Journal of Selçuk University Natural and Applied Science
Online ISSN: 2147-3781
www.josunas.org
4(1):122-131, 2015

# Determination of Outlier in Live-Weight Performance Data of Japanese Quails (*Coturnix coturnix japonica*) by Dffits and Cook’s Distance

Hakan INCİ[1], Burhan BAHADIR[1], Ufuk KARADAVUT[2*]

[1] *Department of Animal Science, University of Bingol, Bingol, Turkey*
[2] *Department of Animal Science, University of Ahi Evran, Kırşehir, Turkey*

**Abstract**

The data on live weight performance usually show a normal distribution. On the other hand, the majority of the statistical analysis methods have been developed on the basis of the assumption that the data under consideration are normally distributed. From time to time, in the analysis of the obtained numerical values, one or more observation is found to be quite far away from other observations. Such values are named as extreme value, deviated value, and outlier value.  This study has been carried out by using Cook's distance and DFFITS criteria, with intent to determine the outliers of the live weight performance data of the Japanese quails, which has been classified according to their sex and have grown until the age of 70 days (10 weeks). In conclusion, it can be said that the DFFITS values are higher than COOK’s distance in male, female, and flock total. DFFITS method is a method that finds more outlier values when compared to COOK’s distance method in terms of finding outlier. DFFITS method should be used if precision is wanted to be enhanced, and COOK’s distance method should be preferred in less precision studies.

**Keywords**: Quail, Live Weight, Outlier Observation, Cook's Distance, Dffits.

## Introduction

The problem of outlier observation appears as a quite old problem. The data obtained from scientific studies generally show a normal distribution. However, sometimes normality gets distorted, and data are not normally distributed. However, since the majority of the statistical analysis methods have been developed on the basis of the assumption that the data under consideration are normally distributed, contradiction is slurred over. As a result of this, reliability of the studies decreases as well. Therefore, normality tests related to the data are required to be made before proceeding to the principal analyses (Bek & Efe, 1987; Akdeniz, 1998). From time to time, in the analysis of the numerical values obtained in consequence of the scientific studies, one or more observation is found to be quite far away from other observations (Bahadır et al., 2014). Outlier observation does not mean falling away from the data (Chen & Liu, 1993). Such observations are named as extreme value, irregular value, discordant observation, vague

[*] Corresponding autor. Tel.: +90 386 280 48 04  Fax: +90 386 48 32    e-mail: ukaradavut@ahievran.edu.tr

observation value, surprise value, dirty data, contaminant, outlier value etc. (Çil, 1990; Billor et al., 2000). They may arise as a result of natural randomness, and may also result from human error, machine error or similar causes (Kaya, 1999).

In some cases, there may be missing parts in our data. The presence of missing data may cause the existence of outlier values in the data. Therefore, first the missing observations should be determined, and then be analyzed accordingly. There may be several causes of missing data (Satman, 2005). The sample taken might not have given the desired answer, and might have been a sample not helpful for explanation of the subject. In other words, a wrong sample might have been selected. Since such samples would lead us to incorrect results, they are not put in process. Accordingly, the data have been taken but have not been entered into the computer. Whatever the reason might be, missing data is an undesirable situation. In fact, if the variable with missing data is our main research subject, the situation is more serious (Liu et al., 2004).

Linear regression analysis gives the best unbiased estimate through Least Squares (LS), when the assumptions that we call standard assumptions about errors are achieved. However, if there are outliers values in the data, distortions may occur in such assumptions (Aydın, 2006). For example, the normality of the error distribution may get distorted; heteroscedasticity problem may appear; and consequently, the estimations may be biased and with big variance (Rousseeuw & Zomeron, 1990). These means observations that may cause such problems in the data should be determined in order to achieve reliable estimates. If there is only one outlier value in the data set, reliable and easy-to-apply techniques are available for their definition. However, if there are more than one outlier values, sometimes they may hide the existence of each other; and even observations that do not constitute any problem in classical estimation methods may be seen as outlier values due to such outlier values (Hadi & Simonoff, 1993).

