# An empirical Bayes approach to polynomial regression under order restrictions

By LEONARD W. DEATON

Summary

An unknown polynomial is to be estimated over a finite interval from $N$ independent, normally distributed observations. A prior distribution is placed on the polynomial coefficients expressing the opinion that the coefficients decrease in absolute value as the degree of the corresponding terms increase. The data are used to estimate the parameters in the prior distribution of the coefficients. A Monte Carlo study is presented which compares the proposed method with the lack-of-fit procedure. This study indicates that the proposed method performs well in terms of minimizing a squared error loss as well as in yielding the correct degree of the polynomial being estimated.

*Some key words*: Bayes rule; Empirical Bayes procedure; Isotonic regression; Least squares estimate; Maximum likelihood procedure; Monte Carlo study; Order restriction; Orthonormal; Polynomial regression.

## 1. Introduction

There are many methods for polynomial regression. Most classical methods are well known. Bayesian approaches have been devised by Guttman (1967), Halpern (1973) and Young (1977). Hager & Antle (1968) studied Guttman's method for determining the degree of a polynomial and concluded that it was not of practical value. They also recommended that future approaches to this problem be compared to the lack-of-fit procedure. Halpern made such a comparison which indicated that his method, using a vague prior on the parameters, was of practical value. However, unless a vague prior is used on the parameters, Halpern's method appears to be computationally cumbersome.

Young's procedure is not designed for determining the correct degree of the polynomial, but is designed only for optimal prediction. Hence, Young's procedure always yields a polynomial of maximal degree. Young's procedure requires numerical methods to approximate a mode.

The initial assumptions of the method proposed here closely resemble those of Young. However, we also attempt to determine the correct degree of the polynomial being estimated. As recommended by Hager & Antle, we have compared our procedure with the lack-of-fit procedure. The results of this comparison are presented in § 4. The computations required for our method are simple and exact. A numerical example is given in § 5.

## 2. The model

We are to estimate the polynomial function $P$. We have $N$ independent observations $Y_i$ of $P$ at the points $x_i$ of the form $Y_i = P(x_i) + \varepsilon_i$ $(i = 1, \ldots, N)$, where the $\varepsilon_i$ are normally distributed with mean zero and unknown variance $\sigma^2$. We write $P$ as the sum of orthonormal

polynomials, that is

$$P(x_i) = \sum_{j=0}^{m} \theta_j \psi_j(x_i) \quad (i = 1, \ldots, N),$$

where $\psi_j$ is a polynomial of degree $j$ such that

$$\sum_{k=1}^{N} \psi_i(x_k) \psi_j(x_k) = \delta_{ij} \quad (i, j = 0, \ldots, m);$$

the coefficients $\theta_j$ are unknown. This assumes that we have at least $m+1$ distinct values of the $x_i$'s. By defining the $N \times (m+1)$ matrix $Q$ with the element of the $i$th row and $j$th column as $\psi_j(x_i)$, we obtain $Q'Q = I$, where $I$ is the identity matrix.

Our assumptions thus far may be expressed in matrix notation as

$$Y \mid \theta \sim N(Q\theta, \sigma^2 I).$$

That is, we observe an $N$ dimensional random vector $Y$ which given the $m+1$ dimensional vector $\theta$ has a multivariate normal distribution with mean $Q\theta$ and covariance matrix $\sigma^2 I$.

The least squares estimator $\hat{\theta}$ for $\theta$ and the error sum of squares $s$ are independent sufficient statistics for the problem. Hence, the components $\hat{\theta}_i$ given $\theta$ are independently normally distributed with mean $\theta_i$ and variance $\sigma^2$, and are independent of $s$ given $\theta$, which is such that $s\sigma^{-2}$ is chi-squared with $n$ degrees of freedom. That is

$$\hat{\theta}_i \mid \theta \sim N(\theta_i, \sigma^2), \quad s/\sigma^2 \mid \theta \sim \chi^2_n, \tag{2·1}$$

where $n = N - m - 1$.

