Advanced Modelling Techniques

Assignment 2 : non-linear models

March 13th, 2006

Soutrik BANERJEE

Introduction

Brief summary results of the LDA first assignment

Using the renal dataset, we previously obtained in the summary statistics (ANCOVA model) that the blood haematocrit level at 10 years is explained by gender and age :

$$hc_{10} = \text{intercept} + \text{age} + \text{male}$$

In the multivariate, 2-stage and random-effects models, we consistently observed the following relationship in all the 3 models, although the parameters estimates in the multivariate model were slightly different from the other two models :

$$\text{Haematocrit} = \text{intercept} + \text{age} + \text{male} + \text{male*year} + \text{male*year}^2$$

Non-linear approach to modelling the renal dataset

With the above results in view, a non-linear modelling approach was tried in this report. The software for the data treatment was SAS® (version 9). The PROC NLMIXED was the principal procedure used to obtain the parameter estimates.

Since detailed description of the dataset was already done in LDA assignments 1 and 2, only comparative results and discussion will be presented in this assignment.

In the non-linear mixed models, it is assumed that the conditional distribution of $\mathbf{Y}_{ij}$, given $\mathbf{b}_i$ belongs to an exponential family.

The mean structure is modelled as [3] :

$$E(\mathbf{Y}_{ij} \mid \mathbf{b}_i) = h(\mathbf{X}_{ij}, \boldsymbol{\beta}, \mathbf{Z}_{ij}, \mathbf{b}_i) \qquad \text{where, } h \text{ is any function}$$

Methods

What non-linear function to be used to model the trend in haematocrit in post-transplant patients ?

The evolution of mean haematocrit over time is shown in Figure 1. It can be seen that there is a sharp increase in blood haematocrit from baseline until 6 months followed by a less pronounced increase of the next 6 months. Thereby a steady-state level was attained, which decreased very slowly over time. This pattern can be explained by the fact that haematocrit level is usually restored in post-transplant patients in 8-10 weeks time on an average [1]. After the attainment of the peak value for nearly 1-2 years in the given dataset, there is an observable slow decline, with fluctuations, until the endpoint (10 years). This could possibly be attributed to the deterioration of renal function in some patients, hence the slow decrease in the mean haematocrit level.

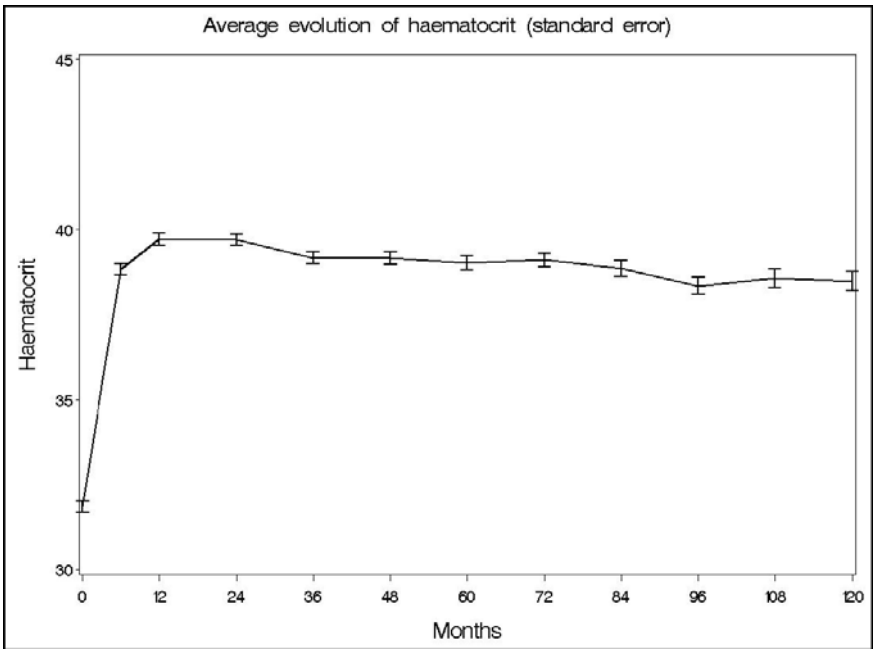


Figure 1. Mean evolution of haematocrit in post-transplant patients in months.