There are many statistical tests developed with intent to be able to determine whether or not observations at a point away from the average are outlier observation. Some of these tests can determine whether or not an observation is statistically outlier, whilst some others can make the same determination for more than one observation. The purpose of this study was to determine if there were outlier values in the data obtained from the quails, according to DFFITS and COOK measurement methods.

## Materials and Methods

Experiments were carried out at quail units at Poultry Units of Animal Science Departments of Agricultural Faculties of Bingol and Ahi Evran Universities. Japanese quails (*Coturnix coturnix japonica*) were used in the experiments. Live-weights were measured twice a week from the hatching until the 10th week of age with a digital balance (±0.01 g). A total of 100 quails (except for initial weight) were used and 20 measurements were performed over each one them and all the measurements were recorded separately. Experiments were carried out in two groups with 5 replications (each replication had 10 quails). That means, measurements were performed over 50 quails of each group. Experiments were performed in a battery cage. Quail grower feed (starter feed containing 23% crude protein and 3100 kcal/kg Metabolic Energy (ME) during the 1st week and grower feed containing 20% crude protein and 3250 kcal/kg ME during the following 10 weeks) for 0-10 weeks was used and ad-libitum feeding was provided. Nutrient composition of the feed ratios was prepared in accordance with NRC (Nutrient Requirements of Poultry) (1994). A total of 100 quails (of which 50 males and 50 females were selected among simultaneously hatched 150 quails after the 4th week of hatching and wing numbers) were installed to chicks after hatching. All these live-weight measurements were used to detect possible outliers. Measurements were evaluated by considering male, female and flock total live-weights.

In this study live-weight data was used to determinate outliers by using the methods Dffits and Cook measurement (SPSS, 1998 16 V package program was used for determination of outlier value).

DFFITS measure is used for the calculation of the estimated values in the new regression to be calculated, when observation i. is taken out of the data. DFFITS measure is independent from the coordinate system, in which the regression equation is established (Belsley et al., 1980). Observation corresponding to big DFFITS values can be considered to be outlier observation. The form that also considers the standard error obtained by taking out observation “i.” is shown with DFFITS. It is expressed by the equation.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}}, i = 1,2,......., n$$

where $S_{(i)}$ is expressed as follows (Hadi & Simonoff, 1993).

$$S^2_{(i)} = \frac{AKT_{(i)}}{n-k-2}$$

Besides, $h_{ii}$ represents the diagonal elements of the transformation matrix (Draper & John, 1981). Where, n and k are described as the number of observations and number of parameters respectively. Absolute value of DFFITS statistic is compared with $2\sqrt{k^{'}/n}$ . Observations with the greater $|\mathrm{DFFITS}_i|$ value among these values are determined to be outlier observations. As another advantage of these statistics, it shows us the outlier values, and on the other hand, gives us the extreme values (Belsley et al., 1980).

COOK measurement (D): It is a method intended for determining outlier value, which has been defined by Cook (1979) for the first time. Squaring the Cook distance as becoming distant from the center constitute the basis of this method. The statistics obtained by this way contains the effects seen on the whole model. Cook distance is expressed by the following equation;

$$D_i = \left[\frac{r_i^2}{k^{'}}\right]\left[\frac{h_{ii}}{1-h_{ii}}\right].\ i=1.\ 2......\ n$$

In this equation, the Cook distance is also affected by both the diagonal elements of the observation distances matrix, and the R-student-type residues, besides the number of parameters in the regression model. At this point, the calculated $D_i$ value is compared with $F_{k^{'},(n-k^{'});0.5}$ critical value, in order to determine whether or not the measure is an outlier value. Where D is greater than the critical value, the observation i. is determined to be an outlier value.

## Results and Discussion

Accordingly, the results obtained with the evaluations were given in two groups. The data obtained in consequence of the DFFITS and COOK distance methods are given in Table 1 and 2, respectively. The data obtained in consequence of the DFFITS method are shown in Figure 1. And the data obtained in consequence of the COOK distance method are shown in Figure 2 (male, female, flock total). The probabilities of identifying outlier observations depend on many factors. An increase or decrease in the number of factors makes the determination of the outlier observations easier of more difficult (Woodruff & Rocke, 1994). In general, since normal distribution tests cannot be made, outlier values cannot be determined. In fact, this is considered as a serious privation.

