Paper 110-2009

A SAS® Macro for Adaptive Regression Modeling

George J. Knafl, University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT

A SAS macro called genreg is available from the author for conducting adaptive regression modeling. It is written primarily in the matrix language PROC IML and supports nonparametric linear, logistic, and Poisson regression modeling of expected values and/or of variances/dispersions in terms of fractional polynomials in one or more predictor variables. Fractional polynomial models are compared using k-fold likelihood cross-validation (LCV) and adaptively selected through heuristic search. The genreg macro supports modeling of independent outcomes under normal, logistic, and Poisson distributions. For the normal distribution case, it also supports modeling of repeated measurements under either order 1 autoregressive or exchangeable (i.e., constant) correlation structures. For the logistic case, it supports polytomous-valued outcomes under generalized logits. An overview of the macro is presented and its use demonstrated through an adaptive growth curve analysis.

INTRODUCTION

A SAS macro called genreg is available from the author at http://www.unc.edu/~gknafl/gcurve.html for conducting adaptive regression modeling. It is written primarily in the matrix language PROC IML and currently consists of over 20,000 lines of code and comments and has an interface consisting of about 125 macro parameters. It supports nonparametric linear, logistic, and Poisson regression modeling of expected values and/or of variances/dispersions in terms of fractional polynomials (Royston & Altman, 1994) in one or more predictor variables. For linear regression, expected values are modeled as linear in associated coefficient parameters. For logistic and Poisson regression, expected values are transformed by logit and log link functions, respectively, and then modeled as linear in associated coefficient parameters. In all cases, logs of the variances/dispersions are modeled as linear in associated coefficient parameters. Fractional polynomial models are compared using k-fold likelihood cross-validation (LCV) and adaptively selected through heuristic search (Knafl et al., 2004; Knafl, Fennie, & O'Malley, 2006). The search process systematically expands the model by adding transforms of predictor variables and then contracts the expanded model by removing extraneous transforms, if any, and adjusting the powers of remaining transforms.

The genreg macro supports modeling of independent outcomes under normal, logistic, and Poisson distributions. For the normal distribution case, it also supports modeling of repeated measurements under either order 1 autoregressive or exchangeable (i.e., constant) correlation structures. For the logistic case, it supports polytomous-valued outcomes under generalized logits. An overview of the macro is presented and its use demonstrated through an adaptive analysis of the classic growth curve data of Porthoff and Roy (1964).

To facilitate its use, genreg contains an extensive header comment describing the functioning of its macro parameters. Execution begins with a thorough error-checking phase, issuing appropriate error and warning messages, and is aborted if parameters are improperly specified. Settings for macro parameters are listed in the output along with formatted modeling results so that the output provides thorough documentation for the requested analysis. The macro has only been used so far within the Windows operating system, but has no Windows-specific code to our knowledge and so should be portable to other operating systems.

SETUP

Assume that the genreg code has been stored in the file called genreg.20090101.sas (version numbers are indicated by dates in the file name) located in the c:\macros folder. This code can then be loaded with the following %include command.

```
   %include "c:\macros\genreg.20090101.sas";
```

Assume that output has been formatted with the following options and title commands.

```
   options linesize=76 pagesize=53 pageno=1 nodate;
   title "Analysis of Dental Measurements";
```

Assume as well that the dental measurements reported by Porthoff & Roy (1964) have been loaded into a SAS data set called DENTDATA in the SAS work library. These data consist of dental measurements in millimeters at 8, 10, 12, and 14 years of age for 27 children, with 11 of them girls and 16 boys, for a total of m=27·4=108 measurements.

Assume that DENTDATA contains variables called SUBJECT containing subject identifiers, AGE containing ages at the time of measurement, SEX containing sex codes of 0 for a girl and 1 for a boy, and DENTMEAS containing dental measurement values at assocated ages.

STANDARD REGRESSION MODELING

A linear regression model with expected values linear in AGE, constant variances, and exchangeable correlations can then be generated using the following command.

```
%genreg(modtype=NORML,datain=DENTDATA,yvar=DENTMEAS,xvars=AGE,matchvar=SUBJECT,
        withinvr=AGE,covtype=EC,foldcnt=10);
```

The modtype parameter indicates the type of likelihood to be used. NORML is used for linear regression analyses including those involving repeated measurements. The other options are LOGIS for logistic regression and POISS for Poisson regression, but those options do not currently support repeated measurements. The name of the source data set is provided by the datain parameter and the name of the outcome (response, dependent) variable by the yvar parameter.

