult to decide if a smooth line is curved "significantly", such that searching for a better model should be helpful and not lead to overfitting the data. In order to help judging the "natural curvature", 19 sets of data are generated by drawing random response values according to the fitted model. The smooths obtained for these sets are also shown (in light blue in the Figure – I hope that this can be seen in all versions of this document). If the smooth determined from the data is clearly the most extreme in some aspect, one may conclude that the model does not fit adequately. The number 19 is chosen to correspond to a "significance level" of 5%. Up to now, such simulated datasets are only generated for ordinary linear regression.

3.4 Augmented Tukey-Anscombe Plot

The plot of residuals against fitted values is the most important diagnostic tool to assesss the validity of model assumptions. It allows for detecting deviations from linearity, homoscedasticity, symmetry and normal distribution of the random deviations.

If there is curvature, one remedy may be to transform the response variable. This can help only if the relation between fitted and response values is monotone – otherwise, quadratic terms are the most straightforward modification of the model. The distinction is easy to see if the response

is plotted on the vertical axis instead of the residuals. This display can also be requested from `plot.regr` (by setting `plotselect=c(yfit=3)`).

In order to avoid the necessity to either call `plot` again or to ask for the plot of response against fitted values routinely, a **reference line** (green dot-dashed line in the Figure) is shown in the Tukey-Anscombe plot. It connects points with equal response values. Since the response can be determined by adding the two coordinates of the plot – fitted plus residual – lines of equal response values have slope -1. The reference line shown is such a line, the one going through the center of the points in the plot. If the smooth never gets steeper than the reference line, then the suggested relation between fitted and response values is monotone, and a transformation of the response may remove the non-linearity.

One might argue that a plot of the response on the fit is more straightforwardly interpreted than the Tukey-Anscombe plot with the reference line. We think that the Tukey-Anscombe plot can show the deviations from model assumptions more clearly and should therefore be preferred, and that the reference line, after a short learning period, will be easy enough to work with and helps avoiding an additional display. Of course, some users may disagree.

3.5 Scale Plot

The absolute values of the residuals are plotted against the fitted values in order to see more clearly than in the Tukey-Anscombe plot whether the scatter of residuals depends on the expected values. The plot shown when plotting `lm` objects uses square roots of absolute residuals instead of the absolute values themselves because they are more symmetrically distributed. This appears as a technical argument to me and it may confuse unexperienced users. While square roots are therefore avoided for plotting, they are used to calculate the smooth that is added to the plot.

3.6 Weights

If weights are given by the user, they should usually reflect different variances of the random errors E_i, and the residuals, divided by the square root of the weights, should have constant variance. The standardized absolute residuals are therefore plotted against the weights. If the size of these residuals is not constant but depends on the weight, the rule or weighting function should be revised.

If weights are given, they are also used as the sizes of the plotting symbols (circles) in other plots.

3.7 Leverage Plot

Influential observations can be spotted in the scatterplot of residuals against leverage values (diagonal elements of the projection matrix). The well-know diagnostic called Cook's distance is a simple function of these two quantities, and level curves are drawn in the plot.

3.8 Sequence Plot

Residuals are plotted against the sequence number in the dataset in order to show any trends and correlation that may be related to this sequence.

3.9 Residuals against explanatory variables

Plotting residuals against explanatory variables serves to detect nonlinearities of the relation between the response and the variable in question. For ordinary regression and some other models, they can show dependencies of the scale of errors on the explanatory variables.

The plots are enhanced by smooths as discussed for the Tukey-Anscombe plot. Also, a reference line is added that helps to decide whether a transformation of the explanatory variable may help to remove any non-linearity shown by the smooth. Again, the reference line intends to connect points of equal response values. Note, however, that they do not really align on a straight line because of the contributions of the other terms in the model to the fit. Therefore, the reference line connects points for which the sum of "component effect" $\widehat{\beta}_j x^{(j)}$ and the residual is equal.

