Identification of Outlying Perturbations

By

Robert Weiss*

School of Statistics, University of Minnesota
Technical Report #594
October, 1993

*Robert Weiss is an Assistant Professor at the UCLA School of Public Health, Los Angeles, CA 90024-1772. This work was partially supported by a UCLA Faculty Career Development Award and by NIH grant #GM50011-01. This work was partially completed while on sabbatical at the University of Minnesota, Department of Applied Statistics. The author wishes to thank Sandy Weisberg for helpful comments from an earlier draft.

Identification of Outlying Perturbations

Robert Weiss*
Department of Biostatistics
UCLA School of Public Health
Los Angeles CA 90024-1772 U.S.A.

Keywords: ANCOVA; Bayes factors; Conditional Predictive Ordinate; Censoring; Diagnostics; Influence analysis.

Abstract

This paper introduces methodology for identifying assumptions that are not supported by the data, without the need for fully fitting a new model. The concept of a perturbation is used to embody various assumptions, and a generalization of the conditional predictive ordinate is developed which identifies perturbations supported by the data.

1 Introduction: Models and Perturbations

Let the likelihood and prior from an initial model M_0 combine to give the posterior

$$p(\theta|Y) \propto \prod_{i=1}^{n} f(y_i|\theta, x_i)p(\theta)\,,$$

where Y represents the entire data set, y_i is the i^{th} response, x_i is the vector of predictors, θ is the parameter vector, $f(y_i|\theta, x_i)$ is the sampling density for the i^{th} case, and $p(\theta)$ is the prior. Often we wish to assess the influence of model assumptions on $p(\theta|Y)$. This can be done using perturbation functions.

A perturbation is a function $h(\theta)$ of parameters θ that multiplies $p(\theta|Y)$. Commonly, $h(\theta)$ will be a function of X or Y, but these arguments are not made explicit. Appropriate choice of $h(\theta)$ can be used to assess the influence of model assumptions. Some basic perturbations $h(\theta)$ are
i. case deletion

$$h(\theta) \propto [f(y_i|\theta, x_i)]^{-1}\,;$$

ii. Sensitivity to y_i

$$h(\theta) \propto \frac{f(y_i+\delta|\theta, x_i)}{f(y_i|\theta, x_i)}\,;$$

iii. Sensitivity to x_i

$$h(\theta) \propto \frac{f(y_i|\theta, x_i+\delta)}{f(y_i|\theta, x_i)}\,;\text{and}$$

iv. Prior sensitivity

$$h(\theta) \propto \frac{q(\theta)}{p(\theta)}\,.$$

These perturbations can be combined through multiplication to build up more interesting and complicated perturbations. Case deletion and prior perturbation are familiar Bayesian perturbations, see for example Kass, Tierney, and Kadane (1986), or Kass and Raftery (1993). For examples of additional perturbations see Cook (1986). Two perturbations that differ only by a constant factor are equivalent. In this paper, $h(\theta)$ is assumed to be a ratio of products of sampling densities $f(y|\theta, x)$ or the ratio of proper priors.

After multiplying $p(\theta|Y)$ by $h(\theta)$, we get a perturbed posterior

$$p_h(\theta|Y) = \frac{p(\theta|Y)h(\theta)}{E[h(\theta)|Y]}\,, \qquad (1)$$

The normalizing constant in the denominator on the right hand side is the posterior expectation of $h(\theta)$. Formula (1) is Bayes theorem for perturbations (Weiss 1993). The perturbed posterior can

*This work was partially supported by a UCLA Faculty Career Development Award, and by NIH grant #GM50011-01. This work was partially completed while on sabbatical at the University of Minnesota, Department of Applied Statistics. The author thanks Sandy Weisberg for helpful comments from an earlier draft.

be thought of the posterior resulting from a competing, perturbed, model M_1.

The influence of the perturbation $h(\theta)$ can be assessed using a divergence measure between posteriors (Johnson and Geisser 1982; Pettit and Smith 1985, Weiss 1993). Weiss (1993) recommends the L_1 divergence

$$L_1(h) = .5 \int |p(\theta|Y) - p_h(\theta|Y)| d\theta$$

to assess influence. The $L_1(h)$ statistic is nonnegative and less than 1. Larger values indicate greater influence. Uninfluential perturbations correspond to assumptions whose exact values are unimportant. Values of $L_1(h)$ close to 1 indicate essentially no overlap in support of the two densities. Values less than .1 indicate only minor influence. See Weiss (1993) for more interpretation.

