Paper 268-28

Smoothing with SAS® PROC MIXED

Alex Pedan, PharMetrics Inc, Watertown, MA

ABSTRACT

Mixed models are an extension of regression models that allows for incorporation of random effects. The application of mixed-effects models to practical data analysis has greatly expanded with consequent development of theory and computer software. It also turns out that mixed models are closely related to smoothing. Nonparametric regression models, especially the general smoothing spline models, are well known for their ability to fit an arbitrary mean response function. This paper describes the use of the MIXED procedure for fitting nonparametric or semi-parametric regression models. Compared with such SAS procedures as PROC LOESS and PROC TPSPLINE, the use of the PROC MIXED allows the fitting of a wide spectrum of complex non-parametric and semi-parametric regression models with simultaneous modeling of trends and covariance structure.

INTRODUCTION

Many statistical models rely on the assumption that the effects of continuous predictors are linear. However, the linearity assumption may be too simple to represent the effects of some risk factors correctly. More specifically, if the linearity assumption is incorrect for a given risk factor, the parametric estimate may underestimate its effect over some range of values or overestimate the effect over some other range, or both.

In the last decade, a number of flexible nonparametric extensions of the conventional linear model have been proposed in the statistical literature (Green and Silverman 1995, Eubank 1999). These nonparametric regression methods eliminate the restrictive linearity assumption and thus allow greater flexibility in modeling the data so that the estimated effects of continuous predictors may follow an arbitrary continuous smooth function. Accordingly the risk of bias is greatly reduced as the estimates depend more on empirical data and less on *a priori* assumptions.

NONPARAMETRIC REGRESSION

The result of a nonparametric regression analysis is a curve fitted to a set of data (y_i,x_i):

$$y_i=f(x_i)+\varepsilon_i, \qquad (1)$$

with f(x) a smooth function and error ε_i, i=1,..,n iid N(0, σ_ε^2)

Starting with Version 7, SAS has incorporated two new procedures for performing non-parametric regression analysis: PROC LOESS (local regression) and PROC TPSPLINE (thin-plate smoothing splines).

In the LOESS procedure the method of weighted least squares is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods whose radii are chosen so that each neighborhood contains a specified percentage of data points.

The TPSPLINE procedure uses a penalized least squares method to estimate multivariate regression surface with thin-plate smoothing splines. The regression spline represents the fit as a piecewise polynomial. The regions that define the pieces are separated by a sequence of knots, and it is customary to force the piecewise polynomials to join smoothly at these knots. The smooth function *f(x)* can be found as result of minimization of the residual sum of squares plus a roughness penalty,

$$\sum_{i=1}^{n}(y(x_i)-f(x_i))^2+\lambda\int(f^{(p)}(x))^2\,dx\,, \qquad (2)$$

where $f^{(p)}(x)$ is the pth derivative of the function *f(x)*.

The resultant curve fitted to the data is a piecewise polynomial of degree 2p-1. The smoothing parameter λ governs the trade-off between smoothness and goodness of fit. This parameter is often unknown in practice and needs to be estimated from the data. A classical data-driven approach to selecting the smoothing parameter is cross-validation, which leaves out one subject's entire data at a time. However, this approach is often computationally intensive and the subsequent inference is difficult.

In addition to the LOESS and TPSPLINE procedures, starting with version 8 SAS introduced PROC GAM, which incorporates a generalized additive model (Hastie and Tibshirani 1990). Generalized additive models are based on the additivity assumption for a multivariate function $f(X_1, X_2,\ldots,X_m)$

$$f(X_1, X_2,\ldots,X_m)= f(X_1) + f(X_2) + \ldots + f(X_m)$$

and allow for a link between $f(X_1, X_2,\ldots,X_m)$ and the expected value of outcome Y. The GAM procedure can utilize either penalized spline or local regression methods to perform smoothing. Different levels of smoothness are possible for different model components, but how smooth each component should be is not an easy question, because cross-validation is very difficult to apply for multiple smoothing parameter problems. The PROC GAM uses a complex iterative procedure, known as backfitting algorithm, to fit the data.

SMOOTHING THROUGH MIXED MODEL

In all the aforementioned methods observations are regarded as independent. However in many longitudinal studies the data collected for each subject are correlated and such correlation should be taken into account in order to produce a valid inference.