We put a prior distribution on $\theta$ and assume that its components $\theta_i$ are independently normally distributed with mean 0 and variance $\sigma_i^2$. We allow $\sigma_i^2 = 0$, that is, some components of $\theta$ may be degenerate at zero.

If a zero mean would contradict our prior opinion for some of the $\theta_i$ we may use something other than zero. Then, the procedures given here would only require a slight adjustment. If one has a vague opinion about $\theta_i$, then a sufficiently large value of $\sigma_i^2$ will negate the effect of assigning a prior mean of zero to it.

Hence,

$$\theta_i \mid \hat{\theta}_i \sim N\{(1 - z_i)\hat{\theta}_i, z_i \sigma_i^2\}, \tag{2·2}$$

where

$$z_i = \sigma^2/(\sigma^2 + \sigma_i^2).$$

Also, the marginal distributions of the $\hat{\theta}_i$ are independent normal distributions with mean 0 and variance $\sigma^2/z_i$. Hence

$$\hat{\theta}_i \sim N(0, \sigma^2/z_i). \tag{2·3}$$

As a result of (2·1) and (2·3) the joint marginal distribution of the $\hat{\theta}_i$ and $s$ is proportional to

$$\frac{s^{\frac{1}{2}(n-2)}}{\sigma^n} \exp\left(-\frac{s}{2\sigma^2}\right) \prod_{i=0}^{m} \frac{z_i^{\frac{1}{2}}}{\sigma} \exp\left(-\frac{z_i \hat{\theta}_i^2}{2\sigma^2}\right). \tag{2·4}$$

The mean of the distribution (2·2) will provide us with the Bayes rule for estimating $\theta_i$ when $\sigma^2$ and $\sigma_i^2$ and hence $z_i$ are known. We proceed in a manner somewhat similar to that of Efron & Morris (1973) and use the data to estimate the $z_i$. However, this procedure differs from theirs in that we shall not assume the $z_i$ are all equal nor shall we use a loss function in obtaining our estimates. Indeed, the process of selecting an appropriate model for regression is accomplished by estimating certain $z_i$ to be 1.

We eventually express complete vagueness in our prior opinion of $\theta_0$, the constant term of the polynomial, by taking $\sigma_0^2 = \infty$. Hence, we use the least squares estimator $\hat{\theta}_0$ to estimate $\theta_0$. The theory at this point will not allow such an assignment. So, temporarily we assume $\sigma_0^2$ is fixed at some large positive value. We also assume that

$$\sigma_1^2 \geqslant \ldots \geqslant \sigma_m^2 \geqslant 0. \tag{2·5}$$

Young (1977) also assumes (2·5) and uses a vague prior on $\theta_0$. The constraint (2·5) reflects a prior opinion that becomes increasingly stronger, as the index $i$ increases, that $\theta_i$ is near zero. The assumptions in (2·5) can be relaxed in varying degrees to the point of being eliminated entirely. However, we believe that (2·5) is appropriate for most practical problems.

In terms of the $z_i$ our assumptions are

$$z_0 = \varepsilon \quad (0 < z_1 \leqslant z_2 \leqslant \ldots \leqslant z_m \leqslant 1), \tag{2·6}$$

where $\varepsilon$ is a known positive number near zero.

Although estimates of the $z_i$ are enough to give us an estimate of $\theta$, it is both practical and convenient also to estimate $\sigma^2$. One way to do this is to use a maximum likelihood procedure and select $\sigma^2$ and the unknown $z_i$ to maximize (2·4) subject to the restrictions in (2·6). First, we make the transformation

$$V_i = z_i \sigma^{-2}, \quad V_{m+1} = \sigma^{-2} \tag{2·7}$$

for $i = 1, \ldots, m$. Then (2·4) may be rewritten as

$$s^{\frac{1}{2}(n-2)} z_0^{\frac{1}{2}} V_{m+1}^{\frac{1}{2}(n+1)} \exp\{-\tfrac{1}{2}V_{m+1}(s + z_0 \hat{\theta}_0^2)\} \prod_{i=1}^{m} \{V_i^{\frac{1}{2}} \exp(-\tfrac{1}{2}V_i \hat{\theta}_i^2)\}. \tag{2·8}$$