A possible way to model this phenomenon would be to use a suitable cumulative distribution function that stabilises over time (as there is saturation of the effect of the rise of blood haematocrit over time) combined with a linear decrease in haematocrit until the endpoint. The rising part of the curve could be modelled using an exponential, Pareto, Raleigh or half-normal distribution, of which, the first two were chosen to model the evolution.

Using the exponential distribution (combined with a linear decreasing function) didn't provide a good fit, because the exponent (λ) of the exponential distribution was not significant in the fitting of the parsimonious model, hence the 'early rise' in haematocrit didn't fit well in this model. Next, it was thought to use a Pareto distribution, which is given as follows :

$$Y = 1 - (b / time)^a \qquad \text{where, time should be positive}$$

It was found that by dropping the exponent (a) in the above equation and using the time variable in both numerator and denominator, by some trial and error method, a good approximation of the true curve can be made. No [year^2] or, [covariates*year, covariates*year^2] interaction terms were considered to avoid computational problems with PROC NLMIXED, using 3 adaptive quadrature points (however in optimal conditions, they must be used to start the full model). Hence, the full (starting) model is described as follows :

$$\mathbf{Y}_{ij} = \frac{[\text{intercept} + a*\text{male} + b*\text{reject} + c*\text{age} + d*\text{cardio} + e*\text{year}]}{[\,0.1 + f*\text{year}]} + \mathbf{b}_i + \boldsymbol{\varepsilon}_{ij}$$

and, it is assumed that,

$\mathbf{b}_i \sim N(0, d_{11})$

$\boldsymbol{\varepsilon}_{ij} \sim N(0, \sigma^2)$

where, $\mathbf{Y}_{ij}$ is the haematocrit level, where a small constant is added to the denominator in order that it doesn't become zero at baseline. The intercept and the coefficients (of the known covariates) are unknown in the above equation, which are to be determined from the data. In addition, a parsimonious model is to be searched for, using the loglikehihood method [2, 3], in order to reduce the number of parameters in the final model (Table 1). The observed and predicted evolutions of haematocrit in the final model are shown in Figure 2 (groupwise plots are shown in the end of the report). The parameter estimates of the parsimonious model (with standard errors) are given in Table 2.

Results

Model	***-2loglikelihood***	***No. of parameters***	***Difference in parameters***	***X^2 difference***	***P-value***
Full	56306	9			
Full - cardio	56306	8	1	0	1
Full – reject & cardio	56309	7	1	3	.083
Full – year (numerator), reject & cardio	73155	6	1	16846	0
Full – year (denominator), reject & cardio	58652	6	1	2343	0

Table 1. It shows the steps of model reduction, where in the final model, the cardio and reject terms were subtracted, but not the year terms (numerator & denominator).

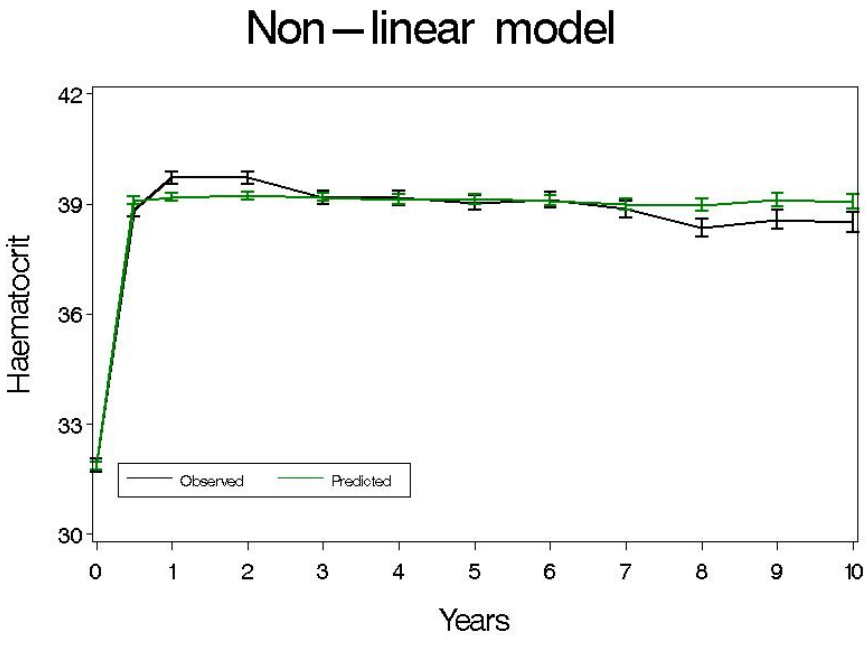


Figure 2. It shows the fitting of the two curves : predicted and observed.