Alternatively, these sums may be used as the vertical axis instead of the residuals for plotting. This display is usually called the "partial residual plot" and can be obtained from `plot.regr`, too. !!! It is related to the plots shown by default in the same way as the response versus fitted plot is to the Tukey-Anscombe plot.

When transformed explanatory variables appear in the model, the residuals are plotted against the original values. The reason is as follows: If any curvature is detected, an improvement of the model may be possible by transforming the variable shown. It is then more natural to search for an enhancement of the original transformation rather than a transformation of transformed values, and this should be easier if the original scale is used. If this appears to be unsuitable in a given application, the user may store the transformed variable under a new name and then use this new variable in the model formula.

3.10 Residuals for an ordinal or binary response

Ordinal regression models are best specified by introducing the idea of a "latent response variable" Z, which is continuous and follows a linear relationship with the explanatory variables. The response Y is considered to be a classification of Z into k classes of possibly unequal width. The most well-known model specifies a logistic distribution for the error term in the linear model for Z. Then, the coefficients and the thresholds or class limits are estimated by maximum likelihood. This is implemented in the function `polr` (proportional odds linear regression) of the `MASS` package, which is the function invoked by `regr` for ordinal regression.

Resdiuals for this model may be defined by considering the conditional distribution of the latent variable Z, given the observation Y and the explanatory variables. This is a logistic distribution with parameters given by the model, restricted to the class of Z given by the observed Y. Residuals are defined as the median of this distribution, and it may help to characterize the uncertainty about the latent residuals by the quartiles of the conditional distribution. The plots show the interquartile range by a vertical line and the median, by a horizontal tick on them.

Note that this definition may not have appeared in the literature yet.

3.11 Residuals for censored responses

If response values are censored, so are the residuals. In the same spirit as for the case of ordered responses, it is straightforward to calculate a conditional distribution, given the fitted value and the censoring value, for each residual. This can again be plotted by showing quartiles.

3.12 Residual plots for multivariate regression.

For multivariate regression, the plots corresponding to each target variable should be examined. They are arranged such that the same type of plot for the different response variables are grouped together, and a matrix of plots is produced for residuals against explanatory variables. In addition, the qq-plot of Mahalanobis norms of residuals is shown as well as the scatterplot matrix of residuals.

4 Further Plotting Functions

4.1 Scatterplot Matrices

Scatterplot matrices are produced by the `pairs` function. In `regr0`, there is a function with more flexibility, called `plmatrix`.

- The set of variables shown horizontally and vertically need not be the same. `plmatrix` can be used to show the dependence of several response variables (vertical axes) on several explanatory ones (horizontal axes).
- A traditional square scatterplot matrix can be split to avoid tiny panels – this is even done by default. For example, if a scatterplot matrix of 15 variables is needed, `plmatrix` produces by default 6 pages of graphical output, the first one showing the upper corner of the lower triangle of the scatterplot matrix, the second, variables 7 to 11 against 1 to 6, the third, 7 to 11 against 7 to 12, and so on, until the whole lower triangle of the full scatterplot matrix is presented.
- Plotting characters and colors may be specified directly (instead of writing a panel function).

5 Utility Functions

5.1 Documentation

For a data analysis, it is often useful to save graphical displays in documents or on paper. In an extended data analysis, one can easily lose control over the precise content of the different plots. `regr0` provides some help for keeping track.

- Every graphical display generated by a graphical function from the package gets a "stamp" in the lower right corner that indicates date and time of its creation and, if specified by the user before that time point, a project title and a step name (by writing `options(project=projecttitle, step=stepname)`). (This stamp can of course be suppressed for producing publication graphics.)
- Data sets may be documented by attaching two attributes, `tit` and `doc` – title and description –, which will be printed with numerical output if desired.

5.2 Multiple Frames

Function `mframe` splits the screen by calling `par(mfrow=...)`. It adds flexibility and sets other defaults for margin width and the like.

This is the end of the story for the time being. I hope that you will get into using `regr0` and have good success with your data analyses.