## 3. Estimating the hyperparameters

We could estimate the hyperparameters $V_i$ by selecting them to maximize (2·8) subject to the restrictions in (2·6). In terms of the $V_i$, (2·6) becomes

$$z_0 = \varepsilon \quad (0 < V_1 \leqslant V_2 \leqslant \ldots \leqslant V_m \leqslant V_{m+1}). \tag{3·1}$$

Thus (2·8) could be maximized subject to (3·1). But we prefer to put a prior on $(z_1, \ldots, z_m, V_{m+1})$ which is proportional to

$$V_{m+1}^{\gamma_{m+1}-1} \exp\left(-\frac{V_{m+1}}{\beta_{m+1}}\right) \prod_{i=1}^{m} z_i^{\gamma_i - 1} \tag{3·2}$$

for the $z_i$ satisfying the restrictions in (2·6) and $V_{m+1} > 0$.

Apart from the restrictions in (2·6) we see that (3·2) is a product of independent beta distributions and a gamma distribution. The choice $\gamma_i = 1$ gives a uniform distribution of $z_i$ and larger values of $\gamma_i$ express stronger opinions that the $z_i$ are near 1. In testing the hypothesis that $\theta_i = 0$, in the classical sense, we essentially express a prior opinion that $\theta_i = 0$ and will stay with that opinion unless sampling evidence is sufficiently strong to reject the hypothesis. Apart from the restrictions in (2·6), we believe that selecting values of $\gamma_i$ larger than 1 is in spirit similar to selecting significance levels less than 50% in testing the hypothesis that $\theta_i = 0$.

The posterior distribution of $(z_1, \ldots, z_m, V_{m+1})$ given $\hat{\theta}$ and $s$ is proportional to the product of (3·2) and (2·8) which can be written as

$$V_{m+1}^{\frac{1}{2}\tilde{n}} \exp(-\tfrac{1}{2}V_{m+1} W_{m+1}) \prod_{i=1}^{m} \{V_i^{\frac{1}{2}(2\gamma_i - 1)} \exp(-\tfrac{1}{2}V_i W_i)\}, \tag{3·3}$$

where

$$\bar{n} = n+1+2\left\{(\gamma_{m+1}-1)-\sum_{i=1}^{m}(\gamma_i-1)\right\},$$

$$W_{m+1} = s+2/\beta_{m+1}+z_0\hat{\theta}_0^2, \quad W_i = \hat{\theta}_i^2 \quad (i = 1, \ldots, m)$$

provided that (3·1) is satisfied.

Our problem is to select $V_i$ to maximize (3·3) subject to the restrictions (3·1). If (3·3) is maximized by taking $V_i = \hat{V}_i$, then we would use (2·7) to solve for the estimates $\hat{z}_i$ of $z_i$ so that our final estimates of $\theta_i$ are $\hat{\theta}_i(1-\hat{z}_i)$. Our estimate of $\theta_i$ is zero provided our estimate $\hat{z}_i$ is one. This occurs provided $\hat{V}_i = \hat{V}_{m+1}$.

We now define $g_i$ by

$$g_{m+1} = \bar{n}/W_{m+1}, \quad g_i = (2\gamma_i-1)/W_i \quad (i = 1, \ldots, m). \tag{3·4}$$

If we ignore the restrictions in (3·1) and if the $g_i$ are positive, then taking $V_i$ to be $g_i$ for $i = 1, \ldots, m+1$ maximizes (3·3). If such $V_i$ satisfy the restrictions in (3·1) our problem is solved. The following theorem gives us a more general solution.

THEOREM 1. *If the $g_i$ in* (3·4) *are positive, then* (3·3) *is maximized by taking $V_i$ to be the isotonic regression of $g_i$ with weights $W_i$ for $i = 1, \ldots, m+1$.*

The proof is too long to be included here. For an account of isotonic regression, see Barlow *et al.* (1972, p. 9).

The next theorem gives a formula for computing the isotonic regression.

