ESTIMATING SEED YIELD USING LINT YIELD AND LINT PERCENTAGE OF SEED COTTON

M. F. Lopez and E. W. Elam
Research Assistant, and Associate Professor, Respectively
Department of Agricultural Economics
Texas Tech University
Lubbock, TX

Abstract

The increasing relevance of cottonseed in a farmer's revenue makes it necessary to have reliable methods to measure its yield. There are two approaches to measuring cottonseed yield. Actual cottonseed yield, obtained from a grab-sample at the ginning mill, and estimated cottonseed yield, obtained from lint yield and lint percentage of seed cotton. These two measures may differ due to various factors. The purpose of this research was to estimate seed yield from lint yield and lint percentage of seed cotton, and to determine biasedness and accuracy of the estimate. A sample of 468 experimental trials, with three Texas locations, two irrigation levels, and 15 varieties was used in the statistical analysis, which consisted of three parts. The first part deter-mined that there is a 9.1 lb. per acre bias in cottonseed yield estimation. The second part determined that variety and irrigation level were significant sources of bias. For the irrigation level effect, dryland trials showed no significant bias, while irrigated trials showed a 21.5 lb. per acre bias. The third part of the analysis showed a close relationship between actual and estimated seed yields (r = .99), and a standard error of the estimate of 78.3 lb. per acre. It was concluded that estimated cottonseed yield (from lint yield and lint percentage of seed cotton) can be used as a reliable estimator of actual cottonseed yield.

Introduction

Cotton is one of the main crops of the southern states of the US. It produces fiber (economically the most important product of this crop) and seed. Cottonseed is used as planting seed and as a source of energy, protein, fiber, and oil for animal and human consumption. Kohel (1978) pointed out that cottonseed has largely been ignored by researchers, producers, and analysts in the past years. Research on cottonseed is mostly concerned with planting seed quality because of its impact on the crop's implantation and on lint yield. Besides, little importance is given to cottonseed as part of a producer's income, even though it represents about 13 percent of the crop's value (Kinard, 1992). Parvin et al. (1978) described the influence of cottonseed quality on a producer's revenue, and how this can be increased by improving seed yield and yield. The authors encouraged economic analysts and planners to include the seed component in the decision making process. Parvin et al concluded that no changes should be made to current research programs in cotton, but they regarded as a helpful advance the development of an accurate and comprehensive testing and reporting program for cottonseed, similar to current reports on the lint component. Walker (1993) provided several reasons why he believes cottonseed usage may increase in the future and have a greater impact on a producer's revenue.

The increasing relevance of cottonseed has to be accompanied by accurate and reliable data on seed yield and quality, so producers can perceive the full benefits of their cottonseed production. There are two approaches to measur-ing cottonseed yield. The first approach is actual cottonseed yield, which is calculated from a grab-sample at the ginning mill and represents the most accurate estimation of seed yield. The second approach is to estimate cottonseed yield from lint yield and lint percentage of seed cotton, (obtained from a boll sample taken before stripping). The interest in estimating cottonseed yield is due to the fact that actual cottonseed yield is seldom reported. For example, the National Cotton Variety Trials (NCVT) is a recognized source of data on cotton lint yield and quality factors for important cotton varieties planted in the Cotton Belt; however, the NCVT does not include data on actual cottonseed yield. The 1994 issue of NCVT reports estimates of cottonseed yield (as a new item) based on lint yield and lint percentage of seed cotton. There are other sources of cotton production data provided by regional agricultural experiment stations, however, these publications do not generally report seed yield. The publication by the Texas Agricultural Experiment Station is noteworthy in that it provides data on actual cottonseed yield (along with lint yield and quality factors).