**Table 1**. The values obtained by DFFITS for males, females and flock totals*

| DFFITS | | | |
|---|---|---|---|
| Measurement Order | Male | Female | Flock Total |
| 1 | 0.886 | 0.761 | 0.415 |
| 2 | 0.715 | 0.816 | 0.587 |
| 3 | 0.851 | 0.413 | 0.452 |
| 4 | 0.915 | 0.286 | 0.318 |
| 5 | 0.458 | 0.317 | 0.852 |
| 6 | 0.692 | 0.588 | 0.817 |
| 7 | 0.471 | 0.744 | 0.645 |
| 8 | 0.568 | 0.851 | 0.765 |
| 9 | 0.706 | 0.717 | 0.298 |
| 10 | 0.815 | 0.971 | 0.307 |
| 11 | 0.809 | 0.983 | 0.793 |
| 12 | 0.796 | 0.812 | 0.644 |
| 13 | 0.899 | 0.413 | 0.811 |
| 14 | 0.296 | 0.374 | 0.409 |
| 15 | 0.199 | 0.506 | 0.371 |
| 16 | 0.106 | 0.514 | 0.501 |
| 17 | 0.278 | 0.651 | 0.479 |
| 18 | 0.451 | 0.703 | 0.551 |
| 19 | 0.388 | 0.549 | 0.307 |
| 20 | 0.315 | 0.308 | 0.418 |

*50 animals were weighted 20 times measured for each measurement. The averages are obtained from the 50 animals.

**Table 2.** The values obtained by Cook distance for males, females and flock totals*

| COOK | | | |
|---|---|---|---|
| Measurement Order | Male | Female | Flock Total |
| 1 | 0.356 | 0.862 | 0.647 |
| 2 | 0.475 | 0.756 | 0.359 |
| 3 | 0.361 | 0.719 | 0.384 |
| 4 | 0.681 | 0.684 | 0.648 |
| 5 | 0.429 | 0.268 | 0.369 |
| 6 | 0.473 | 0.199 | 0.521 |
| 7 | 0.631 | 0.237 | 0.645 |
| 8 | 0.891 | 0.455 | 0.391 |
| 9 | 0.504 | 0.601 | 0.294 |
| 10 | 0.388 | 0.638 | 0.168 |
| 11 | 0.608 | 0.254 | 0.674 |
| 12 | 0.721 | 0.628 | 0.359 |
| 13 | 0.744 | 0.349 | 0.378 |
| 14 | 0.394 | 0.255 | 0.348 |
| 15 | 0.538 | 0.109 | 0.911 |
| 16 | 0.671 | 0.674 | 0.362 |
| 17 | 0.442 | 0.441 | 0.541 |
| 18 | 0.196 | 0.532 | 0.593 |
| 19 | 0.264 | 0.466 | 0.752 |
| 20 | 0.388 | 0.582 | 0.269 |

*50 animals were weighted 20 times measured for each measurement. The average is obtained from the 50 animals.

The measurements made in quails were subjected to three-sided evaluation containing the evaluation of males, females and total of herd. The main reason for doing this is the fact that growth and living conditions vary depending on gender. In addition, outlier values of some statistics in much extended samples cannot be determined exactly (Gentleman & Wilk, 1975). Therefore, attention should be paid for ensuring that the appropriate sample size is in the dimensions suitable for the performance of normality tests. There is a no standard size for it, so it is left to decision.

