Poincaré Plots for Residual Analysis

A.I. MCLEOD
Department of Statistical and Actuarial Sciences,
The University of Western Ontario, London, Ontario N6A 5B7
Canada

June 13, 2003

Abstract:

After some suitable ordering of the residuals, $e_t, t = 1, \ldots, n$, it is suggested that scatter plots of e_{t+1} vs. e_t along with a robust smooth loess trend be routinely examined to check for lack of statistical independence. Such plots may be termed Poincaré plots because of their similarity to plots used in nonlinear dynamical systems. Poincaré plots are helpful in detecting positive correlations in the fitted model which may invalidate statistical inferences. Poincaré plots are common in time series analysis but do not appear to be frequently used in other situations.

Key Words: Bootstrap and informal statistical inference; Diagnostic check; Model inadequacy; Residual autocorrelation; Statistical inference.

1. Introduction and Summary

There are many possible threats to the validity of a statistical model, but one of the most potentially serious in many situations is the possibility of lack of statistical independence in the observations. When positive correlation exists and is not taken into account then the estimators will not be fully efficient in many situations. An even more serious problem is that statistical inferences from the model may be completely wrong. Specifically, under positive autocorrelation, it is well known that in the usual regression model the variances are inflated. This means that the usual confidence limits will be too narrow and that p-values will overstate the statistical significance of the results (Wold, 1952, §13.4, p. 211, Theorem 1). This effect of positive correlation applies potentially to all statistical models and not just to models involving time series data.

Indeed McCullagh and Nelder (1989 §2.2, p.26) voice this concern with respect to generalized linear models when they state: "For the random part we assume independence and constant variance of the errors. These assumptions are strong and need checking".

In the case of time series data or when the order of collection of the observations is known, the residual autocorrelation function is often used as well as related statistical tests such as the Durbin-Watson test. For time series regression models, Draper and Smith (1981, §3.9, p.156) recommend plotting the residual for the t-th observation denoted by e_t as a lagged one scatter plot of e_{t+1} vs. e_t. Plotting e_t vs. t may also be useful but since dependency relationships are usually strongest at lag one, the lagged one scatter plot often best reveals problems due to lack of independence in the

data or residuals.

Scatter plots of e_{t+1} vs. e_t, referred to as Poincaré return maps, are used in non-linear time series analysis (Tong, p.1990) and in nonlinear dynamics for identifying limit cycles (Kaplan and Glass, 1995, §6.6, p.304). When applied for the purpose of diagnostic checking of a statistical model we will refer to this type of plot as a Poincaré plot.

The purpose of this article is to show that Poincaré plots are useful in almost all statistical model building even where the chronological order of the data is either not known or not relevant. Specifically many statistical models assume that given the model specification the residuals are statistically independent. Violation of this assumption indicates that the model is misspecified and this misspecification may result in incorrect statistical inferences. Assuming that the observations are statistically independent, the observations may be ordered in various ways. For example, the observations could be ordered according to some covariate.

Poincaré plots may reveal non-linear forms of dependence or features not well summarized by a correlation coefficient. An informal method of statistical inference is to use a parametric bootstrap of the model to examine a sequence of Poincaré plots simulated when the independence assumption is known to hold. The significance of the residual plot can be informally judged by comparing with these plots. More formally, the Kendall rank correlation between e_{t+1} and e_t provides a statistical test for monotone dependency which may be helpful in some cases.

Cleveland (1979) introduced the residual dependency plot. This is defined as a plot of the residuals vs. a covariate along with a loess smooth

to help visualize whether there is a relationship. The Poincaré plot is recommended as a complement not a replacement to this plot. In the following two examples, residual dependency plots do not suggest any model inadequacy. But the Poincaré plots indicate strong positive dependence. Also the forms of dependency revealed by the Poincaré plots in both examples is more complicated than that simple linear correlation.

2. Generalized Linear Modeling Example

Deviance residuals (McCullagh and Nelder, 1989, §2.4.3) are frequently used for diagnostic checking with generalized linear models. Under the usual assumptions the observations, in a specified model, are statistically independent. This implies that the deviance residuals should also be approximately statistically independent. Consider the logistic regression of 189 births fitted by Venables and Ripley (2002, p.194–197). In this model a response variable, low birth weight, is fitted to 9 explanatory variables. One of the explanatory variables is `age`, which represents the age of the mother in years. Using this `age` variable to order the data, the resulting Poincaré plot of the deviance residuals shown in the lower left panel in Figure 1 indicates very strong positive residual dependence. For comparison, the Poincaré plots for 8 bootstrap simulations of deviance residuals are shown in the other panels. To aid the visualization of the dependence relationship or lack thereof a robust linear loess smooth with a span equal to 1 is shown in each panel. Figure 1 clearly reveals that there is significant strong positive dependence in the residuals and so statistical inferences from the fitted model may not be correct. Figure 2 shows the residual dependency

plot of residuals vs. age. In this plot there is no apparent correlation in the residuals.

