Research Paper Statistics

# Detection Outliers and Influential Observartions in Turkish Red Pine Progeny Trials in Mediterranean Region

**Semra TÜRKAN**

Hacettepe University, Department of Statistics, Beytepe, 06800Ankara, Turkey

ABSTRACT

Open pollinated seeds collected from 168 families in six seed orchards and 140 plus trees are used to establish progeny trials for low elevation Turkish red pine breeding zone (0-400m) in Mediterranean Region. Breeding values for 168 families are obtained by using linear mixed model. In previous studies on progeny trials, outliers and influential observations were removed by using 99% confidence interval prior to estimate breeding values. The aim of this study is to detect outliers significantly affect the estimation of breeding value and accordingly genetic gain of red pine families in Antalya city by using diagnostics which are proposed for linear mixed model. The genetic gain is calculated by removing outliers and influential observations as to mixed model diagnostics and confidence interval to see performance of diagnostics to detect outliers. Using diagnostics is significantly increased in genetic gain with respect to confidence interval.

**KEYWORDS** Turkish red pine; Genetic Gain; Outliers; Influential Observations; Diagnostics.

## 1. INTRODUCTION

Turkish red pine (*Pinus brutia*) is one of the important tree species in Turkey. It covers naturally 20% of Turkey's forest land. It takes the first place among the species preferred to be used for forestation activities in Turkey due to its high genetic diversity, fast growing ability, wood density and flowering at early ages. Turkish Red Pine naturally grows from sea level up to 1200 m, occasionally to 1400 m elevation in the Taurus Mountains along the Mediterranean Coast. It grows on a variety of sites with very different climatic conditions. Turkish red pine is included in 2000's to the National Tree Breeding Program (NTBP) and Seed Production Program of Turkey and National Plan for in-situ Protection of Plant Genetic Diversity which were initiated in 1994 to progeny test as plus trees. The main objective of the NTBP is to increase both the quantity and the quality of wood produced in a unit area. Highest priority has been given to Pinus brutia (Turkish red pine) in NTBP due to its characteristics appropriate for breeding. Tree breeding zone designations and plus tree selections for each breeding zone is determined and progeny tests is established on multiple sites to evaluate the genetic merits of the selected trees in progeny trials. Selection is the best way to increase the quantity and the quality of wood in tree breeding activities and selection efficiency depends on accurate the prediction of genetic values. Progeny trials have been primary concern in the genetic improvement of Turkish red pine in the mediterranean region. Hence, the main aim of progeny trials is prediction of the genetic values (Gülcü and Çelik, 2009; Öztürk vd., 2003).

To estimate breeding values of Turkish red pine families from Mediterranean low elevation region (0-400m), open pollinated seeds are collected from selected plus trees in six populations. Three progeny trials are established in Fethiye, Antalya and Ceyhan by Forest Tree Seeds and Tree Breeding Research Directorate. Each trial site had 168 families and 6 control groups. Completely randomized block design with 4 row plot configuration is used in all trials. At the end of 4th growing season, tree heights are measured. Linear mixed model is used to estimate breeding values for tree heights.

In this study, the data collected from progeny trials which are established in Antalya city is examined. Prior to estimate breeding value, it is investigated whether there are outliers and influential observations in the data by using two approach. The first approach is to use diagnostics developed for linear mixed model and the second approach is to use the confidence interval to find influential observations and outliers. In the previous studies on progeny trials, outliers and influential observations were removed by using 99% confidence interval. Hence, the goal of this study is to show that diagnostics developed for linear mixed model is better than confidence interval criteria to detect outliers. For this purpose, the estimates of breeding values and genetic gain for tree heights are calculated by removing outliers and influential observations by both approach.

