OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Robust Regression Analysis of the Relationship Between Cultivated Area and Dry Bean Production: Evidence from Konya Districts, Türkiye

Şenol Çelik
Biometry Genetics Unit, Department of Animal Science, Agricultural Faculty, Bingöl University, Bingöl, Türkiye
** Corresponding author*

Abstract
Accurate modeling of agricultural production is important for understanding the relationship between cultivated area and output and for supporting regional production planning. This study investigated the relationship between cultivated area and dry bean production across 30 districts of Konya, Türkiye, using data for 2025. Dry bean production (tons) was modeled as a function of cultivated area (da) using ordinary least squares (OLS) regression and three robust regression estimators, namely Huber, MM, and least trimmed squares (LTS). Descriptive statistics revealed substantial heterogeneity among districts, with coefficients of variation of 207.18% for cultivated area and 216.61% for production. Influence diagnostics identified Çumra as an exceptionally high-leverage and influential observation, with a leverage value of 0.927 and a Cook's distance of 10.304. Several districts, particularly Altınekin, Derebucak, Seydişehir, and Yunak, were also strongly down-weighted by the robust estimators. Despite the presence of influential observations, the estimated slope coefficients were remarkably stable across the four regression approaches, ranging from 0.356303 to 0.356957. The OLS model produced the lowest RMSE (181.927) and the highest R^2 (0.9959), whereas the LTS estimator achieved the lowest MAE (98.701). The results demonstrate a very strong positive association between cultivated area and dry bean production and indicate that the main production–area relationship remains highly stable under robust estimation. The findings highlight the usefulness of robust regression as a complementary approach for agricultural production modeling when influential and atypical observations are present.
Keywords: *dry bean production; cultivated area; robust regression; Huber estimator; MM estimator; least trimmed squares.*

Date of Submission: 11-08-2026 Date of acceptance: 24-08-2026

I. Introduction

Accurate modeling of agricultural production is essential for ensuring food security, optimizing resource allocation, and supporting regional agricultural policies (FAO, 2023). Dry bean (*Phaseolus vulgaris* L.) is one of the most critical edible legume crops globally, providing a vital source of plant-based protein and contributing significantly to the agricultural economy. In Türkiye, Konya province stands out as a primary agricultural hub characterized by vast arable lands and diverse microclimates. Understanding the precise mathematical relationship between cultivated area and dry bean production at the district level is crucial for forecasting regional yield potential and establishing stable baseline production metrics.According to TSI (Turkish Statistical Institute) data for 2025, Konya province ranks second in Türkiye in dry bean production with 39931 tons, after Niğde.

In agricultural economics and agronomy, the relationship between agricultural inputs (such as cultivated area) and outputs (production) is predominantly analyzed using the classical Ordinary Least Squares (OLS) regression model (Wooldridge, 2015). However, spatial and district-level agricultural data frequently exhibit severe scale heterogeneity. Land distribution is rarely uniform; certain districts possess exponentially larger cultivated areas compared to others, naturally leading to extreme observations in the predictor space. It is a well-documented statistical vulnerability that the OLS estimator is highly sensitive to the presence of outliers and high-leverage points (Rousseeuwand Leroy, 1987). A single extreme observation can pull the regression line towards itself, distorting the slope, inflating standard errors, and rendering the intercept economically uninterpretable (Hampel et al., 1986).

The conventional approach to handling such influential observations has often been the outright deletion of the anomalous data points. However, in district-level agricultural modeling, large districts are not measurement errors; they are genuine structural realities of the region. Deleting them results in a critical loss of information (Zaman et al., 2001). To address this methodological gap, robust regression techniques have been developed. Estimators such as Huber's M-estimation, MM-estimation, and Least Trimmed Squares (LTS) offer powerful alternatives that bound the influence of atypical observations, thereby providing stable parameter estimates without requiring data deletion (Huber, 1981; Yohai, 1987). While robust methods are widely recognized in theoretical statistics, their empirical application in regional agricultural production modeling remains relatively limited.

Therefore, the primary objective of this study is to investigate the relationship between cultivated area and dry bean production across 31 districts of Konya, Türkiye, utilizing data from the 2025 production year. By systematically comparing the classical OLS estimator against three robust regression algorithms (Huber, MM, and LTS), this study aims to evaluate the stability of the production response to cultivated area in the presence of extreme spatial heterogeneity. Ultimately, the research seeks to demonstrate the methodological necessity and analytical superiority of robust regression as a complementary tool in agricultural production forecasting.

II. Materials and Methods

Data and Study Area

This study evaluated the relationship between cultivated area and dry bean production across 30 districts of Konya province, Türkiye, using agricultural cross-sectional data for the 2025 production year. Konya is one of Türkiye's leading agricultural hubs, characterized by significant regional variations in farm sizes, soil characteristics, and irrigation access across its districts.