Traditional data analysis assesses the influence and outlyingness of individual cases. In a Bayesian paradigm, the conditional predictive ordinate (CPO) is often used to assess outlyingness. A notation slightly different from what we have developed so far is used for case deletion. Write $p_h(\theta|Y) = p(\theta|Y_{(i)})$, with $Y_{(i)}$ indicating that the i^{th} case has been removed from the data set. Bayes theorem for case deletion is

$$p(\theta|Y) = \frac{p(\theta|Y_{(i)}) f(y_i|\theta, x_i)}{f(y_i|Y_{(i)})},$$

where $f(y_i|Y_{(i)})$ is the predictive distribution of the i^{th} observation given the remainder of the data, or the conditional predictive ordinate or CPO

$$\text{CPO}_i \equiv f(y_i|Y_{(i)}) \equiv \{E[h(\theta)|Y]\}^{-1}.$$

The conditional predictive ordinate can be used to identify outliers (Geisser 1980, Pettit and Smith 1985, Geisser 1987); the smaller the CPO is, the more outlying the observation is.

In case analysis, it is useful both to know if the observation is influential and if it outlying. This paper extends the idea of outlyingness to other perturbations besides case deletion. In particular, the CPO is directly generalized. When a set of perturbations is of interest, then the CPO statistic can be computed for each perturbation. Section 2 extends the definition of CPO. Section 3 gives an example. Section 4 closes with discussion.

2 Bayes factors and the Conditional Predictive Ordinate

The Bayes factor in favor of M_0 against M_1 is

$$\begin{aligned} B(M_0, M_1) &= \frac{f(Y|M_0)}{f(Y|M_1)} \\ &= \frac{\int f(Y|\theta) p(\theta) d\theta}{\int f(Y|\theta) p(\theta) h(\theta) d\theta} \\ &= E[h(\theta)|Y]^{-1}. \end{aligned} \quad (2)$$

Equation (2) shows why the version of $h(\theta)$ that is a ratio of products of sampling densities and/or proper priors was chosen; under this restriction, additional constants are eliminated. Use of improper priors is obviously problematic. When $h(\theta)$ corresponds to case deletion, then $B(M_0, M_1)$ is also the CPO. Equation (2) allows us to generalize the CPO immediately to other perturbations. Outlying perturbations correspond to cases with large $E[h(\theta)|Y]$, small CPO, and the corresponding perturbed models are more supported by the data than M_0. Similarly, inlying perturbations have large $B(M_0, M_1)$, small $E[h(\theta)|Y]$, and are not supported by the data.

Two calculations that can aid in interpretation of $E[h(\theta)|Y]$ are the formula for the posterior probability of the perturbed model, and the formula for calculating the unconditional posterior of θ. Suppose that we have prior probabilities p_0 and p_1 for models M_0 and M_1 respectively. Then the posterior probability of model M_1 is

$$p(M_1|Y) = \frac{p_1 E[h(\theta)|Y]}{p_1 E[h(\theta)|Y] + p_0}$$

If $E[h(\theta)]$ is small compared with 1, assuming similarity of the p_j's, the perturbation can be ignored in further analysis. On the other hand, if $E[h(\theta)|Y)]$ is large compared to 1 then a posteriori, it will have large posterior probability or at least, much larger probability a posteriori than a priori.

The unconditional posterior of θ is a weighted mixture of the initial and perturbed posteriors

$$p_u(\theta|Y) = \frac{p_1 E[h(\theta)|Y] p_h(\theta|Y) + p_0 p(\theta|Y)}{p_1 E[h(\theta)|Y] + p_0}. \quad (3)$$

where the subscript u identifies the unconditional posterior. Formula (3) is key for identifying which perturbations are important. There are three ways

for a perturbation to be unimportant so that the data analyst can ignore it. From (3), if M_1 is a priori implausible, then p_1 will be small, and assuming that $E[h(\theta)]$ is not so large to overwhelm the smallness of p_1, then the perturbation is ignorable. Presumably, if $h(\theta)$ is under consideration this has not happened. Second, suppose $E[h(\theta)|Y]$ is very small compared to 1, then the posterior probability of M_1 is small, and it can be ignored. Finally, if $p_h(\theta|Y)$ is similar to $p(\theta|Y)$, then the perturbation causes no change in conclusions, i.e. it is uninfluential and again we don't have to worry about it. The L_1 influence statistic might be used to assess influence at this point. Formula (3) shows that three features of the perturbation, a priori plausibility, support of the data and influence are essentially equally important in assessing the importance of a perturbation.

