Biometrical Letters
Vol. 44 (2007), No. 1, 17-22

Outlying observations in genetic evaluation of chicken

Anna Wolc[1], Małgorzata Twardowska[2], Tomasz Szwaczkowski[1]

[1]Department of Genetics and Animal Breeding, Agricultural University of Poznan, Wolynska 33, 60-637 Poznan, Poland, email: tomasz@jay.au.poznan.pl
[2]Centre for Nucleus Breeding "MESSA" Ltd., Mienia, PL08-775 Cegłów, Poland

SUMMARY

A single outlying observation was excluded from the data in order to analyse its effect on goodness of fit of three egg production curves (McMillan, Ali and Schaeffer, Grossman). The adequacy of functions and analysis of residuals were obtained using several criteria. A numerical example is included.

Key words: egg production curves, outlying observations, breeding value estimation

1. Introduction

The idea of breeding value estimation originates from the fact that to some extent the offspring resemble their parents and other relatives. On that basis, phenotypic variance can be divided into genetic and environmental parts. By solving a system of so-called mixed model equations, each animal can be attributed a breeding value – the part of the genetic value that is expected to be transmitted to offspring. Due to high computational demands, until recent decades laying ability was described as a singular measurement – cumulated egg production up to a given time, e.g. 100 days. This approach, however, ignored the fact that this trait is expressed over a long trajectory of time during which both environmental and genetic effects can vary. To take this into account, the idea of using egg production curves was developed. If mathematical models are fitted to the empirical data, the problem of outlying observations often arises. This may occur in two situations: either when the function is not adequate or when single extreme observations appear. In poultry the frequency of outlying observations may be increased due to high susceptibility to changes in environmental conditions.

The objective of this study was to analyse the effects of outlying observations on the goodness of fit of models potentially useful in breeding value estimation.

2. Egg production curves and outlying observations

Several functions have been proposed to describe egg production:
The McMillan *et al.* model (1970)

$$y = a\left(1 - e^{-b(t-t_0)}\right)e^{-ct},$$

where: y is the egg production in week t, t_0 is the initial week of egg laying, a is the potential maximum daily output of eggs, b is the rate of increase in egg laying, c is the rate of decay of egg production.

Although the model was primarily developed to describe egg production of *Drosophila melanogaster*, later it was proved to fit chicken data well. The important advantage of this model is that the parameters have a biological meaning.

The Ali and Schaffer model (1987)

$$y = a + b\left(\frac{t}{n}\right) + c\left(\frac{t}{n}\right)^2 + d\ln\left(\frac{n}{t}\right) + f\left[\ln\left(\frac{n}{t}\right)\right]^2,$$

where: y is the egg production in week t, a is associated with peak production, b and c are associated with decreasing slope, d and f are associated with increasing slope, n is the number of periods. This model was chosen because it has already been implemented in genetic evaluation of dairy cattle and is relatively easy to fit due to linearity in the parameters.
The Grossman *et al.* model (2000)

$$y = r\left(\frac{yp}{t_1 - t_2}\right)\left[\ln\left(\frac{e^{t/r} + e^{t_1/r}}{1 + e^{t_1/r}}\right) - \ln\left(\frac{e^{t/r} + e^{t_2/r}}{1 + e^{t_2/r}}\right)\right] + rb_4\ln\left(\frac{e^{t/r} + e^{(t_2+P)/r}}{1 + e^{(t_2+P)/r}}\right),$$

where: y is the egg production in week t, t_1 and t_2 are the times at transition, r is the duration of transition, yp is the level of constant production, b_4 is the rate of decline in production, P is the persistency defined as the number of weeks during which constant production is maintained. The model describes egg production as a set of intersecting lines with continuous transition between the slopes.

3. Empirical data

Three egg production curves were applied to the data of 428 hens from strain M55 based on recommendations from the literature (Ananng *et al.* 2001) and preliminary analysis. Data were collected in Centre for Nucleus Breeding Ltd. Mienia in 1999 and summarized into 48 weekly average records (as the data were not collected on weekends, one week consisted of five consecutive days). An iteration method which is a combination of Gauss-Newton and Levenberg-Marquard (Dennis *et al.* 1998) was applied using the NLREG program (Sherrod, 1998). An observation in the 13th week appeared unusual with respect to other Y values; therefore the following case deletion diagnostics were applied using NLIN PROC with MARQUARDT METHOD in SAS (SAS, 2002): analysis of residuals, studentized residuals (STUD), standardized difference in fit (DFFITS). The following rules were applied (Schabenberger and Pierce, 2002): if $|\text{STUD}|>2$ the observation was treated as an outlier; if $|\text{DFFITS}|>2\sqrt{k/n}$, where k denotes the number of parameters and n the number of observations, the observation was concluded to be a highly influential point. Models were compared based on average deviation, proportion of variance explained (R^2), adjusted coefficient of multiple determination (R_a^2), and the Durbin-Watson test for autocorrelation.

4. Results and discussion

The observed values and those predicted by different models were plotted against time (Figure 1). There is fairly good agreement between the shape of observed and predicted lines, however none of the curves followed the production drop in the 13th week.

The studentized residual (Table 1) confirmed that the observation in the 13th week is an outlier and may have an influence on the overall analysis, whereas the standardized difference in fit did not exceed the critical value so it is not necessarily a highly influential point (Schabenberger and Pierce, 2002).