The estimated measure of cottonseed yield (based on lint yield and lint percentage of seed cotton) may differ from actual cottonseed yield responding to a number of physiological and environmental factors. Given the increasing significance of cottonseed at different levels (Parvin et al, and Kohel et al, 1978), and that most seed yield information is obtained from lint yield and lint percentage of seed cotton, the problem is that there is no statistical evidence of whether estimated cottonseed yield is a reliable (unbiased and accurate) estimator of actual cottonseed yield. The purpose of this research was to estimate seed yield using lint yield and lint percentage of seed cotton, and to determine the biasedness and accuracy of this estimate. Since there was a bias in cottonseed yield estimation (found in the first part of the statistical analysis), the second objective was to determine the specific sources of bias. Finally, the third objective was to determine the relationship between actual and estimated seed yield

---

Reprinted from the *Proceedings of the Beltwide Cotton Conference*
Volume 1:486-489 (1996)
National Cotton Council, Memphis TN

measures and the accuracy of individual estimates. If the estimation of cottonseed yield from lint percentage of seed cotton is unbiased and accurate, it can be used in further economic research, cost-benefit, and risk and return analysis that include seed yield as an extra component, using standard sources of data which only provide lint yield and lint percentage of seed cotton (e.g., NCVT). If the estimation is biased and not accurate, alternative methods to estimate cottonseed yield should be found or actual cottonseed yield should be reported.

Measuring Cottonseed Yield

Two approaches to measure cottonseed yield were analyzed and compared in this paper. One of them, actual seed yield, is measured at the ginning mill by grabbing a sample and separating lint and seed with a cotton-ginner. Then, lint and cottonseed are weighted and their yields extrapolated to a per acre basis. The second approach, estimated cottonseed yield, is based on a random, hand-picked, boll sample from the field. This sample is taken up to one week before stripping and is composed of 25 to 100 bolls. When the boll sample is gathered, lint is manually sorted from the seed and weighed along with the total of seed cotton (lint plus seed). Then, lint percentage of seed cotton is determined by dividing lint weight by total seed cotton weight. The purpose of collecting the boll sample is to determine lint percentage. However, data on lint percentage along with lint yield can be used to estimate cottonseed yield using the following formula:

$$E = LY * [(100 - L\%) / L\%] \qquad (1)$$

where E is estimated seed yield in lb. per acre; LY is lint yield in lb. per acre; and L% is lint percentage of seed cotton.

As stated in the Introduction, a bias (average difference) in cottonseed yield estimation, using lint yield and lint percentage of seed cotton, can exist. This bias can arise specifically from the sampling technique, the time span between boll-sample collection and stripping (environmental conditions at play), particular genetic characteristics of each variety, or specific management and production practices. The accuracy of individual estimates can be measured by the standard error of the estimate of actual yields versus estimated yields, for a given location, irrigation level, and variety. The significance of the bias and the accuracy of the estimate were tested later in this research.

Methods and Procedures

The data used in this research were obtained from the Cotton Performance Tests in the Texas High Plains and the Trans-Pecos Areas of Texas (1985-1993), annually published by the Texas Agricultural Experiment Station (Gannaway, et al.). This database provides both actual cottonseed yield and lint yield and lint percentage of seed cotton, making it possible to statistically compare and analyze both yield measures (i.e.: actual vs. estimated), and to determine the reliability of the estimate.

The data (468 observations) were from experimental test plots in three Texas locations (Lubbock, Halfway, and Lamesa), with two irrigation levels (dryland or irrigated) from 1985 through 1993. Fifteen varieties were chosen from the data set, using as selection criteria planted area in the South Plains (Varieties Planted, various issues) and number of years under trial. These criteria allowed the selection of the most important (widely planted) varieties during the study period. No discrimination was intended with respect to any cotton variety, and it was not the purpose of this paper to rank or qualify specific varieties.

The statistical analysis consisted of three parts. The first part tested the overall bias in cottonseed yield estimation by comparing the estimates of seed yield to actual seed yield. If results were significant (e.g., the estimate is biased with respect to the actual cottonseed yield), part two of the analysis, designed to identify the specific sources of bias in cottonseed yield estimation, would be carried out. The third part of the analysis involved the evaluation of the relationship between actual and estimated seed yields and the accuracy of individual estimates of seed yield (i.e., for a given location, variety, and irrigation level).