3 Example

This example comes from an experiment in pediatric pain (Fanurik, Zeltzer, Roberts and Blount 1993). Children immersed their arms in cold water; the length of time that they can tolerate the cold is a measure of pain tolerance. Children can be divided into two groups based on their preferred style of coping with the cold. Attenders paid attention to their arm, the feelings of the water and the experiment. Distracters thought about school, a favorite trip to the beach or the corner of the room. The children participated in the experiment twice at their first visit, with the second trial acting as a baseline measure. On their second visit, they again participated twice. Prior to the last trial, one of three counseling interventions occurred; either counseling to attend, counseling to distract or sham counseling without instruction. The three counseling interventions were known to have different effects on the two coping style groups. The last trial is used as the response. See Fanurik, Zeltzer, Roberts and Blount (1993) for more details. A full Bayesian analysis of this data will be published elsewhere.

The model is an ANCOVA. The residuals from the first ANCOVA displayed non-constant variance suggesting that the data be transformed. Since the baseline and the response are repeated measures, both were transformed. The transformation is unknown and was not fixed by the analysis; posterior uncertainty in the transformation was included in the conclusions. The initial model is

$$\begin{aligned} y_i^{(\lambda)} &= x_i^{(\lambda)}\gamma + Z_i^t\beta + \epsilon_i \\ \epsilon_i &\sim N(0,\sigma^2) \end{aligned}$$

where $x^{(\lambda)} = (x^\lambda - 1)\lambda^{-1}$ is the standard Box-Cox transformation of x, x_i is the baseline measure, Z_i is a vector of six indicator functions to indicate which of the 2×3 groups the child was in. Calculations were performed by sampling from the posterior. First, λ was sampled from a normal approximation to $p(\lambda|Y)$, then the appropriate conditional distribution $p(\gamma, \beta, \sigma^2|Y, \lambda)$ was used to sample the remaining parameters. A total of 2000 samples were used to do the computations.

The baseline and response are pain tolerance in seconds; the length of time that children could keep their arms in cold water. If a child kept their hand in for 240 seconds, the trial was terminated and 240 seconds was recorded as the response. Three children with missing data have been dropped from this analysis. Unfortunately in this model, accounting for the censoring is difficult. Usual censoring would just suggest that the x_i's and y_i's should be treated as random variables with appropriate distributions restricted to being greater than 240 seconds. A competing hypothesis with merit is that those children did not respond to the pain of the cold, and should and would have had shorter response times, if they had responded properly. This alternative hypothesis suggests that these extreme observations should actually be shorter. A third hypothesis suggests that these cases should just be deleted, that they are different from the other children.

The data: ID number, coping style, treatment, baseline x, and response y are given in tables 1 and 2. The ID number runs from 0 to 60, numbering cases as in Lisp. Also given is $100 * \text{CPO}$ and the L_1 influence diagnostic for deleting each observation. Within coping style and treatment groups, observations are ordered by the L_1 statistic. Three cases with small CPO are identified as outlying and influential in this analysis, cases 15, 30 and 58. These three observations were deleted as a group to give one perturbation. The cases with either x_i or y_i equal to 240 were also grouped together and deleted. Summary statistics from these perturbations are given at the bottom of Table 3 labeled cd. Table 3 also gives the results of perturbing the 240 values in the data set up to 300

ID	CS, TMT	x	y	L_1	CPO
0	1 , 1	35.31	11.71	0.284	0.714
33	1 , 1	11.92	44.72	0.263	0.325
53	1 , 1	32.84	25.21	0.076	2.263
35	1 , 1	23.29	20.67	0.074	2.747
7	1 , 1	19.03	30.37	0.073	1.959
13	1 , 1	13.47	15.98	0.073	3.605
54	1 , 1	30.66	38.47	0.064	1.672
3	1 , 1	23.41	31.38	0.059	2.041
31	1 , 1	30.84	37.03	0.059	1.775
10	1 , 1	26.3	28.64	0.050	2.307
34	1 , 2	12.99	34.76	0.182	0.807
28	1 , 2	11.42	27.44	0.138	1.423
60	1 , 2	16.5	11.12	0.128	3.252
57	1 , 2	23.18	14.16	0.126	2.694
41	1 , 2	16.44	12.63	0.106	3.506
47	1 , 2	13.41	21.19	0.067	2.784
21	1 , 2	42.22	41.44	0.060	1.599
26	1 , 2	18.13	19.33	0.059	3.197
24	1 , 2	18.77	20.34	0.057	3.078
59	1 , 2	27.61	27	0.05	2.422
19	1 , 3	**240**	116.68	**0.120**	**0.592**
2	1 , 3	10	8.27	0.118	4.800
23	1 , 3	6.24	7.13	0.114	6.300
46	1 , 3	38.85	48.42	0.113	1.050
45	1 , 3	33.54	22.65	0.075	2.538
16	1 , 3	20.03	26.82	0.071	2.197
25	1 , 3	9.63	15.28	0.071	3.778
6	1 , 3	11.05	13.86	0.070	4.241
55	1 , 3	11.19	15.51	0.067	3.834
37	1 , 3	16.87	18.88	0.059	3.286