## 2. MATERIALS AND METHODS

### 2.1. Experimental Design

Genetic breeding studies are established to provide the increase in yield and quality by changing the gene frequencies in the desired direction in the production populations. To accomplish this, the gene pool is constituted by genotypes having desired genes. Gene pool at the beginning of studies of genetic breeding in forest trees is selected plus trees from the forest. Plus trees are considered as trees having desired genotypes. It is unknown that if the selected trees have the desired genotype because of the only measurable value is the phenotypic values of the trees in the selection of plus tree. Whether an individual has the desired genes revealed by the breeding value which means that the total effect of genes the individual has. The genetic test method used to estimate breeding values in genetic breeding studies is progeny trials. Progeny trials constitute the most important part of breeding studies in National Tree Breeding and Seed Production Program. Progeny trials could be realized as a result of a joint operation of a large number of institutions at the different stages such as collection of pine cones, seedlings growing, the establishment of the trials, the implementation of regular maintenance and conservation work and measurement of trials. In the study conducted by Ozturk et al.(2003), red pine is given priority due to features such as lack of facilities, rapid growth, the potential for afforestation, wood materials are suitable for various use, genetic diversity is high.

This study is based on the study conducted by Ozturk et al.(2003). The data collected from the progeny trials in low elevation Turkish red pine breeding zone (0-400m) for Antalya city in Mediterranean Region are used in this study. Randomized block design was used in all trial. However, forest trees are large volume. Therefore, the number of families to be tested, block number and the area required for a block depending on the number of trees in each plot may be very large. In this case, increasing environmental variance and the increase causes covering part of the genetic variance and reduction of selection efficiency. This drawback is eliminated by sub-blocking. Hence, B type (set in rep) sub-blocking proposed by Shutz and Cockherham (1966) is used to decrease environmental effect. In B type (set in rep) sub-blocking, all the test material is divided into sets of specific families and four set are used in each block. The control material is included in each set. B type (set in rep) sub-blocking is shown in Figure 1 and the distribution of families and control material in sets of each block is given Table 1.

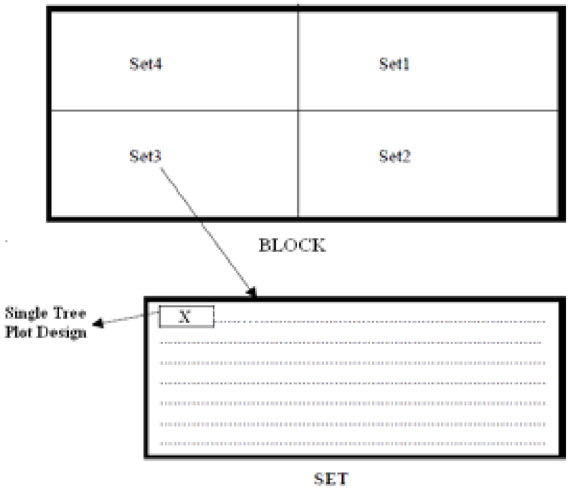

**Figure 1. B type (set in rep) sub-blocking (Alan, 2006)**

**Table 1. The distribution of families and control material in set s(Öztürk et al., 2003)**

| Population | Sets | | | | Total |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 3 | 14 | 10 | | | 24 |
| 7 | 14 | 16 | | | 30 |
| 4 | 14 | 16 | 5 | | 35 |
| 5 | | | 15 | 14 | 29 |
| 11 | | | 11 | 14 | 25 |
| 16 | | | 11 | 14 | 25 |
| Total | 42 | 42 | 42 | 42 | 168 |
| Control material | 6 | 6 | 6 | 6 | 6 |

## 2.2. Statistical Analysis

In forest tree breeding, tree breeder needs to know the best individuals that have economically important characteristics. One of the best way to know these individuals is to estimate breeding values. Many methods are used to estimate breeding values in forestry. Average, least square and weighted least square methods are used when variance is the same in progeny trials. However, standart transformation, logarithm transformation, performance level and weighted performance level methods could be used to estimate breeding values when there is variance difference in progeny trials. In these methods, the families are considered as fixed effect. Recently, BLUP method which is based on estimate both fixed effects and random genetic effects at the same time in mixed model analysis is used to estimate breeding values if families variances are different and the data is unbalanced. To estimate breeding values by BLUP maximizes probability of the sort of families effect approximate to real value and correlation between estimated breeding value and real breeding value. While breeding value is estimated by using BLUP, families are randomly taken and other effects are assumed fixed (White and Hodge, 1989). Hence, the linear mixed model which is combination of fixed effects and random effects is used to estimate breeding value by using BLUP. The estimates of fixed effect and random effect simultaneously are obtained by using BLUP method. The BLUP estimates of each family are breeding values (BV). In breeding program, genetic gain is calculated from breeding values. Control groups are considered as separate family calculating genetic gain. The formula for genetic gain is