New methods have recently been developed in which inference for all model components can easily proceed in a unified linear mixed model framework. These additive models describe complex covariate effects in each subject, while allowing for unexplained population heterogeneity and serial or spatial correlation among repeated measurements.

Let $k_1,\ldots,k_K$ be a set of distinct numbers inside the range of the x_i's and let $x_+=\max(0,x)$.

A random coefficient linear regression spline model for *f(x)* is

$$f(x)=\beta_0+\beta_1 x+\sum_{k=1}^{K}u_k(x-\kappa_k)_+\,, \qquad (3a)$$

where $\mathbf{u}=[u_1,\ldots,u_K]^T \sim N(\mathbf{0},\ \sigma_u^2\,\mathbf{I})$ is independent of $\varepsilon=[\varepsilon_1,\ldots,\varepsilon_n]^T$. Set of functions such as $(x-\kappa_k)_+$ is called a linear spline basis and the values of κ_k are referred as a knots. When σ_u^2 =0, *f(x)* is linear, but for σ_u^2 >0, the truncated lines $(x-\kappa_k)_+$ flexibly allow for nonlinearities in *f*. More smoothness could be attained by instead using quadratic (or cubic) polynomials:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K} u_k (x - \kappa_k)_+^2, \qquad \text{(3b)}$$

We can combine equations (1) and (3b) in one model

$$\mathbf{y}=\mathbf{X}\beta + \mathbf{Zu} + \varepsilon, \quad [\mathbf{u}, \varepsilon]^T \sim N(\mathbf{0}, \text{diag}(\sigma_u^2 \mathbf{I}, \ \sigma_\varepsilon^2 \mathbf{I})), \qquad \text{(4)}$$

where

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_n & x_n^2 \end{bmatrix}, \quad Z = \begin{bmatrix} (x_1-\kappa_1)_+^2 \ldots (x_1-\kappa_K)_+^2 \\ . \quad\quad\quad\quad . \\ . \quad\quad\quad\quad . \\ . \quad\quad\quad\quad . \\ (x_n-\kappa_1)_+^2 \ldots (x_n-\kappa_K)_+^2 \end{bmatrix} \qquad \text{(5)}$$

Equation (4) is nothing but a normal linear mixed model and, for any given σ_u and σ_ε, the estimated best linear unbiased predictor (EBLUP) of **y**,

$$\hat{f} = X \hat{\beta} + Z \hat{u} \qquad \text{(6)}$$

Unbiased refers here to the property that the average value of the estimate is equal to the average value of the quantity being estimated, that is E($\hat{f}$)=E(*f*). Equation (6) can be rewritten (McCulloch and Searle 2001) as

$$\hat{f} = C(C^T C + \lambda^2 D) C^T y, \qquad \text{(7)}$$

where C=[**X Z**], **D**=diag($\mathbf{0}_{p+1}$,$\mathbf{I}_K$) and $\lambda^p=\sigma_\varepsilon/\sigma_u$ for the p^th degree of penalized spline model (p=2 in the case of the quadratic spline model (3b))

It has been shown (Wang 1998, Brumback and Rice 1998) that the EBLUP estimates (7) evaluated at design points are the same as the penalized regression spline solution to Equation 2. Thus it turns out that the nonparametric smoothing spline regression is equivalent to a mixed-effects model (4). In this representation the smoothing parameter is related to the ratio of variance components: $\lambda^p=\sigma_\varepsilon/\sigma_u$. One can now fit models of the form seen in Equation 4 using PROC MIXED.

DATA SET

As a motivating example, we will use data from a clinical trial, which was conducted to evaluate a novel approach to hemodialysis vascular access offered by the LifeSite Hemodialysis Access System (Vasca, Inc., Tewksbury, MA). The LifeSite® Hemodialysis Access System is a subcutaneous access device designed to provide superior blood flow while minimizing access-related complications. The study was an open-label, prospective, multicenter, randomized clinical trial (Schwab et al. 2002) The study population consisted of patients with end-stage renal disease (ESRD) who required hemodialysis and were appropriate candidates for both the LifeSite and the Tesio-Cath devices (control device). Seventy patients were randomized using a central randomization scheme. Of these, 36 patients were implanted with the LifeSite and 34 patients were implanted with Tesio-Cath devices. There were 7 visits resulting in collection of study data in a three-month evaluation period. The performance of each device was determined by its capacity to provide adequate blood flow rate during dialysis. The blood pumps were set to achieve target machine-indicated blood flow rates (QB) of 100, 150, 200, 250, 300, 400, 500 and 600 mL/min. At these blood pump settings, true blood flow rates (BFR) and return (arterial) pressures (RP) were recorded. The primary goal of the trial was to estimate the growth curves for BFR vs. RP for the two treatment groups and to test the equality of the curves.