Table 1: Data and case diagnostics, part 1: attenders. ID is Id number; CS, coping style; TMT is treatment; x is baseline tolerance; y is response tolerance; L_1 is the $L_1(h)$ influence statistic; and CPO is 100 times the conditional predictive ordinate corresponding to case deletion. Coping style is either 1=attend, 2=distract. Treatment is either 1=attend, 2=distract, or 3=sham. The baseline and response tolerance are measured in seconds. Within CS/TMT group, cases are ordered by their influence. Cases and measurements referred to in the text are in bold, along with the associated CPO and L_1 statistics.

ID	CS, TMT	x	y	L_1	CPO
27	2 , 1	10.51	22.8	0.149	1.645
49	2 , 1	52.01	20.16	0.145	1.733
43	2 , 1	12.42	8.06	0.140	4.300
56	2 , 1	**240**	104.5	**0.095**	**0.685**
4	2 , 1	85.91	60.3	0.073	1.129
50	2 , 1	14.47	14.53	0.070	4.034
44	2 , 1	12.58	15.63	0.069	3.772
17	2 , 1	17.53	21.73	0.068	2.724
36	2 , 1	49	43	0.067	1.496
11	2 , 1	23.93	20	0.059	3.101
15	2 , 2	41.72	**240**	**0.668**	**0.001**
58	2 , 2	36.43	**180.19**	**0.402**	**0.021**
9	2 , 2	11.75	13.29	0.267	1.091
1	2 , 2	24.22	20.3	0.241	0.721
22	2 , 2	44.94	35.97	0.185	0.727
38	2 , 2	25.13	31.04	0.134	1.358
52	2 , 2	**240**	**240**	**0.130**	**0.307**
48	2 , 2	42.58	48.94	0.093	1.143
8	2 , 2	41.2	78	0.068	0.935
29	2 , 2	29.51	63.12	0.063	1.131
42	2 , 2	29.35	55.27	0.051	1.314
30	2 , 3	88.89	6.67	**0.684**	**0.016**
20	2 , 3	44.16	65.42	0.265	0.191
14	2 , 3	45.41	44.31	0.140	0.927
12	2 , 3	41.2	40.78	0.133	1.041
32	2 , 3	10.12	7.62	0.098	6.543
51	2 , 3	24.51	12.19	0.091	4.081
5	2 , 3	18.95	20.35	0.088	2.549
18	2 , 3	20.29	11.89	0.083	4.512
40	2 , 3	16.75	14.66	0.069	3.987
39	2 , 3	38.89	20.9	0.061	2.961

Table 2: Data and case diagnostics, part 2: coping style = distractors. Key is the same as table 1.

seconds and down in 30 or 60 second increments to 120 seconds. Four observations 15, 19, 52 and 56 were always perturbed; for the 150 and 120 second perturbations, case 58 with $y_i = 180$ seconds was also perturbed. These perturbations are called recensoring which is abbreviated as rec in the table.

Influence analysis tells us that changing all 240 second values to 300 is roughly as influential as changing them to 180 seconds, and that as we decrease the new value to 150 and then to 120 seconds the influence increases. Adding in case 58 with its response of 180 seconds to the set of perturbed values substantially increases the influence. Outlier analysis tells us that changing the 240 second values to 180 is supported by the data by a factor of roughly $13 = 1/.077$, over the null perturbation, while the change to 300 seconds is not supported compared to the null perturbation by a factor of 8. Thus the data support decreasing the times of these pain tolerant children.