$$\Delta G = \frac{\overline{BV_f} - BV_K}{MBV_K} x100 \qquad (1)$$

where $\overline{BV_f}$ is mean of estimated breeding values of families, $BV_K$ is estimated breeding value of control group and MBV is absolute breeding value for control group. MBV is obtained adding mean of weight to breeding value of control group.

$$y_{ijk} = \mu + \alpha_i + \theta_{j(i)} + f_k + \varepsilon_{ijk} \qquad (2)$$

where
$y_{ijk}$ is measurement of observation in the $i^{th}$ block, $j^{th}$ set and $k^{th}$ family,
$\mu$ is the overallmean;
$\alpha_i$ is the effect of $i^{th}$ block, (i=1,...,25);
$\beta_{j(i)}$ is the $j^{th}$ set effect in $i^{th}$ block; (j=1,..,4);
$f_k$ is the $k^{th}$ family effect, (k=1,..,168).

The model in Equation 2 can be expressed in matrix form

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \qquad (3)$$

where $\boldsymbol{\beta}_{(25+4+1)x1}$ is a vector of fixed unknown parameters; $\mathbf{X}_{nx(25+4+1)}$ is a known incidence matrix of the fixed effect factors associated with $\mu$, $\alpha$ and $\theta_{j(i)}$ ; $\mathbf{Z}_{nx168}$ is a known incidence matrix for the random effect factor associated with $f_k$; $\mathbf{u}_{1x168}$ is a vector of parameters (random effects); $\varepsilon$ is a nx1 vector of error terms. Usually one assumes that $u_i \sim N(0, \sigma_i^2 I)$ and this implies that $\mathbf{u} \sim N(0, \sigma_e^2 \mathbf{D})$ where $\mathbf{D}$ is block diagonal with $i^{th}$ block is $\gamma_i I_{q_i}$, for $\gamma_i = \sigma_i^2 / \sigma_e^2$ i=1,2,...,r and $\boldsymbol{\varepsilon} \sim N(0, \sigma_e^2 I)$. $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are independent. $\mathbf{y}$ has a multivariate normal distribution with $E(y) = \mathbf{X\beta}$ and $\mathbf{V} = V(y) = \sum_{i=1}^{r} Z_i Z_i' \sigma_i^2 + \sigma_e^2 I_n = \sigma_e^2 (I_n + \sum_{i=1}^{r} Z_i Z_i' \gamma_i) = \sigma_e^2 \mathbf{H}$. The estimates of $\boldsymbol{\beta}$, u and $\sigma_e^2$ are respectively,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X'H^{-1}X})^{-1}\mathbf{X'H^{-1}y}, \\ \hat{\mathbf{u}} &= \mathbf{DZ'H^{-1}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \hat{\sigma}_e^2 &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n \end{aligned} \qquad (4)$$

where $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and $\hat{\mathbf{u}}$ is BLUP of $\mathbf{u}$ (Zewotir and Galphin 2005).