[Figures 1 and 2]

3. Loess Fitting Example

Cleveland (1993, §3.6, pp.122–127) fits a loess curve to some sunlight polarization data. The response variable is the Babinet point and the explanatory variable is the concentration of particulate matter in the atmosphere. In Cleveland's final fit the Babinet point is regressed on the cube-root of concentration using a robust loess linear regression with a span of 1/3. The data are ordered according to the concentration variable. The resulting Poincaré plot for this fit indicates very strong positive dependency in this data exists. It should be noted Cleveland (1993, p.126, Figure 3.37) found the usual residual dependence plot satisfactory and there is no indication of positive correlation and dependency in this plot. An improved fit can not be obtained simply by changing the loess smoothing parameters in this case. Choosing the span to be zero or close to zero can remove the positive dependence in the Poincaré plot but at the expense of increasing the variance and degrading the overall fit.

Close inspection of Figure 3 shows that many points follow the 45^{deg} line. This means they are exactly equal and hence that both the dependent and independent variable are tied. Such ties are not consistent with the

hypothesis of independent and continuously distributed data. Removing data values corresponding to ties in both variables and refitting the loess curve, as shown in Figure 4, does improve the Poincaré plot a little but there still remains strong positive dependence due to many nearly tied values.

[Figures 3 and 4]

4. Concluding Remarks

Many other examples could be given which indicate the presence of strong positive dependency in the residuals of published statistical models fitted to data. The presence of such strong positive correlation or dependency invalidates statistical inferences from these models. It may also suggest possible sources of variation that can be included in the model to remove this effect.

One should avoid using Poincaré residual plots for data which has been ordered by the response variable. Since positive dependence is expected in the Poincaré residual plot even when the assumption of independence holds in this case.

Many published statistical results in medicine seem to overstate their statistical significance (Matthews, 1998). One possible reason for this apparent lack of robustness in medical statistics, at least in some cases, could simply be model misspecification due to lack of statistical independence.

Further research is needed on ways to improve the statistical model when statistical dependence is found.

ACKNOWLEDGEMENTS

I wish to thank Duncan Murdoch for valuable and insightful comments on this paper.

REFERENCES

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994), *Time Series Analysis: Forecasting and Control*, 3rd Ed., San Francisco: Holden-Day.

Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association* 74, 829–836.

Cleveland, W.S. (1993), *Visualizing Data*, Summit: Hobart Press.

Draper, N.R. and Smith, H. (1981), *Applied Regression Analysis*, 2nd Ed., New York: Wiley.

Kaplan, D. and Glass, L. (1995), *Understanding Nonlinear Dynamics*, New York: Springer-Verlag.

Matthews, R. (1998), "The great health hoax", *Sunday Telegraph*, September 13. Available at `http://www.sepp.org/controv/healthhoax.html`.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd Ed., Boca Raton: Chapman & Hall/CRC.

Tong, H. (1990), *Non-linear Time Series*, Oxford: Oxford University Press.

Venables, W.N. and Ripley, B. (2002), *Modern Applied Statistics with S*, New York: Springer-Verlag.

Wold, H. (1952), *Demand Analysis: A Study in Econometrics*, New York: Wiley.