The base model for the expected values in this case contains an intercept parameter since the default value for the xintrcpt parameter is Y (for yes), as opposed to N for a no-intercept model. The xvars parameter setting requests that this model also includes the untransformed predictor AGE. The base model for the variances is the default constant (i.e., intercept-only) model corresponding to "`vintrcpt=Y`" and "`vvars=`" (i.e., an empty setting). The xpowers (vpowers) parameter can be used to request transformation of xvars (vvars) variables. For example, "`xvars=AGE`" with "`xpowers=0.5`" requests a square root transform of AGE while "`xvars=AGE AGE`" with "`xpowers=1 2`" requests the standard degree 2 polymomial model in AGE. The variable SEX can be added to the xvars list to control for a main effect to SEX.

For repeated measures data, observations are assumed to be in matched sets corresponding to different values of the matchvar variable, in this case to the different values of the SUBJECT variable (which are indexes from 1 to 27), and with repeated condition values for those sets corresponding to values of the withinvr variable, in this case to different values of the AGE variable (which are 8, 10, 12, and 14 years for each child). Setting "`covtype=EC`" requests exchangeable correlations and, for the case of constant variances, is the same as the standard repeated measures model called compound symmetry. It can be changed to autoregression (of order 1) with "`covtype=AR1`", which is equivalent to the setting "`type=SP(POW)(AGE)`" for the repeated statement of PROC MIXED, not to the setting "`type=AR(1)`".

The "`foldcnt=10`" setting requests a 10-fold LCV. In other words, matched sets of measurements (corresponding to subjects in this case) are randomly partitioned into 10 disjoint subsets of observations called folds. Likelihoods for data in folds are computed using maximum likelihood parameter estimates based on data in complements of folds, multiplied together, and normalized by taking the mth root where m is the total number of measurements (m=108 in this case). Higher LCV scores indicate better models, more compatible with the data in terms of the fold assignment. Such "subject"-wise deletion with all the measurements of a matched set in the same fold is the default option. Measurement-wise deletion with folds based on random subsets of measurements can be requested by "`measdlte=Y`". This can be useful for assessing the impact of missing outcome values, but there are none for the dental measurement data.

Page 2 of the output for the above macro invocation is displayed in Table 1. In this case, the model for the expected values has estimated intercept of 16.76 and estimated slope for AGE of 0.66. The model for the log of the variances has estimated intercept of 1.84 which translates into an estimated constant standard deviation of 2.51. Maximum likelihood estimates of variance are generated by genreg to be consistent with LCV model evaluation, rather than unbiased estimates. The constant correlation is estimated to be 0.68 and the 10-fold LCV score is 0.11859 (the last entry of the output).

Table 1
Example Output
Model for the Expected Values Linear in AGE

```
                      Analysis of Dental Measurements                      2

                                 base model


model correlation structure:                                           EC
# of matched sets:                                                     27
maximum # of distinct values within matched sets:                       4
m, the number of measurements:                                        108
lower bound on correlation:                                          -0.3


                         base expectation component


                  predictor     power    estimate

                  XINTRCPT          1   16.761111
                  AGE               1   0.6601852


                        base log variance component


                  predictor     power    estimate

                  VINTRCPT          1    1.843992


iterations:                                                             5
estimated correlation:                                          0.6799041
MLE of response variance:                                       6.3217242
MLE of response standard deviation:                             2.5143039
log likelihood:                                                 -221.6948
-2 log likelihood:                                              443.38956
average log likelihood:                                         -2.052729
mth root of the likelihood:                                      0.128384


average deleted square error:                                   6.6354624
standard deleted prediction error:                              2.5759391
log likelihood using deleted predictions:                       -230.2688
-2 log likelihood using deleted predictions:                    460.53762
average log likelihood using deleted predictions:               -2.132119
mth root of the likelihood using deleted predictions:           0.1185858
```