The variables of analysis were selected to determine whether there is bias in cottonseed yield estimation. All independent variables were set in a dummy variable system, with three possible values, 1, 0, -1. This setting would make it possible to compare the bias caused by each effect to the overall bias. Additional examples of this setting can be found in Neter, Wasserman, and Kutner (1990, chp. 20). The group of independent variables was composed of main effects and interaction effects. The main effects (location (L), irrigation level (I), and variety (V)) for the regression analysis are shown in Table 1. The selected effects summarize, within the limitations of the data, most of the potential sources of bias discussed earlier in this paper.

The interaction effects were derived from the cross-product of the two main effects, e.g., the interaction effect “location x irrigation” (L x I) was obtained by multiplying each location by each irrigation level, and in similar manner, location x variety (L x V), and irrigation x variety (I x V). The year effect was not considered in the analysis because it was assumed that the relation between L, I, and V, and their interactions, remained unchanged over the years. The dependent variable (A-E) was the difference between actual yield (A) and estimated yield (E).

The first part of the analysis consisted of a t-test to determine whether there is an overall bias in cottonseed yield estimation. The model is represented by:

$$A\text{-}E = \delta + \varepsilon \qquad (2)$$

where (A-E) is the difference between actual (A) and estimated (E) cottonseed yield; k is the overall bias (the sum of all the effects); and ε is the error term. The least squares estimator k = (A-E) represents the mean difference between actual and estimated seed yields. The calculated t-value is $t = k / s_k$, where s_k is the standard deviation of the series (A-E) divided by the square root of the number of observations. The null hypothesis was no bias in the estimation (i.e., H_0: $k = 0$), while the alternative hypothesis was bias in the estimation (i.e., H_a: $k \neq 0$). If the null hypothesis is not rejected, there would be no bias in the estimation, implying that none of the effects (L, I, and V) considered in this analysis are significant sources of bias and that estimated seed yield from lint percentage of seed cotton is an unbiased estimator of actual seed yield.

As results in part one were significant, part two of the analysis was carried out. It consisted of a series of tests to determine the specific sources of bias in cottonseed yield estimation. A regression approach for testing factor effects was used because sample sizes were unequal, and a fixed-effects analysis of variance (ANOVA) was performed because the selection of the variables was not random (Neter, Wasserman, and Kutner, 1990, chp. 20). The full regression model is represented as:

$$A\text{-}E = f(L, I, V, L \times I, L \times V, I \times V, \varepsilon) \qquad (3)$$

Equation 3 represents the bias in cottonseed yield estimation (A-E) as a function of main effects (L, I, and V), their interactions, and the error term (ε). As previously explained, if there is no bias in the estimation, the value of all effects would be equal to zero, and the difference between actual and estimated yields would only be a function of the random error term. The identification of the specific sources of bias was done by fitting the following reduced regression models (equations 4 through 9).

Test for I x V: $A\text{-}E = f(L, I, V, L \times I, L \times V)$ (4)
Test for L x V: $A\text{-}E = f(L, I, V, L \times I, I \times V)$ (5)
Test for L x I: $A\text{-}E = f(L, I, V, L \times V, I \times V)$ (6)
Test for V: $A\text{-}E = f(L, I, L \times I, L \times V, I \times V)$ (7)
Test for I: $A\text{-}E = f(L, V, L \times I, L \times V, I \times V)$ (8)
Test for L: $A\text{-}E = f(I, V, L \times I, L \times V, I \times V)$ (9)

An F-test determined the significance of each effect by measuring the increase in sum of squares error for the reduced model with respect to the full model (Neter, Wasserman, and Kutner, 1990). The calculated F-value is:

$$F = [(SSE(R) - SSE(F)) / (df_r - df_f)] / [SSE(F) / df_f] \qquad (10)$$

where SSE(R) is the sum of squares error for the reduced model (equations 4 to 9); SSE(F) is the sum of squares error for the full model (equation 3); and df_r and df_f are the degrees of freedom for the reduced and full models, respectively. In each test, the null hypothesis was that the specific effect is equal to zero (e.g., H_0: $L = 0$; $I = 0$; $V = 0$; etc.), while the alternative hypothesis was that the specific effect is different from zero (e.g., H_a: $L \neq 0$; $I \neq 0$; $V \neq 0$; etc.). If the null hypothesis is not rejected, the effect is not a significant source of bias in cottonseed yield estimation; whereas if the null hypothesis is rejected, the specific effect is a source of bias.