Deleting the 4 observations gives a Bayes factor of 2e-12, while deleting the 3 outliers gives a Bayes factor of 4e-14, which is apparently supported by an extra factor of 200. Unfortunately, the case deletion CPO's are not easily compared to the CPO's of other perturbations because of scaling problems, and multiple case deletions are not comparable to single case deletions for the same reason. A solution to this is based on realizing that most observations must not be outliers. The 25^{th} percentile of the single observation CPO's is .01, which is a convenient round number to use in further computation. Thus the suggestion is that CPO for individual cases be multiplied by a factor of 100 before comparing it to the null perturbation, and that CPO for perturbations that correspond to deleting k cases be multiplied by 100^k. After this adjustment, the recensoring perturbation that changes 240 to 120 seconds is about equally outlying and equally influential as the perturbation which deletes the 4 cases with times of 240 seconds. In contrast, deleting the 3 outliers is much more influential, and the adjusted CPO is $4e-8$, almost 10^4 smaller than the other perturbations considered. Further analysis in this data set should consider the impact of deleting the 3 outliers.

Perturb	No. of cases	new value	L_1	CPO
rec	4	300	.3	8.12
rec (null)	4	240	0	1
rec	4	180	.365	.077
rec	4	150	.511	.022
rec	4	120	.597	.0090
rec	5	150	.648	.0044
rec	5	120	.805	.00028
cd	3		.944	3.94e-14
cd	4		.664	2.32e-12

Table 3: Summary of perturbation results. Includes all perturbations considered except single case deletion. Perturb stands for type of perturbation, rec=recensoring, cd=case deletion, null=no change; Recensoring is the change of long tolerances to a different value. The number of cases involved: 4 cases indicates cases 15, 19, 52 and 56; 5 cases indicates cases 15, 19, 52, 56 and 58; and 3 cases indicates cases 15, 30 and 58; new baseline or response values for changed values; L_1 influence statistic; CPO outlier statistic. The recensoring with a changed value of 240 is the null perturbation.

4 Discussion

Past use (Geisser 1987, Pettit and Smith 1987, Pettit 1990) of the individual case CPO uses an internal norming to try to identify outliers. That is, the CPO are compared amongst themselves, without comparison to the null model. In contrast, in this paper I would like to be able to at least roughly compare CPO from case deletion to CPO from other perturbations. A procedure can be done for the delete three and delete four perturbations, sampling sets of three or four observations and calculating their CPO. Doing this for the delete four perturbation leaves the observed value of deleting the four cases as roughly the 12^{th} percentile of the CPO's of the sets of four. In contrast, the delete three perturbation CPO's are substantially more outlying than any randomly selected subset of three cases.

That CPO is not absolutely interpretable should be clear in linear regression. The statistic CPO_i is equal to the expected value of $(2 * \pi\sigma^2)^{.5} \exp(.5 * \sigma^{-2} * (y_i - x_i\beta)^2)$. Multiplying all y_i in the regres-

sion by k will increase σ to $k\sigma$, without changing how we should feel about case i being an outlier. In the current transformation model, there is an additional factor of $y_i^{1-\lambda}$ due to the Jacobian of the Box-Cox transformation. Thus there is an additional choice that needs to be made. In the current example, it was decided to analyze on the seconds scale since seconds are more interpretable than log seconds or square root seconds.

References

[1] Cook, R. D. (1986). Assessment of local influence. *JRSS-B*, 48, 133-155.

[2] Fanurik, D., Zeltzer, L. K., Roberts, M. C. and Blount, R. L. (1993). The relationship between children's coping styles and psychological interventions for cold pressor pain. *Pain*, in press.

[3] Geisser, S. (1980). Comments on "Sampling and Bayes' inference in scientific modeling and robustness" *JRSS-A*, 143, 416- 417.

[4] Geisser, S. (1987). Influential observations, diagnostics and discordancy tests. *Journal of Applied Statistics*, 14, 133-142.

[5] Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. *Statistics and Probability: Essays in Honor of C. R. Rao*, pp. 343-358. Elsevier/N. Holland.

[6] Kass, R. E. and Raftery, A. E. (1993). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, to appear.

[7] Kass, R. E., Tierney, L. & Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, 76, 663-674.

[8] Pettit, L. I. & Smith, A. F. M. (1985). Outliers and Influential Observations in Linear Models. In *Bayesian Statistics 2*, Eds. J. M. Bernardo, M. DeGroot, D. Lindley, and A. F. M. Smith, pp. 473-94 Amsterdam: North Holland.

[9] Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *JRSS-B*, 52, 175-184.

[10] Weiss, R. E. (1993). Influence assessment using divergence measures. Submitted for publication.