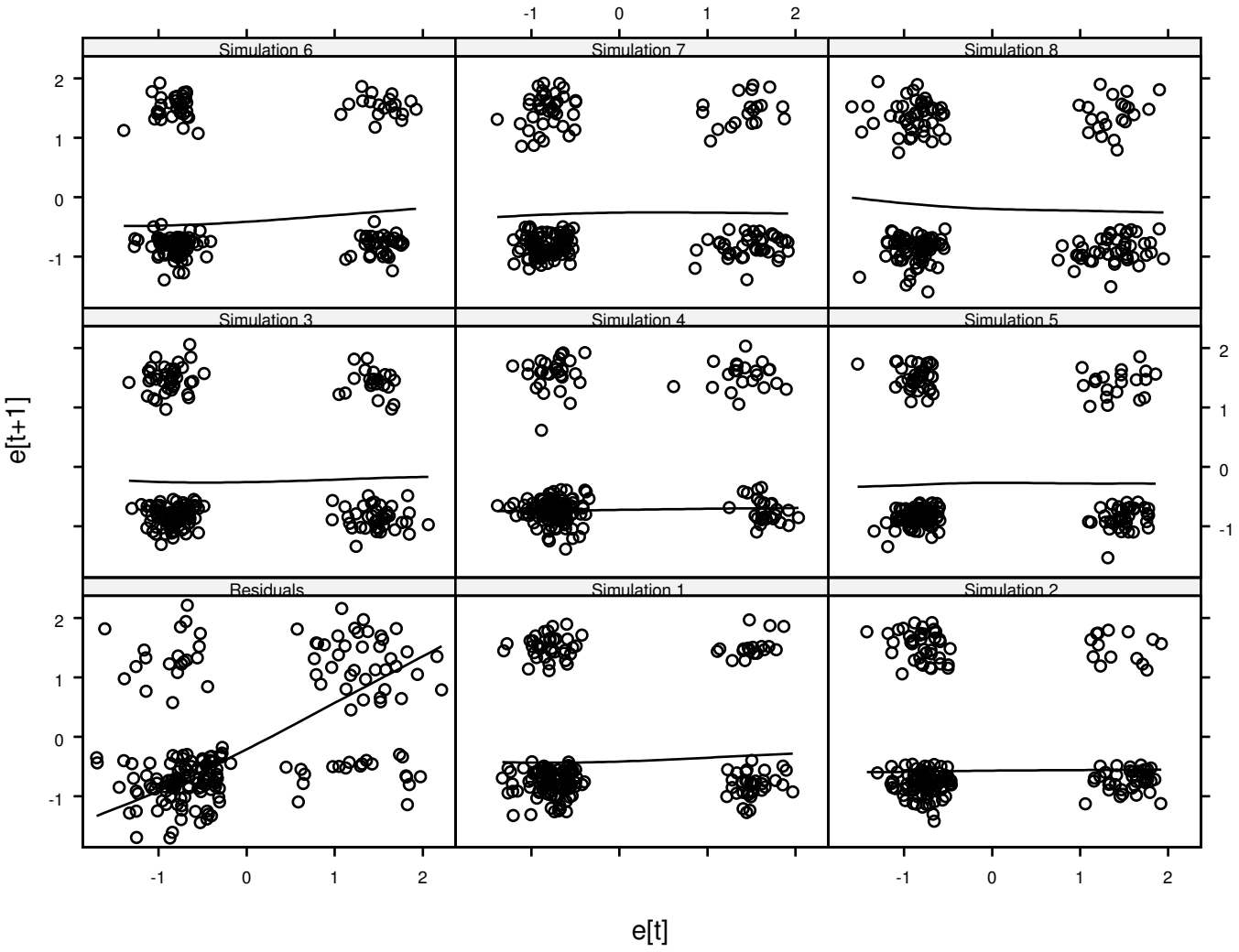

Figure 1: Poincaré plot of deviance residuals in the logistic regression of low birth weight on 9 explanatory variables and 8 parametric bootstrap simulations.