Linear mixed models are considerably sensitive to outliers and influential observations. The presence of the influential observations and the outliers in data yield different inferences, or violate the assumptions of the statistical model (Schabenberger, 2004). So, being aware of outliers and influential observations is very important step in model validation. Therefore, prior to estimate breeding value by using BLUP method, outliers and influential observations in data should be removed. In previous studies on progeny trials, outliers and influential observations were removed by using 99% confidence interval. In this study, Cook's distance which is adapted to linear mixed model is used to detect influential observations that affect the estimate of variance and vectors of parameter and studentized residuals are used to detect outliers in the model. These diagnostics are, respectively,

$$\begin{aligned} CD_i(\boldsymbol{\beta}) &= (\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})'\left[Var(\hat{\boldsymbol{\beta}})\right]^{-1}(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})/p\hat{\sigma}^2 \\ CD_i(\mathbf{u}) &= (\hat{\mathbf{u}} - \hat{\mathbf{u}}_{-i})'\left[Var(\hat{\mathbf{u}})\right]^{-1}(\hat{\mathbf{u}} - \hat{\mathbf{u}}_{-i})/\hat{\sigma}_e^2 \\ CD_i(\gamma) &= (\hat{\gamma}_{-i} - \hat{\gamma})'\left[Var(\hat{\gamma})\right]^{-1}(\hat{\gamma}_{-i} - \hat{\gamma}) \end{aligned} \qquad (5)$$

$$r_i = \frac{e_i}{\sqrt{Var(e_i)}} = \frac{e_i}{\hat{\sigma}_e\sqrt{k_{ii}}}$$

where $e_i$ is $i^{th}$ residual and $\mathbf{K} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X'H^{-1}X})^{-1}\mathbf{X'H^{-1}}$ is the symmetric matrix that transforms the observations into residuals. Large values of $CD_i\_(\gamma)$, $CD_i\_(u)$ and $CD_i\_(\beta)$ show observations that affect the estimates of variance ratios, the random efects and the fixed effects. The influential observations are detected using scatter plots of $CD_i\_(\gamma)$, $CD_i\_(u)$ and $CD_i\_(\beta)$. In general, the observations which are the values of Studendized residuals, $r_i$ are not between (-2, 2) could be considered as potential outliers (Schabenberger, 2004).

## 3. RESULTS

In this study, BLUP method is used to estimate of breeding values since the data is unbalanced. Firstly, outliers and influential observations which affect substantially the estimates of breeding value are determined by using scatter plots of $CD_i\_(\beta)$, $CD_i\_(u)$, $CD_i\_(\gamma)$ and $r_i$. The scatter plots of diagnostics are shown in Figure 2

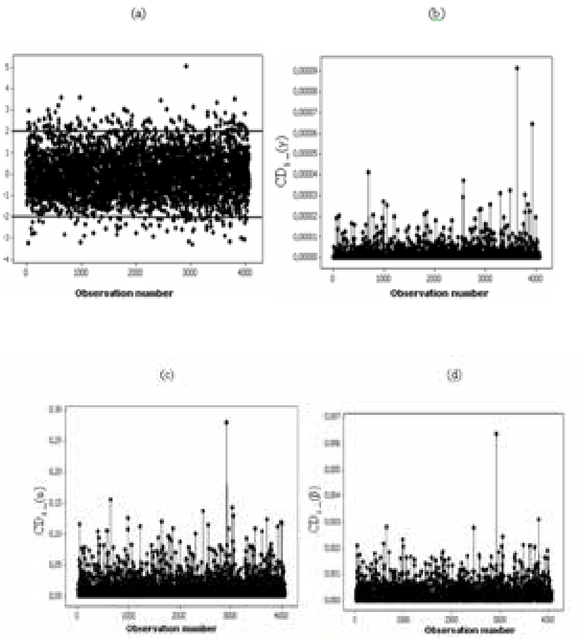

*Figure 2. (a) Scatter plot of studentized residuals ($r_i$), (b) Scatter plot of Cook distance for variance Components ($CD\_\gamma$) (c) Scatter plot of Cook distance for random effect ($CD\_u$) (d) Scatter plot of Cook distance for fixed effect ($CD\_\beta$)*

The value of studentized residuals for some observation which are not between (-2,2) as seen from Figure 2 are considered as outlier. The large values of $CD_i\_\gamma$, $CD_i\_\beta$ and $CD_i\_u$, show influential observations that affect the estimates of variance ratios, random effects and fixed effects, respectively. According to these diagnostics, outliers and influential observations are removed from the data. Then, the estimates of breeding value for each family are obtained. In parallel with, genetic gain is calculated for height growth. For once, observations which are outside of 99% confidence interval are considered as outliers or influential observations as in previous studies. With respect to %99 confidence interval, outliers and influential observations are removed from the data. Then, genetic gain is calculated for height growth, too. As the same way, The estimate of error variance ($\sigma_e^2$) and family variance ($\sigma_f^2$) are obtained. The results are given in Table 2 and Table 3.