The parsimonious model is thus given by :

$$\mathbf{Y}_{ij} = \frac{[\text{intercept} + a*\text{male} + c*\text{age} + e*\text{year}]}{[0.1 + f*\text{year}]} + \mathbf{b}_i + \boldsymbol{\varepsilon}_{ij}$$

Parameters	***Estimates***	***Standard errors***	***t - value***	***P-value***
'Intercept'	3.02	0.05	60.27	<.0001
a (male)	-0.11	0.03	- 4.08	<.0001
c (age)	0.005	0.001	5.05	<.0001
e (year - numerator)	329.26	171.76	1.92	.0555
f (year - denominator)	8.39	4.38	1.91	.0558
Variance ($\boldsymbol{\varepsilon}_{ij}$)	16.67	0.17	129.60	<.0001
Variance (d_{11})	14.43	0.7	20.69	<.0001

Table 2. Parameter estimates in the parsimonious model. It can be noted the the random-effects' variance (d_{11}) and residual variance ($\boldsymbol{\varepsilon}_{ij}$) are nearly equal.

Discussion

Consistently, in our first LDA assignment and in this report, one finds that *age* and *male* are significant covariates in all the models (ANCOVA, multivariate, 2-stage, random-effects, and finally, in non-linear models). This consistency indicates a significant effect of age and gender on the evolution of haematocrit in the post-transplant patients.

The assumption of normality of the random-effects and error terms were considered for the analysis. The squared residuals (variance function) are approximately uniformly distributed with respect to time, which is seen as almost horizontal 'loess' smoothing line (except for an initial decrease due to a greater baseline variability) and squared residuals plot in Figure 3. The latter shows no particular trend.

An adaptive Gaussian quadrature of 3 points was used in this report, however better results would be provided by refitting with higher Gaussian quadrature.

In addition, random effects could also be added to the numerator of the mean structure instead of treating separately.

Figure 3 showing 'loess' smoothing and squared residual plot over time (*year*).

Next, one would like to interprete the significance of the parameters in this non-linear parsimonious model. The 'intercept' and the coefficients for *age*, *male* are mainly responsible for the starting values, whereas the coefficients of the time variable (*year*) are responsible for the manner of rise (of fall) of the curve, followed by a steadying 'plateau' effect. The rapidity of rise (or fall) is directly proportional to the numerator coefficient of *year*, and inversely proportional to the denominator coefficient of *year*. These two *year* coefficients also control the final height attained at the endpoint (along with the 'intercept').

In the first attempt to model the evolution of the response with an exponential distribution combined with a linear slope didn't succeed and hence, this second approach was used. However, it goes without saying that different possible functions could be explored to better fit the current dataset in future.

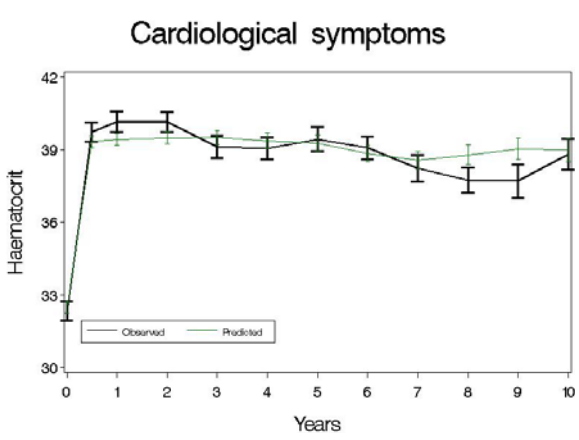
References :

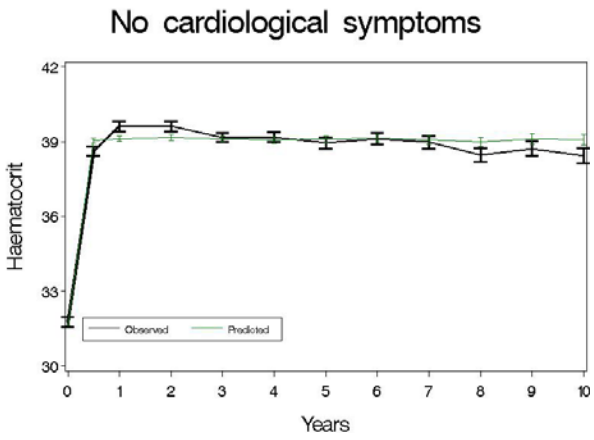
[1] Kessler M. Erythropoietin and erythropoiesis in renal transplantation. *Nephrol Dial Transplant* 1995;**10**(Suppl 6):114-116