Finally, the third part of the analysis consisted of a simple regression analysis to determine the relation between actual and estimated seed yield. Estimated cottonseed yield was the independent variable, and actual cottonseed yield was the dependent variable. The standard error of the estimate (SE) was used to evaluate the accuracy of individual estimates of seed yield for a given L, I, and V.

Results and Discussion

The analysis was performed using a sample of 468 observations, with three main effects (L, I, and V) and three interaction effects, over nine years (1985-93). The first part of the statistical analysis was designed to test the overall bias in cottonseed yield estimation. Equation (2) was estimated obtaining k = 9.1 lb. ($p < .05$). The value of the estimated coefficient k indicates that actual cottonseed yield is greater than estimated cottonseed yield by an average 9.1 lb. per acre. This significant difference implies that there is a bias in the estimation of cottonseed yield. Yet, the overall level of the bias (9.1lb. per acre) represents a less-than-one percent error in the estimation of cottonseed yield, based on an average seed yield of 1134 lb. per acre.

Table 2 contains the ANOVA table presenting the results from fitting the reduced model for each main and interaction effect (part 2 of the analysis, equations 4 through 9). In this manner, the specific sources of bias were identified, and the following results obtained. The ANOVA indicates that L and the interaction effects are not significant sources of bias in cottonseed yield estimation. However, V and I are significant sources of bias ($p < 0.05$). Among the different varieties, Deltapine 50 was the only significant variety. This can be attributed to specific genetic characteristics of this variety or limitations of the data. In this manner, the relevance of this effect as a source of bias can be minimized. In fact, a statistical test which did not include this variety nor its interactions showed non-significant results for the variety effect.

Irrigation level was a significant source of bias ($p < .05$) in cottonseed yield estimation (Table 2). For this effect, a ±12.4 lb. per acre bias in cottonseed yield estimation with respect to the overall bias was found. This implied a 21.5 lb. per acre bias for irrigated trials, and a -3.9 lb. per acre bias for dryland trials. Further, the bias was significant for irrigated trials ($p < .01$), but not significant for dryland trials. The significance of this effect has no clear explanation. However, it was speculated that the physiology of irrigated trials (differences in plant size and maturity, etc.), and environmental conditions (interacting with the time span between sampling and stripping) might be causing the bias.

Finally, the third part of the analysis determined the relation between actual and estimated cottonseed yields, and the accuracy of individual estimates. A simple regression analysis was performed to determine the relation between both variables (Figure 1). The correlation was 0.99, indicating, on average, a close relation between both measures. The SE from the regression (78.3 lb. per acre) provided an indication of the accuracy of estimated cottonseed yield. Assuming a normal distribution, approximately 2/3's of the differences between actual and estimated seed yields would fall within one SE of the overall mean difference (i.e., A-E = 9.1 ± 78.3 lb. per acre). The high correlation of the regression and the relatively low SE (less than 7% variation on an average 1134 lb. per acre seed yield) support the fact that estimated cottonseed yield can be used as a reliable estimator of actual seed yield.