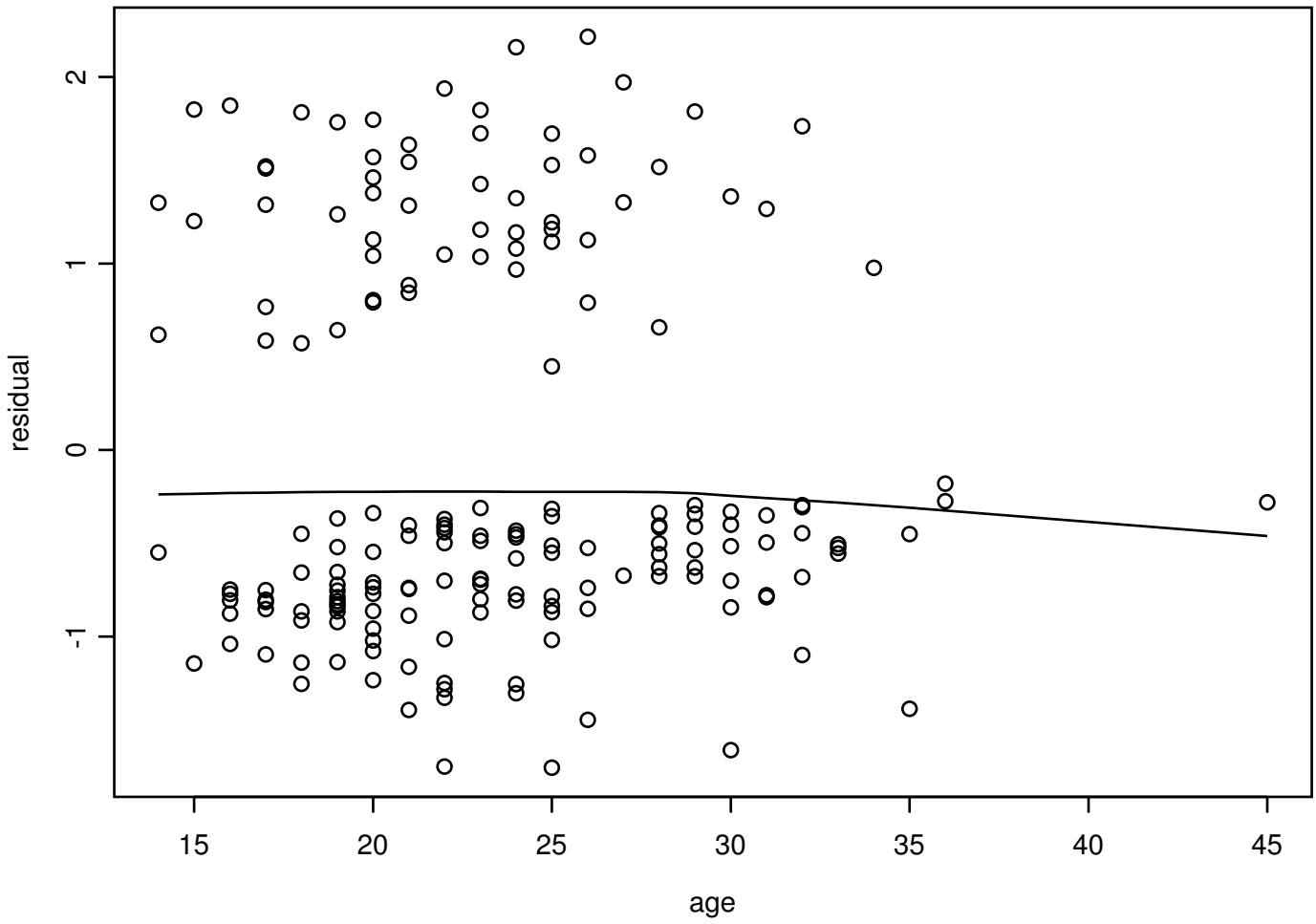

Figure 2: Residual dependency plot of residuals vs. age.