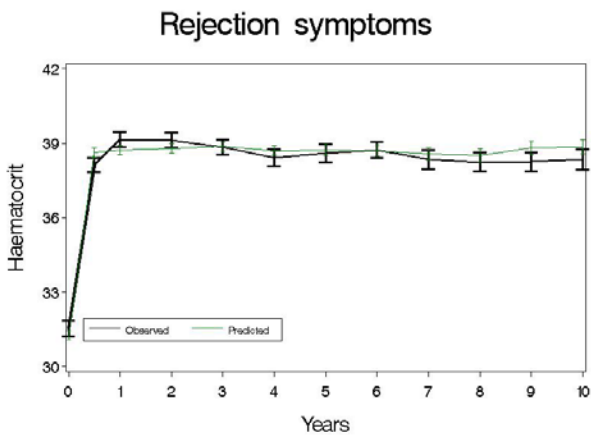
[2] Molenberghs G and Verbeke G. *Longitudinal Data Analysis* (courses notes), Universiteit Hasselt, 2005-2006

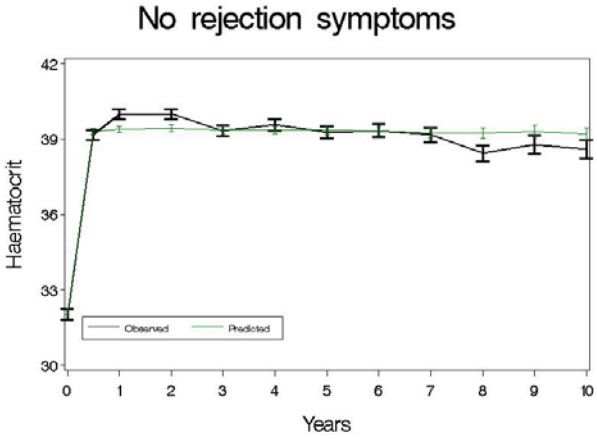
[3] Verbeke G and Molenberghs G. *Advanced Modelling Techniques* (courses notes), Universiteit Hasselt, 2005-2006

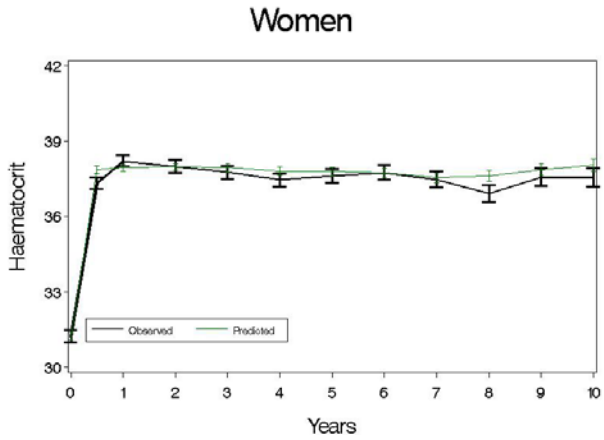
Groupwise plots : predicted versus observed