ADAPTIVE REGRESSION MODELING

An adaptive growth curve model for the expected values of the dental measurements in terms of children's ages can be requested as follows.

```
%genreg(modtype=NORML,datain=DENTDATA,yvar=DENTMEAS,matchvar=SUBJECT,withinvr=AGE,
        covtype=EC,foldcnt=10,expand=Y,expxvars=AGE,contract=Y);
```

An expansion is requested by "`expand=Y`", in this case with "`expxvars=AGE`" specifying the one variable to consider for expansion of the model for the expected values. The base model for the search in this case is the default constant model (i.e, corresponding to the default settings of "`xintrcpt=Y`" and "`xvars=`"), which can be changed if desired. By default "`exptrans=Y`", meaning consider power transforms of the expansion variables (as opposed to "`exptrans=N`" for generating linear models), and "`multtrns=Y`", meaning consider inclusion of multiple transforms of expansion variables (as opposed to "`multtrns=N`" for limiting the model to at most 1 transform per expansion variable). No variables are considered for expansion of the model for the variances since the expvvars parameter is not set and so has its default empty specification.

A contraction is requested by setting "`contract=Y`" and always follows the expansion (unless "`expand=N`" and there is no expansion). Intercept terms of base models are considered for removal in the contraction, but all model terms in the base model for the expected values can be held fixed by setting "`nocnxbas=Y`". Alternately, the "`nocnxint=Y`" setting holds the intercept for the expected values fixed while allowing all other terms of the model for the expected values to be considered for removal in the contraction. By default, "`cnretrns=Y`" requesting retransformation of transforms remaining in the model at each step of the contraction.

A crucial issue for the contraction is when to stop it. Stopping too soon leaves in extraneous transforms while continuing for too long removes valuable transforms. The default approach for stopping the contraction is based on LCV ratio tests analogous to likelihood ratio tests (Knafl, Fennie, & O'Malley, 2006). The genreg macro computes this in terms of an allowable percent decrease in LCV scores, which for these data with m=108 measurements is 1.76% (the associated proportion is reported in the output as the "stop contraction tolerance"). In other words, a percent decrease smaller than 1.76% is insubstantial indicating that the model with the lower LCV score is a competitive alternative to the one with the larger LCV score. On the other hand, a percent decrease larger than 1.76% is substantial indicating that the model with the larger LCV score provides a distinct improvement over the model with the lower LCV score. For example, the model linear in AGE with "`covtype=AR1`" has LCV score 0.10483, 11.6% lower than the score with "`covtype=EC`", indicating that the standard repeated measures approach with constant correlations for all pairs of ages is distinctly more appropriate for these data than autoregressive correlations weakening the further apart the ages are.

The above macro invocation starts by generating a constant base model for the expected values, expands it to include the single transform of AGE raised to the power 2, and then contracts this model by removing the intercept and adjusting the power for AGE to 0.31. The LCV score for this model is 0.12020. In comparison, the linear model in AGE has a 1.34% lower LCV score, indicating that the expected dental measurements are not distinctly nonlinear in AGE.

Our experience is that the setting of the number of folds does not have much of an effect on the results, especially if it is not too small. One way to balance the need for the number of folds to be large enough to produce consistent results while small enough to limit the computations, is to pick a benchmark analysis like the above one, vary foldcnt from 5 by increments of 5 until the LCV score achieves a local maximum, and use that local maximum in subsequent analyses of the data. The first local maximum for the selection of an adaptive model for the expected dental measurements is 10 folds, which is why we used it in the above macro invocation. Moreover, exactly the same model is selected for foldcnt set to 5, 10, and 15, suggesting that results for other analyses of these data are likely to be reasonably robust to the setting of the number of folds.

Parameter estimates are computed by genreg directly, solving associated equations using the matrix language of PROC IML, rather than by invoking standard SAS procedures. Results for the associated SAS procedure, in this case PROC MIXED (or PROC REG for independent normal data, PROC LOGISTIC for binary logistic data, PROC CATMOD for polytomous logistic data, and PROC GENMOD for count data), can be requested by the setting "`procmod=Y`". This can be used to generate p-values for tests of zero coefficients. However, those p-values are of questionable value for adaptively selected models. They are typically significant as a consequence of the heuristic search process and often highly significant.