Challenges in modeling the blood flow rate in hemodialysis include the incorporation into the analysis the complex functional form of BFR as well as the longitudinal characteristic of the measurements. The use of non-linear parametric models could lead to misspecification of the model (Pedan 2001). The alternative solution is to use the nonparametric approach, which offers a convenient way to accommodate the non-linear behavior of BFR without making any *a priori* assumptions.

We will use the mixed effect approach to get non-parametric estimate of the growth curve for the BFR vs. RP. In this case the nonparametric model has the form:

$BFR_i = s(RP_i) + error_i$

Here s() - smooth function of return pressure. Note that this model does not contain an intercept term because this is absorbed into the function s().

The input data set to the initial DATA step is the longitudinal data set containing 2061 observations. There are 6 variables in the data set:

SUBJ_NO = subject number
VISIT = visit number
GROUP = 0, if control group
1, if test group
QB = machine-indicated blood flow
BFR = true blood flow rate
RP = return pressure

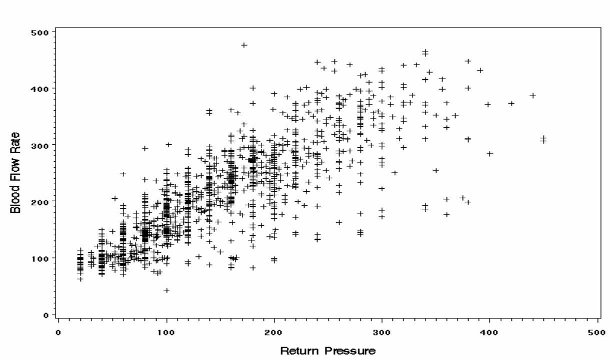
In the example below, we will restrict ourselves to analysis of data for the control device (GROUP=0). Part of the data set is shown in Table 1.

Table 1. Part of the VASCA data set with the observations for subject 33 at visits 1 and 2.

OBS	SUBJ_NO	GROUP	VISIT	QB	BFR	RP
126	33	0	1	100	111.4	40
127	33	0	1	150	156.0	80
128	33	0	1	200	204.6	120
129	33	0	1	250	247.0	160
130	33	0	1	300	282.0	200
131	33	0	1	400	356.0	280
132	33	0	1	500	414.8	340
133	33	0	1	600	447.3	380
134	33	0	2	100	80.1	40
135	33	0	2	150	139.4	180
136	33	0	2	200	191.4	182
137	33	0	2	250	225.2	220
138	33	0	2	300	267.7	260
139	33	0	2	400	327.7	300
140	33	0	2	500	356.8	340
141	33	0	2	600	373.6	360

Figure 1 shows the noisy scatter plot of BFR vs. RP from the VASCA data set.

Figure 1: BFR measurements for Control device from VASCA clinical trial plotted against Return Pressure



SMOOTH CURVE BUILDING

Below, for the sake of simplicity, we will assume that observations are not correlated.

In order to avoid numerical problems when fitting the data, variables RP and BFR were re-scaled from units of mm Hg and mL/min to the units 100mm Hg (RP_S) and L/min (BFR_S) correspondingly:

```
RP_S=RP/100;
BFR_S=BFR/1000;
```

To generate a non-parametric curve using PROC MIXED we have to create a set of basis functions $(x-\kappa_k)_+$ for linear regression spline models or $(x-\kappa_k)_+^2$ for quadratic regression spline model. This is performed by the following programming code:

```
array rp_{8};
 do k=1 to 8;
   *rp_{k}=max(0,rp_s-k*0.5);
     rp_{k}=max(0,rp_s-k*0.5)**2;
end;
```

Here we selected set of K=8 equally spaced knots k_k=0.5,1,…,4. Because smoothing is mainly controlled by the penalty parameter, λ, the number of knots, K, is not crucial . A set of new variables rp_1, rp_2, …, rp_8 for the observations presented in the Table 1 are shown in Table 2.