THEOREM 2. *Let $\hat{V}_i$ be the isotonic regression of $g_i$ with weights $W_i$ for $i = 1, \ldots, m+1$. Then*

$$V_i = \max_{s\leqslant i}\min_{t\geqslant i}\mathrm{Av}\,(s,t) \quad (i = 1, \ldots, m+1),$$

*where*

$$\mathrm{Av}\,(s,t) = (\Sigma g_r W_r)/(\Sigma W_r)$$

*where both sums are from $r = s$ to $r = t$.*

The formula for $\hat{V}_i$ in Theorem 2 is called a max–min formula and is given by Barlow *et al.* (1972, p. 19).

It is rather difficult to make any simple statement as to how many coefficients $\hat{\theta}_i$ will be eliminated by this procedure. If a given coefficient is eliminated, then all coefficients of higher degree orthonormal polynomials will also be eliminated. As for those coefficients which are left in, the shrinkage factors increase with the degree of the polynomial. Roughly speaking, a coefficient $\hat{\theta}_i$ is eliminated provided $g_i$ is 'significantly' larger than $g_{m+1}$. The previous statement is an oversimplification since all $g_i$ and their weights $W_i$ must be taken into account. For maximum elimination, we would want $g_{m+1}$ small and all other $g_i$ large. As shown in the Monte Carlo study in § 4, it seems difficult to overeliminate when some elimination is required. In that study $g_{m+1}$ was essentially zero and excellent results were obtained.

The assumption in Theorem 1 that the $g_i$ are positive is met provided that

$$\sum_{i=1}^{m}\gamma_i < \tfrac{1}{2}\{n-1+2(m+\gamma_{m+1})\}, \quad \gamma_i > \tfrac{1}{2} \quad (i = 1, \ldots, m).$$

The requirement that $\gamma_i > \frac{1}{2}$ for $i = 1, \ldots, m$ does restrict our ability to express a strong prior

opinion that any $z_i$ is near zero. However, if we have a very strong prior opinion that $z_i$ is very near zero, we can simply take $z_i = 0$ as we do with $z_0$. It would seem rather rare that Theorem 1 could not be applied in any practical case.

By comparing (3·3) to (2·8) we see that the problem of maximizing one is equivalent to maximizing the other. In fact, if in (3·3) and (3·4) we take $\beta_{m+1} = \infty$ and the $\gamma_i = 1$ the problems are equivalent.

We now consider the problem of expressing complete vagueness in our prior opinion of $\theta_0$. We note that there is no mathematical difficulty encountered with simply taking $z_0 = \varepsilon = 0$ in (3·1), (3·3) and (3·4). An appropriate continuity result justifies the process of taking $z_0 = 0$ when selecting the $V_i$ to maximize (3·3) subject to (3·1). This procedure can be applied to express vagueness for any $\theta_i$. For example, if we wanted our estimate to be at least a quadratic, we could take $z_0 = z_1 = z_2 = 0$.

## 4. Monte Carlo studies

The Monte Carlo study consisted of adding a $N(0,1)$ deviate to a polynomial $P(x)$. Two observations were made at each of the seven points $x = 0, \pm 1, \pm 2$ and $\pm 3$. With these 14 observations the following estimates of $P(x)$ were computed:

(i) Hager & Antle's (1968) lack-of-fit test using a 5% level of significance for each $F$ test.

(ii) The isotonic regression rule proposed in this paper in which the assumed restrictions on the $V_i$ were $V_1 \leqslant \ldots \leqslant V_{m+1}$. The parameters in (3·2) were selected to express as strong an opinion as possible that the $z_i$ were near 1 while still keeping the $g_i$ positive. In fact

$$\gamma_1 = \gamma_2 = 1, \quad \gamma_3 = 1{\cdot}5, \quad \gamma_4 = \gamma_5 = 2{\cdot}0, \quad \gamma_6 = 2{\cdot}49999999, \quad \gamma_7 = 1, \quad \beta_7 = \infty, \quad z_0 = 0.$$

The strange selection for $\gamma_6$ is because, once the others were selected, $g_7$ is positive provided $\gamma_6 < 2{\cdot}5$. This selection of the parameters will almost certainly eliminate the sixth degree term of the polynomial unless the sixth degree fit is essentially perfect.