Summary and Conclusions

A test to determine the bias and accuracy in the estimation of cottonseed yield using lint yield and lint percentage of seed cotton was conducted. The data used in the analysis were obtained from Texas Agricultural Experiment Station publications (Gannaway, et al) which provided actual cottonseed yields. Three main effects (location, irrigation level, and variety) and their interactions were tested as potential sources of bias (first part of the analysis). The results showed a significant 9.1 lb. per acre overall bias in cottonseed yield estimation, which represents a less than one percent error based on mean cottonseed yield. The second part of the analysis was designed to identify the specific sources of bias. Location and interaction effects showed non-significant results. However, V and I were significant sources of bias in cottonseed yield estimation. The bias generated by V was due to one particular variety, and thus its relevance as a source of bias was minimized. The only source of significant and scientifically important bias (beyond the overall bias) was I. Irrigated trials had a significant bias (21.5 lb. per acre), while dryland trials had a non-significant bias. There is no clear explanation as to why there is a bias in irrigated trials but not in dryland trials. However, note that the irrigation level effect is a local effect to the South Plains of Texas (with two possible levels, irrigated or dryland), not common to other regions of the Cotton Belt. Finally, the third part of the analysis evaluated the relationship between actual and estimated seed yield, and the accuracy of individual estimates. The correlation between actual and estimated seed yield measures was 0.99, with a SE of 78.3 lb. per acre. It was concluded that estimated cottonseed yield (from lint yield and lint percentage of seed cotton) is a reliable estimator of actual cottonseed yield, with the caveat that individual estimates will fall within the ±78.3 lb. per acre range 2/3's of the time.

This research has implications for economic analysis of cotton. With data on seed yield, it will be possible to include the seed component in cost-benefit, risk and return analysis, and cotton variety selection. Data on seed yield by variety should be useful to the crushing industry in analysis and decision making. Undoubtedly, there are many other uses for cottonseed yield data.

However, there is need for additional research to evaluate the accuracy of the estimation procedure (used in this research) for other regions of the Cotton Belt. The results reported here are for the South Plains of Texas, and it is not known whether they can be generalized to other regions. The immediate problem with conducting this type research for other regions is the lack of actual seed data.

References

1. Kinard, D. 1993. Seed index: key to higher value cotton crop? Unpublished outline of presentation to Beltwide Cotton Improvement Conference.

2. Kohel, R. J. 1978. Breeding for cottonseed quality. Proc. Beltwide Cott. Prod. Res. Conf., 223-225.

3. Neter, J., W. Wassermam, and M. H. Kutner. 1990. Applied Linear Statistical Models. Richard D. Irwin, Inc., Boston, MA.

4. Parvin, Jr., D. W., F. T. Cooke, Jr., E. A. Stennis. 1978. The economics of cottonseed quality. Proc. Beltwide Cott. Prod. Res. Conf., 210-212.

5. Texas Agricultural Experiment Station. 1985-1994. Cotton Performance Tests in the Texas High Plains and Trans-Pecos Areas of Texas. Texas A & M University System, College Station, TX.

6. United States Department of Agriculture. Various Issues. Cotton Varieties Planted. Agricultural Marketing Service, Cotton Division, Memphis, TN.

7. Walker, M. H. 1993. Markets for cottonseed. Proc. Beltwide Cott. Prod. Res. Conf., 97.

Table 1: *Set of dummy variables for (a) location effect, (b) irrigation level effect, (c) variety effect.*

(a) *Location effect.*

L	L1	L2
Lubbock	1	0
Halfway	0	1
Lamesa	-1	-1

(b) *Irrigation level effect.*

I	I
Irrigated	1
Dryland	-1

(c) *Variety effect.*

V	V 1	V 2	V 3	V 4	V 5	V 6	V 7	V 8	V 9	V 10	V 11	V 12	V 13	V 14	V 15
PM 147	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Atlas	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
PM HS26	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
PM 145	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
BB-53	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
SM1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1517-88	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Quicke	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
HQ95	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CABCS	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
DP 50	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
DP 90	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
DP SR383	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
PM HS200	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CD3H	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1

Table 2. *ANOVA table for cottonseed yield estimation (based on equations 3-9).*

Source of variation	SS	d. f.	MS	F*	(p-value)
Irrigation-Variety Interaction	52,880.8	14	3,777.2	0.63	($p<$ 0.84)
Location-Variety Interaction	84,504.8	28	3,018.0	0.51	($p < 0.98$)
Location-Irrigation Interaction	23,166.3	2	11,583.1	1.94	($p < 0.15$)
Variety	150,388.8	14	10,742.1	1.78	($p < 0.04$)
Irrigation	45,321.8	1	45,321.8	7.59	($p < 0.01$)
Location	3,236.8	2	1,618.4	0.27	($p < 0.76$)
Error	2,424,246.5	406	5,971.1		