Table 2: Set of quadratic spline basis functions rp_1, rp_2, …, rp_8 with knots κ_k at 0.5,1,…,4

OBS	rp_1	rp_2	rp_3	rp_4	rp_5	rp_6	rp_7	rp_8
126	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
127	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
128	0.49	0.04	0.00	0.00	0.00	0.00	0.00	0.00
129	1.21	0.36	0.01	0.00	0.00	0.00	0.00	0.00
130	2.25	1.00	0.25	0.00	0.00	0.00	0.00	0.00
131	5.29	3.24	1.69	0.64	0.09	0.00	0.00	0.00
132	8.41	5.76	3.61	1.96	0.81	0.16	0.00	0.00
133	10.89	7.84	5.29	3.24	1.69	0.64	0.09	0.00
134	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
135	1.69	0.64	0.09	0.00	0.00	0.00	0.00	0.00
136	1.71	0.66	0.11	0.00	0.00	0.00	0.00	0.00
137	2.89	1.44	0.49	0.04	0.00	0.00	0.00	0.00
138	4.41	2.56	1.21	0.36	0.01	0.00	0.00	0.00
139	6.25	4.00	2.25	1.00	0.25	0.00	0.00	0.00
140	8.41	5.76	3.61	1.96	0.81	0.16	0.00	0.00

Now we can fit the mixed model (4) by using the following code:

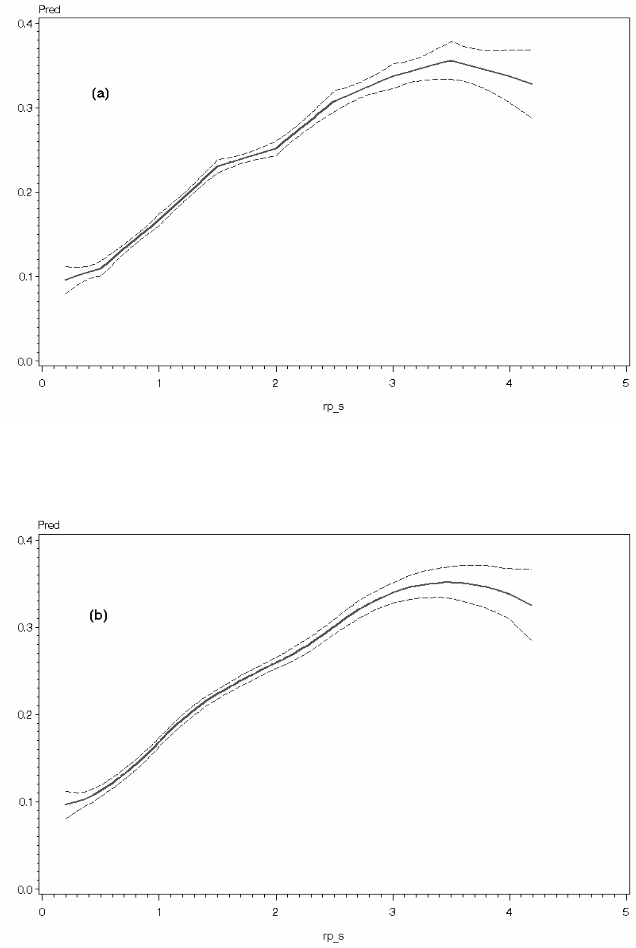
```
proc mixed data=Vasca method=reml;
model tfr_s=rp_s|rp_s/solution outpred=Smooth;
parms (0.0015) (0.015);
*parms (0.0015) (0.015)/hold=1;
random   rp_1-rp_8 /solution type=toep(1);
where group=0;
quit;
run;
```

The statements and options used in this PROC MIXED (SAS Institute 1999) are described below:

DATA= names the input data set (Vasca) for nonparametric growth curve estimation
METHOD= specifies the estimation method for the covariance parameters. The most popular methods being REML (restricted maximum likelihood) the default method, ML(maximum likelihood), MIVQUE0 (minimum variance quadratic unbiased estimation)
MODEL specifies the equation for fixed effects. An intercept is included in the fixed –effects model by default. When bar | is used, the right- and left-hand sides become effects, and the cross of them becomes an effect. Multiple bars are permitted. In the example above rp_s|rp_s is equivalent to rp_s + rp_s*rp_s. An option **SOLUTION** requests that a solution for the fixed-effects parameters (β_0, β_1, β_2) be produced and **OUTPRED =** specifies an output dataset (Smooth) containing EBLUP (*pred*) and related quantities (*Lower*, *Upper* – lower and upper t-type confidence limits for estimated smooth curve (EBLUP) at all study points, *Alpfa* – number defining the confidence level (1- Alpha) for constructing confidence interval, *DF*-degrees of freedom for the t-type confidence limits, *StdErrPred* – standard error of EBLUP and *Resid* - residual). (see Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) for details)
PARMS specifies initial values for the covariance parameters and **HOLD**= specifies which parameters values should be held equal to the specified values
RANDOM defines the random effects constituting **u**=(u_1, u_2,…,u_8)^T vector in the mixed model (4) and **SOLUTION** requests that a solution for the random-effects parameters (u_1,u_2,…,u_8) be produced, **TYPE=** specifies the covariance structure of **G**=*var*(**u**) and **TOEP(1)** specifies σ_u^2 ***I*** structure, where ***I*** is an identity matrix. Multiple RANDOM statements are possible.

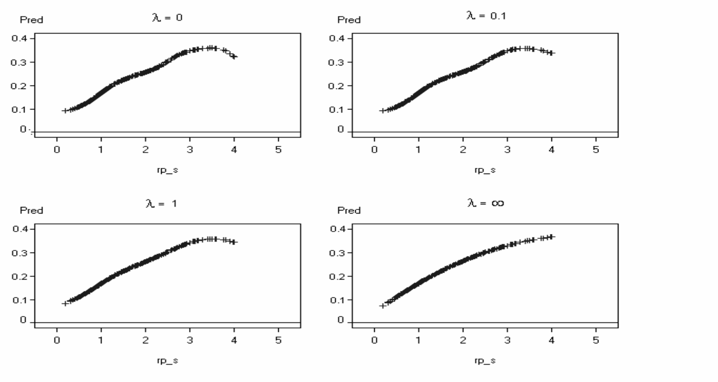
In our example the smoothing parameter $\lambda=\sqrt{\sigma_\varepsilon / \sigma_u}$ was selected via REML estimation of the variance components: $\hat{\sigma}_\varepsilon$ = 0.002237 and $\hat{\sigma}_u$ = 0.007310, which gives us λ=0.55. Sometimes, due to numerical instability, such automatic smoothing parameter selection could give us a zero variance component $\hat{\sigma}_u$, resulting in an over-smoothing of the curve. In order to make this procedure more stable one can use MIVQUE0 method of estimation rather than REML or ML. Another option is to fit various models while keeping one or more of the variance components fixed to a specific value and then to choose the best model. This could be done by using option HOLD=1 in the PARMS statement to fix σ_u , HOLD=2 to fix σ_ε or HOLD=1,2 to fix both the σ_u and σ_ε parameters.
The effect of the degree of the polynomial on estimated function *f* is illustrated in Figure 2. Graph (a) shows REML fit of the linear polynomial model (λ=0.69) and graph (b) shows REML fit of the quadratic polynomial (λ=0.55) model. It is easily seen from graph (a) of Figure 2 that the fitted growth curve *f(x)* in the case of the linear polynomial model consists of piecewise straight lines connecting at the knots. As result, the structure has multiple artificial corners and does not look very smooth. We can see from graph (b) of the Figure 2, that the use of the quadratic polynomial model eliminates corners, making the fit much smoother.

Figure 2: Solid lines: estimated regression curves using (a) linear spline smoother (b) quadratic spline smoother. Dashed lines: 95% confidence intervals



The effect of smoothing parameter λ to the fit is illustrated in Figure 3. On the first graph quadratic spline (3) fit without penalty (λ=0) is shown. Estimate of the curve *f* in this case is a somewhat wiggly piecewise quadratic function, which strongly depends on the number of the knots and their locations. As the smoothing parameter λ increases, the importance of the knot locations and their number decreases and estimates of *f* become smoother, but with loss of some finer details. The case of $\lambda = \infty$ represents the simple least square quadratic regression.

Figure 3: Quadratic penalized spline regression fits to VASCA data for λ values of 0, 0.1, 1, ∞.