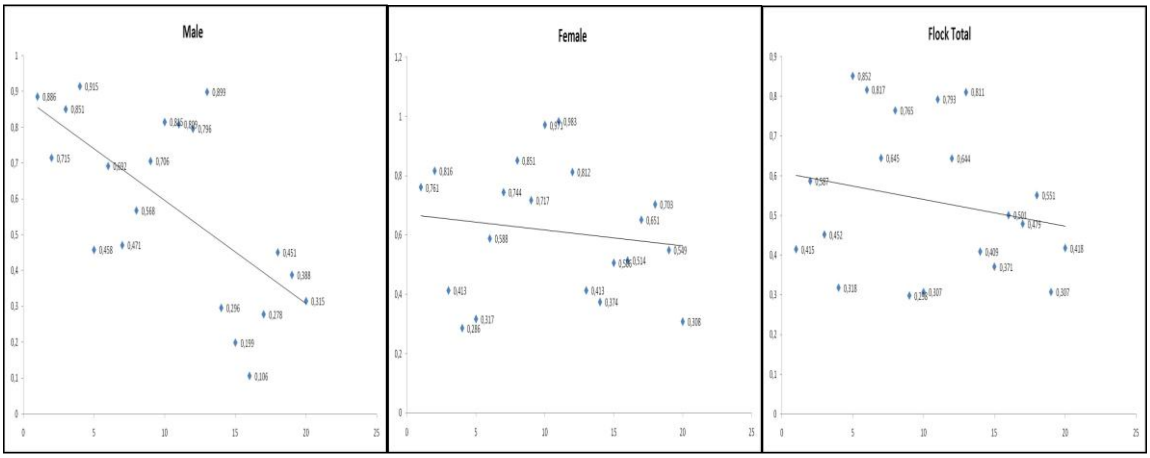

**Figure 1.** Graph of outliers obtained by DFFITS for males, females and flock total

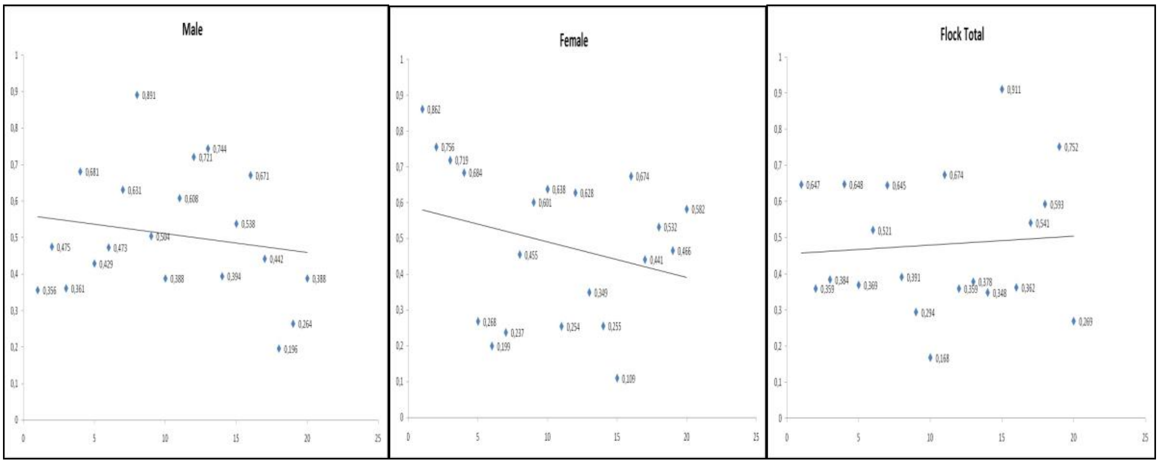

**Figure 2.** Graph of outliers obtained by COOK for males, females and flock total

DFFITS and COOK distance methods become successful in small-sized samplings for outlier observations in the direction of x. Although COOK distance seems to be successful in deviations in the direction of x, its ratio for identifying clear observations as outlier observation was found to be very high. These results are in parallel with the findings of Satman (2005) and Aşıkgil (2006).