ADAPTIVE INTERACTION EFFECTS

There is a likely effect to SEX on dental measurements. This can be addressed by adding SEX to the expxvars list, but that only address a possible main effect to SEX rather than an interaction effect. A more general approach is to create the interaction variable "AGEBOYS=SEX*AGE;" in the DENTDATA data set and then use the following macro code to adaptively address both main and interaction effects for SEX.

```
%genreg(modtype=NORML,datain=DENTDATA,yvar=DENTMEAS,xvars=,matchvar=SUBJECT,
        withinvr=AGE,covtype=EC,foldcnt=10,expand=Y,expxvars=AGE SEX AGEBOYS,
        contract=Y);
```

Distinct main and/or interaction effects to SEX exist if the selected model contains SEX and/or one or more transforms of AGEBOYS. In this case, the selected model for the expected values has two terms, AGE raised to the power 0.21 representing the dependence on age for girls and AGEBOYS raised to the power 2 representing the change in that dependence for boys compared to girls. The LCV score for this model is 0.12787. In comparison, the score of 0.12020 for the model without consideration of a SEX effect is 6.00% smaller. Consequently, there is a substantial interaction effect with the dependence of expected dental measurements on AGE different for boys than for girls, indicating that dental measurement change on the average over time differently for boys than for girls.

ADAPTIVE VARIANCE MODELING

The genreg macro also supports adaptive modeling of the log of the variance (or the dispersion for logistic and Poisson regression) in terms of fractional polynomial models. The genreg macro supports four ways to expand and contract the model (X) for the expected values and the model (V) for the variances. These alternatives are requested through the expordr and contordr parameters. For example, the following code requests that the expected value model be expanded before the variance model ("expordr=XV") as well as contracted before the variance model ("contordr=XV") while addressing the possibility of an effect to AGE as well as main and interaction effects to SEX on both the expected values and the variances.

```
%genreg(modtype=NORML,datain=DENTDATA,yvar=DENTMEAS,matchvar=SUBJECT,withinvr=AGE,
        covtype=EC,foldcnt=10,expand=Y,expxvars=AGE SEX AGEBOYS,
        expvvars=AGE SEX AGEBOYS,expordr=XV,contract=Y,contordr=XV);
```

This produces a model with LCV score of 0.13539. In comparison, the associated constant variance model has the 5.56% smaller LCV score of 0.12787, indicating that the dental measurement variances are substantially nonconstant. The LCV scores for the four expansion/contraction alternatives range from 0.13346 to 0.13539, with the smallest score 1.43% smaller than the largest score, so that all four models are competitive alternatives for the dental measurement data, but that need not always be the case. The default settings are "expordr=." and "contordr=.", meaning to search adaptively through the four possible alternatives. Under this default approach, the first component to be expanded is the one whose expansion generates the best LCV score, and the first component to be contracted is the one whose contraction generates the best LCV score. For the dental measurements, the default procedure generates the same model as for "expordr=XV" and "contordr=XV", with the largest score among the four alternatives. These results suggest that the default approach is likely in general to produce at least a competitive model compared to the best model for the four alternatives, if not the best model itself. The model generated by the default approach has two expected value terms, AGE raised to the power 0.23 and AGEBOYS raised to the power 2, about the same as for constant variances, and two variance terms, AGEBOYS raised to the power –0.8 and AGE raised to the power 0.5.

OUTPUT DATA

The genreg macro generates several output data sets. The dataout parameter is used to name a data set containing a copy of the datain data set along with several generated output variables describing the results for the selected model. Macro parameters are also available for setting the names of these output variables. This data set can be exported to an Excel file and used to produce plots of generated results as in Figure 1 for the model based on the default expansion/contraction alternative for the dental measurement data. An inspection of Figure 1 indicates that expected dental measurements are larger for boys than for girls at any age and increase as children get older in nearly linear patterns that become further apart with age. Moreover, dental measurements are more variable for boys than for girls at any age, with the variability decreasing for boys as they age and increasing for girls as they age.