CORRELATED ERRORS

In the example above we assumed that the observations are independent. This assumption is definitely violated for the VASCA dataset due to the longitudinal character of data. As we mentioned before, for each subject there were up to 7 visits resulting in collection of study data in a three-month evaluation period. It was shown (Wang 1998) that correlation has a great effect on the selection of smoothing parameters, which are critical to the performance of smoothing spline estimates. Very often failure to take into account correlations will result in undersmoothing. One of the big advantages of the use of mixed model representation for smoothing is that it allows to estimate the smoothing and the correlation parameters simultaneously. For example, to take into account within subject correlations in the VASCA dataset we need to add REPEATED or RANDOM statement(s) with specified correlation structure to the PROC MIXED:

```
random /solution type=AR(1) subject=subj_no;
```

Here we assumed the first-order autoregressive (AR(1) with parameter α) model for errors. The estimates of the parameters σ_ε, σ_u and α are 0.002280, 0.000132 and 0.6069 respectively. The estimate of smoothing parameter λ_α under this covariance structure is 4.2, which is significantly bigger than under assumption of independent errors (α=0).

CONCLUSION

Curve data arise frequently in scientific studies and are an active topic of current statistical research. Non-parametric regression analysis is a powerful tool when there is little prior information about such data or we want to describe new features that parametric analysis ignores. Recent developments in the non-parametric fitting of the non-linear data can be related to the mixed model representation of the penalized spline scatter plot smoothing. This connection suggests a way of fitting non-parametric models using existing software for linear mixed-effects models.

Among the advantages of using a mixed model approach for fitting non-parametric regression (Ruppert, Wand and Carroll 2003) are that this approach utilizes such basic principles of statistics as maximum likelihood and best prediction and that the estimation and inference can be performed in a unified fashion within the mixed model framework, by appealing to the likelihood ratio principles. The mixed model representation can be used to facilitate fitting, inference and model selection. This approach makes it easy to fit more complex models by including additional fixed or random effects or interactions (Coull, Ruppert and Wand 2001). Such a model

assumes that each group has its own mean response curve. Various error structures may be assumed for random errors to account for spatial or longitudinal correlations. In addition, the final model could be chosen by comparison of various goodness-of-fit statistics, such as -2·log-likelihood, AIC or BIC.
SAS PROC MIXED provides an excellent work environment for implementing all such analyses.

REFERENCES

Brumback, B.A. and Rice, J.A. (1998), “Smoothing Spline Models for the Analysis and Crossed Samples of Curves,” *Journal of American Statistical Association*, 93 (443): 961-976

Coull, B.A., Ruppert, D., Wand, M.P. (2001), “Simple incorporation of interactions into additive models,” *Biometrics*, 57(2):539-45.

Eubank, R.L. (1999), *Nonparametric Regression and Spline Smoothing*, New York: Marcel Dekker.

Green, P.J. and Silverman B.W. (1995), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall

McCulloch, C.E. and Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, New York: John Wiley & Sons

Pedan, A. (2001), “A Comparison of Non-Linear Random Effect Model and Semiparametric Generalized Additive Model for Estimation Treatment Effect in Hemodialysis Trial,” *Controlled Clinical Trials*, 22(2S): 97S

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric Regression, New York, Cambridge University Press, (to be published)

SAS Institute Inc. (1999). *SAS/STAT® User’s Guide*, Version 8, Cary, NC: SAS Institute Inc.

Schwab, S.J.,Weiss, M.A.,Rushton, F., Ross, J.P., Jackson, J., Kapoian, T., Yegge, J.,Rosenblatt, M., Caridi, J.M., Reese, W.J., Soundararajan, R. , Work, J., Ross, J., Stainken, B.Pedan, A., Moran, J.A. (2002), “Multicenter Clinical Trial Results with the LifeSite® Hemodialysis Access System,” *Kidney Int*, 62 (3):1026-1033

Wang, Y. (1998), “Smoothing Spline Models With Correlated Random Errors,” *Journal of the American Statistical Association*, 93, 341-348

ACKNOWLEDGMENTS

I thank Prof. Matt Wand for valuable comments.

CONTACT INFORMATION

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

Alex Pedan
PharMetrics, Inc
150 Coolidge Avenue
Watertown MA 02472
Work Phone: 617-972-8590
Fax: 617-972-8587
Email: apedan@pharmetrics.com
Web: www.pharmetrics.com