(iii) This is the same as (ii) except that null values for the parameters were used. That is, all $\gamma_i = 1$, $\beta_7 = \infty$ and $z_0 = 0$. This expresses a vague prior opinion on the $z_i$.

Thus the rules (ii) and (iii) provide two extreme prior opinions on the $z_i$.

To compare the accuracy of each estimate $\hat{P}(x)$ of $P(x)$, the loss function $L(\hat{P}, P)$ was used where $L(\hat{P}, P) = \frac{1}{6}\int\{P(x) - \hat{P}(x)\}^2\,dx$ and the limits on the integral were $-3$ and 3. Hence, for each estimate, a loss was observed. The process was repeated for a total of 121 observations of the loss for each rule. Also observed was the number of times in the 121 trials each rule yielded the correct degree of $P(x)$.

The entire process was repeated for nine different polynomials $P(x)$ which ranged from degree two to degree four. For each rule it was assumed that the degree of the polynomial was known to be between one and six inclusive. The results are summarized in Table 1. For example, in case 3 the coefficient of the orthonormal polynomial of degree zero was 4, of degree one was 10, of degree two was 20 and zero for the others. Thus case 3 dealt with a quadratic polynomial. The smallest average loss was obtained by the rule (ii) which was 0·1874. The lack-of-fit rule yielded the correct degree 116 times in 121 trials. Although not shown in the table, the average loss for the least squares estimator of the full model, a sixth degree polynomial, was also computed and was 0·8069 for every case.

One can see that in terms of loss, (ii) did better than lack-of-fit in every case. In terms of degree, (ii) beats lack-of-fit in 6 out of the 9 cases. Method (iii) beats lack-of-fit in 4 cases in terms of degree and in 4 cases in terms of loss.

The five cases numbered 3, 4, 5, 7, 9 have coefficients that appear rather unlikely on the basis of the prior assumptions for the rules (ii) and (iii). Of those, only in two cases did lack-of-fit outdo rule (ii) in terms of degree.

Table 1. *Monte Carlo results*

| Case | Actual $\theta$ | Number of times correct degree (i) | (ii) | (iii) | Average loss $\times 10^4$ (i) | (ii) | (iii) |
|---|---|---|---|---|---|---|---|
| 1 | 5, 2, 1, 0, 0, 0, 0 | 6 | 63 | 21 | 2592 | 2110 | 3370 |
| 2 | 15, 0·5, 0·05, 0, 0, 0, 0 | 1 | 30 | 12 | 1683 | 1320 | 2484 |
| 3 | 4, 10, 20, 0, 0, 0, 0 | 116 | 95 | 36 | 2067 | 1874 | 3306 |
| 4 | 10, 10, −20, 30, 0, 0, 0 | 117 | 109 | 47 | 2655 | 2392 | 3932 |
| 5 | 10, 0, 0, 3, 0, 0, 0 | 68 | 108 | 47 | 6300 | 2738 | 3999 |
| 6 | 10, 1, 2, −2, 0, 0, 0 | 35 | 100 | 47 | 7102 | 3371 | 4188 |
| 7 | 1, 1, 2, 3, 4, 0, 0 | 108 | 116 | 62 | 5330 | 3643 | 4691 |
| 8 | 1, 5, 4, 3, 2, 0, 0 | 44 | 93 | 59 | 9075 | 5109 | 5090 |
| 9 | 1, 50, 25, 50, 10, 0, 0 | 116 | 116 | 62 | 3810 | 3207 | 4759 |

(i) Lack-of-fit; (ii) and (iii), isotonic regression rules.

The rules (ii) and (iii) might have been improved by using something other than vague knowledge on $\sigma^{-2}$ or $V_{m+1}$. Since there were repeat measurements available, we could substitute for the error sum of squares $s$, the sum of squares for pure error (Draper & Smith, 1966, p. 26). This may improve the rules (ii) and (iii), since $g_{m+1}$ as given in (3·4) is used as an initial estimate for $V_{m+1} = \sigma^{-2}$. These ideas have not been tested in Monte Carlo studies.