**Table 2. Genetic Gain ( ÄG )**

| | |
|---|---|
| $\Delta G$ obtained by removing outliers using $CD_i\_\gamma$, $CD_i\_\beta$, $CD_i\_u$ and $r_i$ | 0.21 |
| $\Delta G$ obtained by removing outliers using %99 confidence interval | 0.15 |

**Table 3. Estimates of Error Variance and Family Variance Removed Outliers**

| | Error Variance ($\sigma_e^2$) | Family Variance ($\sigma_f^2$) |
|---|---|---|
| Removing outliers using $CD_i\_\gamma$, $CD_i\_\beta$, $CD_i\_u$ and $r_i$ | 736.54 | 76.66 |
| Removing outliers using confidence interval | 783.93 | 53.89 |

As seen from Table 2, the genetic gain calculated by removing outliers and influential observations as to $CD_i\_\gamma$, $CD_i\_\beta$, $CD_i\_u$ and $r_i$, , and $r_i$ diagnostics and 99% confidence interval respectively is 0.21 and 0.15. Hence, using diagnostics would lead to an increase of 6 % in the genetic gain.

As seen from Table 3, the estimate of error variance obtained after outliers and influential observations removed using confidence interval is 783.93 and removed using diagnostics is 736.54. The estimate of family variance obtained after outliers and influential observations removed using confidence interval is 53.89 and removed using diagnostics is 76.66. The decrease in the estimate of error variance or the increase in estimate of family variance implies that the composed model is estimated more correctly.

## 4. CONCLUSION

Statistically, using diagnostics to detect outliers and influential observation in progeny trails lead to significantly increase in the genetic gain as to confidence interval. In progeny trials, if genetic gain is 15-20% , economic return will be 68-260%. Hence, increase of 6% in genetic gain leads to nearly 260% economic return. If genetic gain obtained after breeding is economically insufficient, breeding studies will become unnecessary. On the other hand, pine plantations to be made with high genetic gain may become even more attractive investment vehicle. Additionally, using these diagnostics leads to the decrease in the estimate of error variance and the increase in estimate of family variance. In other words, the decrease in the estimate of error variance or the increase in estimate of family variance implies that the composed model is estimated more correctly and BLUP estimates of families are more realiable. Also, outlier and influential observations which affect estimates of fixed effect, random effect and variance parameter are determined clearly using these diagnostics.

## REFERENCES

Alan, M. (2006). Estimation of breeding values of Turkısh Red Pine (Pinus Brutia Ten.) Families in Seed Stands of Aegean Region (0-400 M). Ph. D. Thesis. Ankara University, Ankara, Turkey Gülcü, S., Çelik, S. (2009). Genetic variation in Pinus brutia Ten. seed stands and seed orchards for growth, stem form and crown characteristics. African Journal of Biotechnology, 8, 18, 4387-4394. Öztürk, H., Şıklar, S., Alan, M., Ezen, T., Korkmaz, B., Gülbaba, A.G., Sabuncu, R.,Tulukçu, M., Derilgen, S.I. (2003). Akdeniz Bölgesi alçak ıslah zonunda (0-400 m) kızılçam (Pinus brutia Ten.) döl denemeleri, Teknik Bülten, No:12. Schabenberger, O. (2004). Mixed model influence diagnostics, Data Analysis Papers, SAS Institute, Paper: 189-29. Zewotir, T., Galphin, J.S. (2005). Influence diagnostics for linear mixed models. Journal of Data Science, 3, 153-177.