When the tables are analyzed, how a change the measurement values have undergone is seen. When analyzed in terms of the average values in particular, changes are seen between the average values of the males, females and total of herd. This change is known to be an expected change (Gürcan & Çobanoğlu, 2012). Some researchers state that this can be said for not only growth and development but also for some other features (Oğuz, 1994; Saatçi et al., 2005). The average of the values obtained in consequence of DFFITS was determined to be 0.5807, 0.6138 and 0.5370 for males, females and herd respectively (Table 1). When the values obtained in consequence of COOK distance test were observed, the average values were determined to be 0.5057, 0.4578 and 0.4692 for males, females and herd respectively (Table 2). When an evaluation was made for males, the values were found to be 0.5807 and 0.5057 for DFFITS and COOK distance respectively. Accordingly, the DFFITS values were higher than COOK distance values. When an evaluation was made for females, the values were found to be 0.6138 and 0.4578 for DFFITS and COOK distance respectively. Accordingly, DFFITS values were higher, and COOK distance values were lower.

When an evaluation was made for herd average, the values were found to be 0.6138 and 0.4692 for DFFITS and COOK distance respectively. Accordingly, DFFITS values were higher again, as in males and females. Outlier observations obtained in consequence of DFFITS can be expressed as follows: 1st, 13th, 14th, 15th, 16th and 17th observation values for males; 4th, 10th, and 11th observation values for females; and 5th and 9th observation values for herd average were determined to be outlier values. Outlier observations obtained in consequence of COOK distance can be expressed as follows: 8th, 18th, and 19th observation values for males; 1st, 2nd and 15th observation values for females; and 10th and 15th observation values for herd average were determined to be outlier values. In recognition of finding outlier value, DFFITS was the method that found more outlier values, when compared to COOK distance method. The method required to be used can be determined according to the precision of the work to be carried out. If the precision is wanted to be increased, DFFITS method is recommended. However, if precision is not required to be cared about, COOK distance method may be chosen.

Outlier observation has been a research subject for a long time. Previously, the subject of what type observations in one-dimensional data may be outlier has been considered. Afterwards, the issue has been raised up to the level of multidimensional data, and their identification has been easier by means of the faster computers and affective algorithms. The methods developed for the diagnosis of only one outlier observation can give accurate results, when applied for all possible subsets in the data containing more than one observation. However, according as the number of observations increases in brute force techniques that function as in Gentleman & Wilk (1975) algorithm, calculation becomes impossible. There are a great number of studies on the identification of outlier observations, determination of their effects, and taking them under control. However, the studies carried out are generally on economic data (Charles & Darne, 2005). There are a limited number of studies on biological data.

## Conclusion

Outlier observation will be no longer a technical problem, with the designing of processors that can rapidly and separately analyze large sample sizes and large-scale data types (Satman, 2005). Of course, detection of the outlier observation is not adequate to solve the problem. The main problem is the determination of whether or not the outlier values are required to be included in the process, because of the fact that observations identified as outlier observation may occasionally give very valuable information. If a variation is demanded for the subject we study on; and if the variation is demanded to be as wider as possible, as a feature demanded in improvement works, outlier observations may give us very valuable information. Otherwise, exclusion of the outlier values from the process would be useful.

In case of its inclusion in the process, parametric processes to be made in data, normality assumption of which have been distorted, would lead us to incorrect conclusions. Another thing to ensure is the validity of the method that we used or planned to use in the determination of outlier value. We must distinguish between works with high precision and works with low precision. The method to be applied can be determined accordingly. An observation that seems to be outlier according to the method used in a multivariate observation might not be considered to be an outlier observation according to another method and in another work.

## Acknowledgments

This manuscript was summarized in a part of the Master's Degree Thesis. Thesis name’s “Determination of Live Weight Determination of The Performance of Outlier Data in Japanese Quails (*Coturnix coturnix japonica*)”. Thesis authors names Burhan BAHADIR- (Master of Science Student); Hakan İNCİ-(Supervisor Asst.Prof.Dr.); Ufuk KARADAVUT (Co- Supervisor Assoc.Prof. Dr.).

## References

Akdeniz, F., (1998). Olasılık ve İstatistik. Baki Kitapevi. Adana. pp 546.