## 5. A numerical example

As a particular example we consider one of the results from case 6 of the Monte Carlo study. The computations required to compute the estimate of $\theta$ using rule (ii), as described in § 4, are given in Table 2. The table is composed of two parts. In the top part the index on the variables runs from zero to six. In the bottom part, the index on the variables runs from one to seven. For example, the estimate of $\theta_2$ given by rule (ii) is 2·5259 while the value of $\hat{V}_2$ is 0·0973.

Table 2. *Numerical example for results from Case* 6 *of Table* 1, *application of method* (*ii*)

| | | $i = 0$ | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ | $i = 5$ | $i = 6$ |
|---|---|---|---|---|---|---|---|---|
| Actual | $\theta_i$ | 10 | 1 | 2 | −2 | 0 | 0 | 0 |
| Least squares | $\hat{\theta}_i$ | 9·8422 | 1·3619 | 2·9031 | −2·1913 | 1·0629 | −0·5344 | 0·2356 |
| | $\hat{Z}_i$ | 0 | 0·1299 | 0·1299 | 0·2782 | 1 | 1 | 1 |
| Rule (ii) | $\hat{\theta}_i$ | 9·8422 | 1·1850 | 2·5259 | −1·5816 | 0 | 0 | 0 |

| | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ | $i = 5$ | $i = 6$ | $i = 7$ |
|---|---|---|---|---|---|---|---|
| $W_i$ | 3·7096 | 16·8556 | 9·6034 | 2·2596 | 0·5712 | 0·1110 | 10·4187 |
| $g_i$ | 0·2696 | 0·0593 | 0·2083 | 1·3277 | 5·2525 | 36·0380 | $19 \times 10^{-10}$ |
| $\hat{V}_i$ | 0·0973 | 0·0973 | 0·2083 | 0·7485 | 0·7485 | 0·7485 | 0·7485 |

The value $W_8$ is the residual sum of squares for the full model or $s$. For $i = 1, \ldots, 7$, $W_i = 2\hat{\theta}_i^2$. This is different from formula (3·3). The doubling of the weights is a direct result of making two observations at each point instead of one. Thus, the assumption on $\hat{\theta}_i$ in (2·1) must be modified by replacing $\sigma^2$ with $\frac{1}{2}\sigma^2$. Similar modifications must be made for those parts of the

remaining formulae in which $\sigma^2$ appears through an assumption regarding the $\hat{\theta}_i$. However, where $\sigma^2$ appears as a direct result of the assumptions on $s$, the formula should be left unchanged. For example, in (2·4), we replace $\sigma$ by $\sigma/\sqrt{2}$ when and only when it occurs after the product sign. The definitions of the $z_i$ are changed to $z_i = \sigma^2/(\sigma^2 + 2\sigma_i^2)$, but (2·7) would be left as it is.

The actual computation of the $\hat{V}_i$ from the $g_i$ and $W_i$ can be done in under five minutes with a hand calculator using the minimum violator algorithm of Barlow *et al.* (1972, p. 19).

The paper is based primarily on my Ph.D. thesis supervised by Professor H. D. Brunk at Oregon State University. I thank Professor Brunk and the referees for their invaluable assistance with the paper.

## References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. New York: Wiley.

Draper, N. R. & Smith, H. (1966). *Applied Regression Analysis*. New York: Wiley.

Efron, B. & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Am. Statist. Assoc.* **68**, 117–30.

Guttman, I. (1967). The use of a concept of a future observation in goodness-of-fit problems. *J. R. Statist. Soc.* B **29**, 83–100.

Hager, H. & Antle, C. (1968). The choice of the degree of a polynomial model. *J. R. Statist. Soc.* B **30**, 469–71.

Halpern, E. F. (1973). Polynomial regression from a Bayesian approach. *J. Am. Statist. Assoc.* **68**, 137–43.

Young, A. S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika* **64**, 309–17.