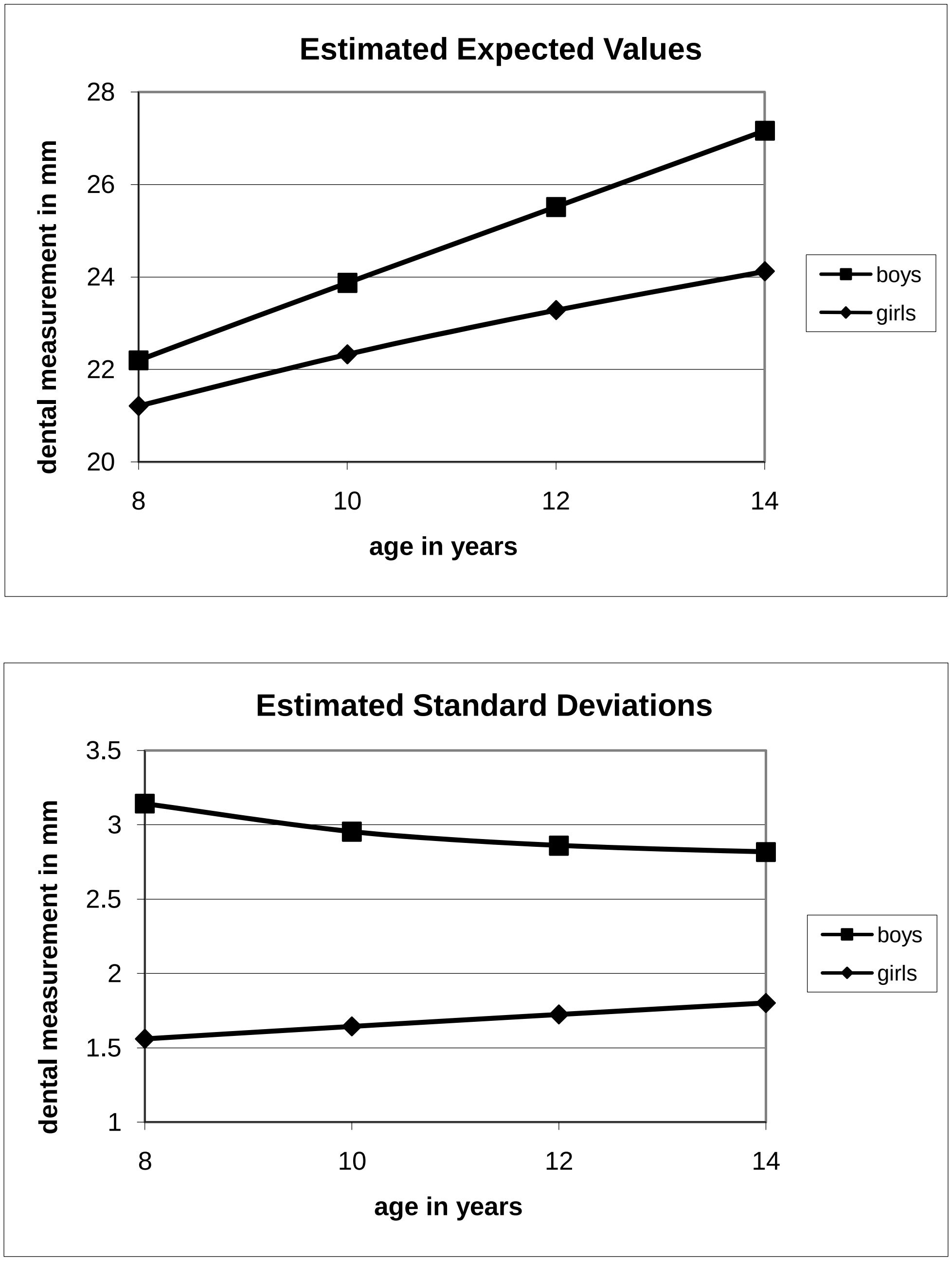

Figure 1
Estimated Dependence of Dental Measurements on Age by Sex of the Child

RESIDUAL ANALYSIS

The genreg macro also supports residual analyses requested by the setting "`ranlysis=Y`". For example, a residual analysis for the adaptively selected model for the expected values and variances of the dental measurements is requested as follows.

```
%genreg(modtype=NORML,datain=DENTDATA,yvar=DENTMEAS,xintrcpt=N,xvars=AGE AGEBOYS,
        xpowers=0.23 2,vintrcpt=N,vvars=AGEBOYS AGE,vpowers=-0.8 0.5,
        matchvar=SUBJECT,withinvr=AGE,covtype=EC,foldcnt=10,ranlysis=Y);
```