Aşıkgil, B., (2006). Çoklu Doğrusal Regresyonda Aykırı. Etkili Değerlerin Araştırılması ve Bir Uygulama. Yülsek Lisanas Tezi. Mimar Sinan Güzel Sanatlar Üniversitesi Istanbul, Turkiye

Aydın, A. (2006). Grafik Yöntemlerle Etkin Gözlemlerin ve Aykırı Degerlerin Tespiti. Yüksek Lisans Tezi., 19 Mayıs Üniversitesi, Samsun,Turkiye

Bahadir, B., İnci, H., Karadavut, U., (2014). Determination of Outlier In Live-Weight Performance Data of Japanese Quails (*Coturnix coturnix japonica*) by Dfbeta And Dfbetas Techniques. Italian Journal of Animal Science 13 (1): 151-154.

Bek, Y., Efe, E., (1987). Araştırma Deneme Metotları 1. Ç.Ü. Ziraat Fakültesi Ofset ve Teksir Atölyesi. Adana. pp 395.

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

Billor, N., Hadi, A.S., Velleman, P.V., (2000). BACON: Blocked Adaptive Computationally Efficient Outlier Nominators”. Computational Statistics & Data Analysis. 34: 279-298.

Charles, A., Darne, O., (2005). “Outliers and GARCH Models in Financial Data”. Economics Letters. 86:347-352.

Chen, C. L., Liu, M., (1993). “Forecasting Time Series with Outliers”. Journal Of Forecasting. 12:13–35.

Cook, R. D., (1979). “Influential Observations in Linear Regression”. Journal of the American Statistical Association. . 74(365): 169- 174.

Çil, B., (1990). Regresyon Analizinde Tek Bir Sapan Değerin “outlier’ın” Belirlenmesine İlişkin Metodların Mukayesesi. Doktora Tezi., Ankara Üniversitesi. Ankara, Turkiye

Draper, N. R., John, J.A., (1981). Influential Observation and Outliers in Regression. Technometrics. 23(1):21-26.

Gentleman, J. F., Wilk, M. B., (1975). “Detecting Outliers. II. Supplementing the Direct Analysis of Residuals”. Biometrics. 31: 387-410.

Gürcan, E. K., Çobanoğlu, O., (2012). Japon Bıldırcınlarında (*Coturnix coturnix japonica*) Çıkım Ağırlığı ve Boyu ile Canlı Ağırlık Performansı Arasındaki İlişkiler Yüzüncü Yıl Üniversitesi Tar. Bil. Derg. 22(2): 85-90.

Hadi, A. S., Simonoff, J. S., (1993). Procedures for the Identification of Multiple Outliers in Linear Models. Journal of the American Statistical Association. 88(424): 1264-1272.

Kaya, A., (1999). Zaman Serilerinde Sapan Değerlerin Analizi Üzerine Bir Araştırma". Doktora Tezi. Dokuz Eylül Üniversitesi. Izmir, Turkiye

Liu, H., Shah, S., Jiang, W., (2004). "On-line Outlier Detection and Data Cleaning". Computers & Chemical Engineering. 28:1635–1647.

NRC, (1994). Nutrient Requirements of Poultry. Ninth Revised Edition. National Research Council National Academy Press Washington, D.C.

Oğuz, İ., (1994). Japon bıldırcınlarında (*Coturnix coturnix japonica*) canlı ağırlık için yapılan seleksiyonun bazı parametrelere etkisi. Doktora Tezi, Ege Üniversitesi. İzmir, Türkiye

Rousseeuw, P. J., Zomeren, B. C., (1990). Unmasking outliers and leverage points. Journal of the American Statistical Association. 85 (411): 633-639.

Saatçi, M., Kırmızıbayrak, T., Aksoy, A., Tilki, M., (2005). Egg weight. shape index and hatching weight and interrelationships among these traits in native Turkish geese with different coloured feathers. Turk J. Vet. Anim. Sci. 29: 353–357.

Satman, M. H., (2005). Doğrusal Regresyonda Aykırı Gözlemlerin Teşhis Yöntemleri. Yüksek Lisans Tezi., İstanbul Üniversitesi. Istanbul, Turkiye

SPSS, SPSS for Windows. (1998). Base System User's Guide, Release 9.05. SPSS Inc., Chicago. 1998.

Woodruff D. L., Rocke, D.M., (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. Journal of the American Statistical Association. 89: 888-896.