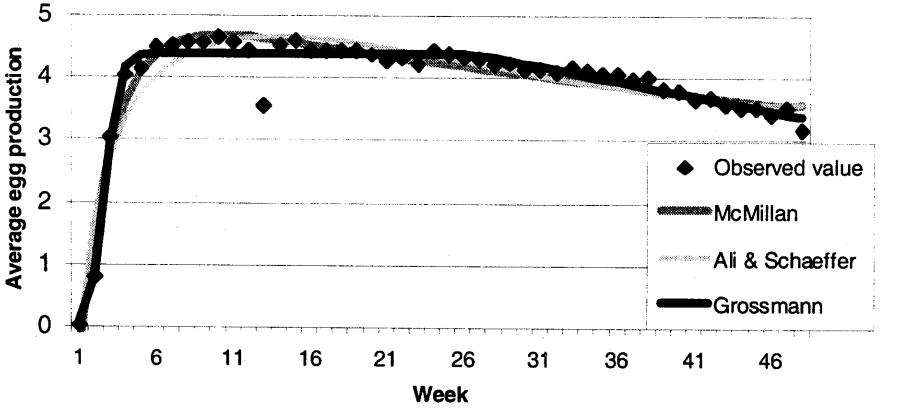

Figure 1. Observed egg production and its predictors from three egg production curves

Table 1. Case deletion diagnostics

Model	OBS	PRED	RES	STUD	conclusion	DFFITS	conclusion
McMillan et al.	3.55	4.57	-1.02	-4.23	outlier	-1.13	in range
Ali and Schaffer	3.55	4.66	-1.11	-3.66	outlier	-0.90	in range
Grossmann et al.	3.55	4.39	-0.84	-5.05	outlier	-1.14	in range

Notes on symbols: OBS - observed value, PRED - predicted value, RES - residual, STUD-studentized residual, DFFITS - standardized difference in fit

After removing the outlying observation, the goodness of fit criteria, except for the Durbin-Watson coefficient of the Grossman *et al.* model, were improved (Table2).

Table 2. Goodness of fit criteria

Model	$\bar{x}_{obs-pred}$		R^2		R_a^2		D – W	
	5 days	5 days *	5 days	5days *	5 days	5days *	5 days	5 days *
McMillan et al.	0.15	0.13	91.8	95.1	91.2	94.7	1.72	1.78
Ali and Schaffer	0.20	0.18	87.3	91.1	86.1	90.3	1.20	1.15
Grossmann et al.	0.10	0.08	96.6	98.5	95.9	98.4	1.79	1.03

Notes on symbols: $\bar{x}_{obs-pred}$ - average deviation, R^2 - proportion of variance explained, R_a^2 - adjusted coefficient of multiple determination, D –W - Durbin-Watson test for autocorrelation, * - after removal of outlying observation

The question arose as to whether this unusual observation should have been excluded from the data file or if it deserved special attention and another model should have been used to take it into account. The answer would probably depend on the purpose for which the curve was to be used. For the prediction of total production based on past record, excluding an outlier seemed reasonable as the regression line was pulled towards it and underestimated the total production. Without that observation the models gave an adequate description of the data. However in estimation of genetic parameters the situation was not that obvious. Aziz *et al.* (2002) suggested eliminating observations which deviated by more than 3 standard deviations from the mean, as well as data found to be outliers based on several criteria. On the other hand, the level of production in unfavourable conditions (which probably caused the unexpected observation) could be a point of interest for the breeder. If only total production was analysed, the fall in egg production did not equally affect the family groups, therefore the increase in variance was partly attributed to additive genetic factors which led to overestimation of overall heritability. If weekly egg production is treated as a series of repeated measurements, fixed regression models can be used.

5. Final comment

The question remains: should we not be thinking about a more flexible model that could follow the actual data rather than trying to fit data to existing models?

Acknowledge

Anna Wolc acknowledges the scholarship from Foundation for Polish Science (contract 113/2007).

REFERENCES

Ali T.E., Schaffer L.R. (1987). Accounting for covariances among test day milk yield in diary cows. Canadian Journal of Animal Science 67, 637-644.

Anang A., Mielenz N., Schüler L. (2001). Monthly model for genetic evaluation of laying hens. I Fixed regression. British Poultry Science 42, 191-196.

Aziz M.A., Schoeman S.J., Jordaan G.F. (2002). The influence of outliers on a model for the estimation of crossbreeding parameters for weaning weight in a beef cattle herd. South African Journal of Animal Science 32, 164- 170.

Belsley D.A., Kuh E., Welsh R. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York John Wiley and Sons.

Grossman M., Grossman T.N., Koops W.J. (2000). A model for persistency of egg production. Poultry Science 79, 1715-1724

McMillan I., Fitz-Earle M., Butler L.,Robson D.S. (1970). Quantitative genetics of fertility I. Lifetime egg production of Drosofila melanogaster. Genetics 65, 349-353.

SAS Institute Inc. 2002-2003. The SAS System for Linux version 9.1, Cary, NC 27513-2414 USA.

Schabenberger O., Pierce F.J. (2002). Contemporary statistical models for the Plant and Soil Sciences. CRC PRESS. 126-130.

Sherrod P.H. (1998). Nonlinear Regression Analysis Program, NLREG version 4.1. Philip H. Sherrod. Nashville. TN.

Takeuchi H. (2002). Assessment of the influence of individual observations on prediction mean square errors in variable selection problems. Journal of Japan Statistical Society 32, 43-55.