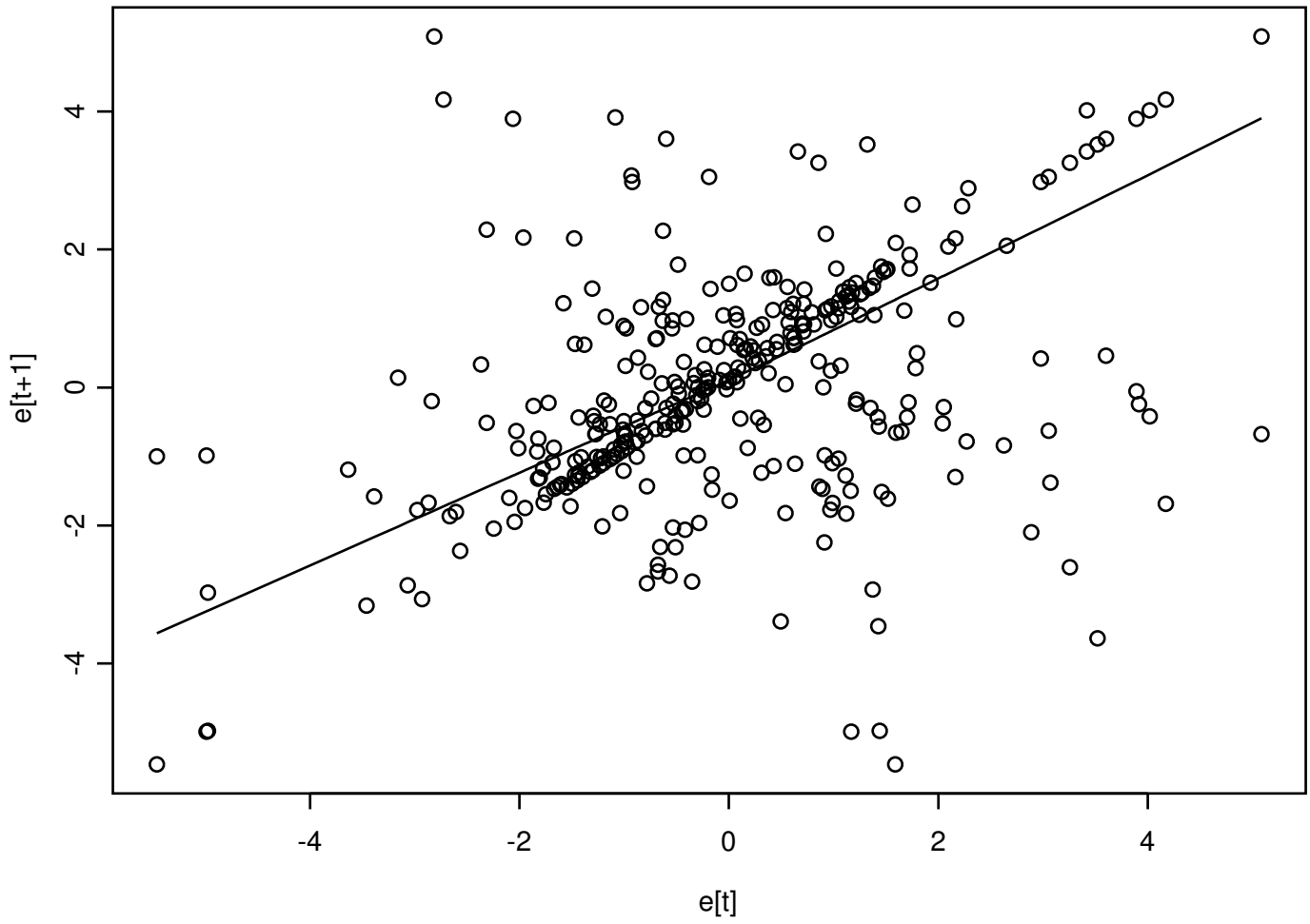

Figure 3: Poincaré plot of residuals in the robust loess fit of Cleveland (1993, §3.6) to the polarization data.