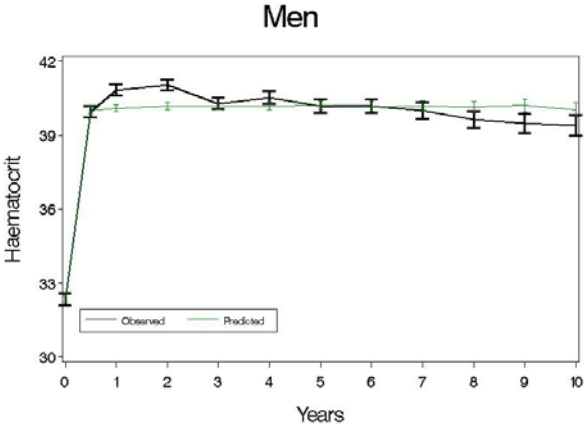












SAS codes :

```
libname Mit "C:\Documents and Settings\Soutrik Banerjee\My documents\Mit\Biostat
master uhasselt\Hausaufgaben\Longitudinal data analysis\Datasets";
proc contents data = Mit.Renal;
run;


/*transposing the horizontal data to vertical*/

data renal0 (keep = id hc jahr age male cardio reject);
set Mit.Renal;
id = _n_;
array zeit {12} hc0 hc06 hc1 hc2 hc3 hc4 hc5 hc6 hc7 hc8 hc9 hc10;
do i = 1 to 12;
     jahr = i;
     hc = zeit {i};
     output renal0;
end;
run;

data renal1;
set renal0;
if jahr = 1 then year = 0;
if jahr = 2 then year = 0.5;
if jahr = 3 then year = 1;
if jahr = 4 then year = 2;
if jahr = 5 then year = 3;
if jahr = 6 then year = 4;
if jahr = 7 then year = 5;
if jahr = 8 then year = 6;
if jahr = 9 then year = 7;
if jahr = 10 then year = 8;
if jahr = 11 then year = 9;
if jahr = 12 then year = 10;
run;


/*deleting missing values for the response and predictor variables*/

data renal2;
set renal1;
where hc ^= . & age ^= . & cardio ^= . & male ^= . & reject ^= .;
run;


/*non-linear full model*/
/*a small constant is added to the denominator to make it +ve*/

proc nlmixed data = renal2 qpoints = 3 maxiter = 500;
parms num1 = 2.67 num2 = 0.12 num3 = 0.05 num4 = 0.01 num5 = -0.03 num6 = 329.29
      den1 = 8.42
      sigma = 6.05 d11 = 1;
numerator = num1 + num2*male + num3*reject + num4*age + num5*cardio + num6*year;
denominator = 0.1 + den1*year;
ratio = (numerator / denominator) + error;
model hc ~ normal (ratio, sigma**2);
random  error ~ normal (0, d11) subject = id;
predict ratio out = estimated;
run;
```

```
/*non-linear final model*/

proc nlmixed data = renal2 qpoints = 3 maxiter = 500;
parms num1 = 3.02 num2 = -0.10 num4 = 0.01 num6 = 329.26
      den1 = 8.39
      sigma = 4.08 d11 = 14.43;
numerator = num1 + num2*male + num4*age + num6*year;
denominator = 0.1 + den1*year;
ratio = (numerator / denominator) + error;
model hc ~ normal (ratio, sigma**2);
random  error ~ normal (0, d11) subject = id;
predict ratio out = estimated;
run;


/*graph plotting*/

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = estimated;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 0.8 mode = include;
symbol2 c = red v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'Non-linear model';
run;quit;


/*residual plot*/

data estimated1;
set estimated;
residual = pred - hc;
residual2 = (pred - hc)**2;
run;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = estimated1;
plot residual2*year = 1 residual2*year = 2 / overlay skipmiss haxis = axis1
vaxis = axis2;
symbol1 c = black i = std1mjt w = 2 mode = include ;
symbol2 c = black v = dot h = 0.2 mode = include ;
axis1 label = (h = 2 'Years') value = (h = 1.5) minor = none;
axis2 label = (h = 2 A = 90 'Squared residuals') minor = none;
title h = 2 'Squared residuals versus years';
run;
quit;


/*Groupwise plots*/

data male;
set estimated;
where male = 1;
run;

data female;
```

```
set estimated;
where male = 0;
run;

data reject;
set estimated;
where reject = 1;
run;

data nonreject;
set estimated;
where reject = 0;
run;

data cardio;
set estimated;
where cardio = 1;
run;

data noncardio;
set estimated;
where cardio = 0;
run;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = male;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'Men';
run;quit;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = female;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'Women';
run;quit;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = reject;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
```

```
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'Rejection symptoms';
run;quit;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = nonreject;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'No rejection symptoms';
run;quit;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = cardio;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'Cardiological symptoms';
run;quit;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = noncardio;
plot hc*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend =
legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside)
offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor =
none;
axis2 label =(h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by
5) minor = none;
legend1 value = ('Observed' 'Predicted' h = 2) label = none frame position =
(bottom left inside) offset = (3,3);
title h = 3 'No cardiological symptoms';
run;quit;
```