For repeated measurements like these, genreg generates independent standardized residuals by transforming the residuals using the inverse of the transpose of the Cholesky root of the estimated covariance matrix. These are called scaled residuals by PROC MIXED (but PROC MIXED can only handle constant variance models). The standardized residual plot generated by genreg is displayed in Table 2. The standardized residuals fall reasonably between ±2 except for two outliers. Listings are also generated that can be used to identify such outliers. The outlier with standardized residual of –3.01 corresponds to the small dental measurement of 16.5 mm for a girl at 8 years old. The model over-estimates this observation. The outlier with standardized residual of 3.69 corresponds to the large dental measurement of 31.0 mm for a boy at 12 years of age. The model under-estimates this observation, a possible consequence of the fluctuating dental measurements for this subject, first decreasing from 23 mm at age 8 to 20.5 mm at age 10, then increasing to 31.0 mm at age 12, and finally decreasing again to 26 mm at age 14.

CONCLUSION

These analyses of the dental measurements demonstrate the use of genreg for adaptive growth curve modeling. The results demonstrate the need for adaptive modeling of variances as well as of expected values. The macro supports adaptive modeling in a variety of other settings as well, but it is still in need of further extension. For example, support is needed for modeling repeated measurements under logistic and Poisson distributions. There is also a need for support of more general correlation structures including adaptively selected random coefficient models.

REFERENCES

Knafl, G. J., Fennie, K. P., Bova, C., Dieckhaus, K., & Williams, A. B. (2004). Electronic monitoring device event modelling on an individual-subject basis using adaptive Poisson regression. *Statistics in Medicine, 23*, 783-801.

Knafl, G. J., Fennie, K. P., & O'Malley, J. P. (2006). Adaptive repeated measures modeling using likelihood cross-validation. In B. Bovaruchuk (Ed.), *Proceedings second IASTED international conference on computational intelligence* (pp. 422-427). Anaheim, CA: ACTA Press.

Porthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika, 51*, 313-326.

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, *43*, 429-467.

ACKNOWLEDGMENTS

This work has been supported in part by Award Number R01 AI57043 from the National Institute of Allergy and Infectious Diseases (NIAID) and by Award Number R03 MH086132 from the National Institute of Mental Health (NIMH). The content is solely the responsibility of the author and does not necessarily represent the official views of the NIAID, the NIMH, or the National Institutes of Health.

Table 2
Example Output
Residual Analysis for Adaptive Model for the Expected Values and Variances in AGE

```
        |
   4.00 +
        |
        |                                          *
        |
        |
        |
   3.00 +
        |
        |
        |
        |
S       |   *
t  2.00 +             *             *                            *
a       |                      *
n       |            *
d       |   *
a       |                             *            *             *
r       |   *        *         *      *
d  1.00 +            *                             *             *
i       |                      *      *            *             *
z       |            *              *
e       |            *              *                            *
d       |            **               *                          *
        |   *        **             * *                          *
R  0.00 +            **             * *            *
e       |   *        *         *                   *             *
s       |             *        *                   *
i       |   *                  *    * *            *             *
d       |            **        *    * *            *             *
u       |   *                  *    *
a -1.00 +                      *    * *            *             *
l       |                      *    *              *             *
        |                             *                          *
        |
        |            *
        |                           *
  -2.00 +             *             *
        |
        |
        |
        |
        |
  -3.00 +   *
        |
        --+--------+--------+--------+--------+--------+--------+--------+--
         21       22       23       24       25       26       27       28

                                 Predicted Outcome
```

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

George J. Knafl
Professor, School of Nursing
Carrington Room 5014, Campus Box 7460
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7460
Work Phone: 919-843-9686
Work Fax: 919-943-9969
E-mail: gknafl@email.unc.edu
Web: http://www.unc.edu/~gknafl/