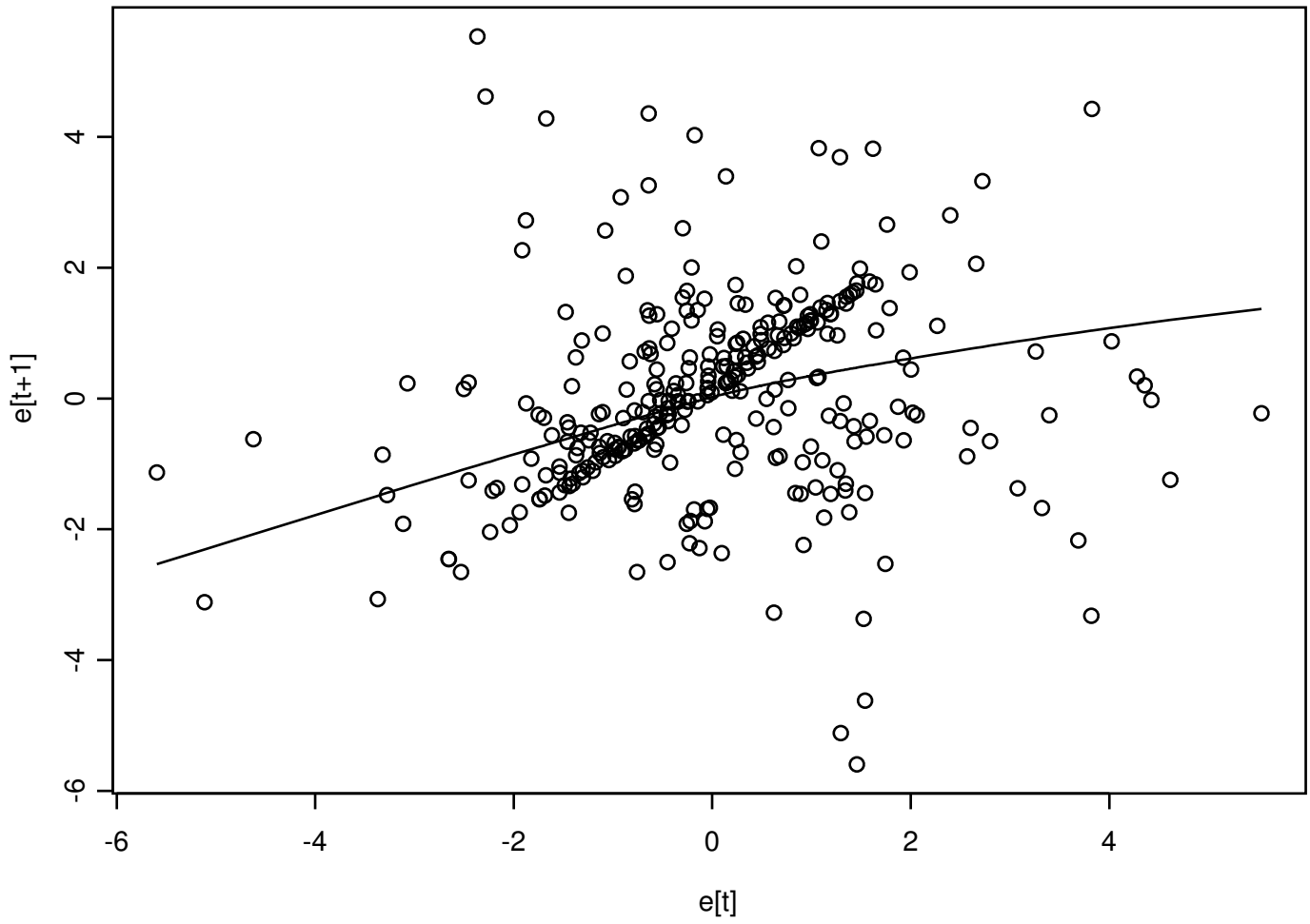

Figure 4: Poincaré plot of residuals in the robust loess fit of Cleveland (1993, §3.6) to the polarization data with ties in both variables removed.

Note For Referee

The purpose of this note is to describe exactly how the nonparametric bootstrap computation reported in §2 was done. The linear component of the logistic model may be written,

$$\zeta_i = \alpha_0 + \sum_{j=1}^{9} \alpha_i x_{i,j}, i = 1, \ldots, n, \tag{1}$$

where $n = 189$ and the parameters, $\alpha_0, \ldots, \alpha_9$ are the estimates obtained from the S `glm` function. The probabilities are given by,

$$\pi_i = \frac{e^{\zeta}}{1 + e^{\zeta}}, i = 1, \ldots, n. \tag{2}$$

The following S function simulates one bootstrap realization of the data,

```
> simulate.birthwt.data
function()
{
    beta <- coef(b.glm)
    zeta <- model.matrix(b.glm) %*% beta
    p <- exp(zeta)/(1 + exp(zeta))
    rbinom(length(p), 1, p)
}
```

The following S Script generates the plot in Figure 1,

```
graphics.off()
o.age<-order(bwt$age)
bwt.age<-bwt[o.age,]
```

```
b.glm<-glm(formula = low ~ ., family = binomial, data = bwt.age)
PoincarePlot(resid(b.glm))
title(main="birthwt.glm, p.195, ordered by age, deviance residuals")
e<-resid(b.glm)
emat<-PoincarePairs(e)
n<-length(e)
nsim<-8
set.seed(181)
for (i in 1:nsim){
    y<-simulate.birthwt.data()
    bwty<-bwt
    bwty$low<-y
    by.glm<-glm(formula = low ~ ., family = binomial, data = bwty)
    ans<-PoincarePlot(resid(by.glm))
    e<-resid(by.glm)
    emat<-rbind(emat, PoincarePairs(e))
    invisible()
}
which <- rep(c("Residuals",paste("Simulation",1:8)),rep(n-1,9))
e.df<-data.frame(emat,which)
names(e.df)<-c("et","etp1","which")
trellis.device(color=F)
xyplot(et~etp1|which, data=e.df,xlab="e[t]",ylab="e[t+1]",strip= function(...)
panel=function(x,y){
    panel.xyplot(x,y)
```

```
    panel.loess(x,y,span=1)
})
```

The S dataframe `bwt` above is obtained as indicated in Venables and Ripley (2002, p.195).