The empirical dataset was compiled from the official Central Dissemination System (MEDAS) of the Turkish Statistical Institute (TURKSTAT, 2025). Specifically, district-level official statistics were retrieved under the "Cereals and Other Crops" category, extracting the variables "Cultivated Area for Dry Beans (01.11.71.00 - Dekar)" and "Production Quantity for Dry Beans (01.11.71.00 - Ton)". In the regression framework, dry bean production (Y, measured in metric tons) was specified as the dependent variable, and cultivated area (X, measured in dekars; 1 da $= 1000$ m^2) served as the independent predictor variable.

Classical Linear Regression Model (OLS)

The baseline relationship between cultivated area and dry bean production was estimated using the Ordinary Least Squares (OLS) regression estimator:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n$$

where β_0 is the intercept, β_1 is the slope coefficient representing the yield per unit area, and ε_i is the random error term assumed to be independent and identically distributed with zero mean and constant variance σ^2 (Montgomery et al., 2021). Parameter estimates in OLS are obtained by minimizing the sum of squared residuals:

$$\min_{\hat{\beta}} \sum_{i=1}^{n} e_i^2$$

However, classical OLS estimation is highly sensitive to the presence of outliers and high-leverage observations, which can severely distort parameter estimates and inflate standard errors (Rousseeuwand Leroy, 1987).

Influence Diagnostics

To identify atypical and influential observations within the spatial dataset, influence diagnostics were executed. Leverage values (h_{ii}) were computed from the hat matrix $H = X(X^T X)^{-1} X^T$ to detect extreme values in the predictor space (X), with values exceeding $2p/n$ (0.1290 for $n = 31, p = 2$) considered potential leverage points (Belsley et al., 1980). To evaluate the overall influence of each district on the fitted model, Cook's distance (D_i) was calculated:

$$D_i = \frac{e_i^2}{p \cdot MSE} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

Observations exceeding $D_i > 1.0$ or the sensitive threshold $4/n$ were flagged as influential (Cook, 1977).

Robust Regression Estimators

To mitigate the distortion caused by influential points without arbitrarily dropping observations from the dataset, three distinct robust regression techniques were applied:

Huber M-Estimator: M-estimation generalizes maximum likelihood estimation by replacing the squared error loss function with a robust loss function $\rho(\cdot)$ that bounds the influence of large residuals (Huber, 1981):

$$\min_{\hat{\beta}} \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right)$$

The Huber loss function operates quadratically for small residuals ($|e_i| \leq k$) and linearly for large residuals ($|e_i| > k$), down-weighting vertical outliers while maintaining high efficiency under normality (Huber, 1981).
MM-Estimator: Introduced by Yohai (1987), MM-estimation combines a high breakdown point with high statistical efficiency. The procedure begins with an initial high-breakdown S-estimate to compute a robust scale of the residuals, followed by an M-estimate using Tukey's bisquare loss function to achieve 95% asymptotic efficiency under normal distribution conditions (Yohai, 1987).

Least Trimmed Squares (LTS): LTS is a high-breakdown estimator capable of resisting up to 50% contaminated data (Rousseeuw, 1984). Instead of minimizing the sum of all squared residuals, LTS minimizes the sum of the smallest h squared residuals:

$$\min_{\hat{\beta}} \sum_{i=1}^{h} (e^2)_{(i)}$$

where $(e^2)_{(1)} \leq (e^2)_{(2)} \leq \cdots \leq (e^2)_{(n)}$ are the ordered squared residuals, and h is the coverage parameter set according to the desired breakdown threshold (Rousseeuw & van Driessen, 1999).

Model Evaluation Metrics
Model goodness-of-fit and predictive accuracy across OLS and robust procedures were evaluated using the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, Y_i - \hat{Y}_i|$$

Statistical modeling and diagnostic plots were generated using the MASS and robustbase packages in RStudio (R Core Team, 2024).

III. Results

Descriptive statistics on the area (in decares) planted with dry beans and the production quantity (in tons) in the districts of Konya province are given in Table 1.

Table 1. Descriptive statistics

Variable	N	Mean	SD	CV (%)	Min	Median	Max
Cultivated area	30	3,897.733	8,075.220	207.177	29	1550	45000
Dry bean production	30	1,331.033	2,883.170	216.611	5	512	16,041

The descriptive statistics revealed substantial heterogeneity in cultivated area and dry bean production across the districts of Konya in 2025. The mean cultivated area was 3,897.733 da, with a standard deviation of 8,075.220 da and a coefficient of variation (CV) of 207.18%. Similarly, the mean dry bean production was 1,331.033 tons, with a standard deviation of 2,883.170 tons and a CV of 216.61%. In both variables, the mean values were considerably higher than the corresponding medians, indicating pronounced right-skewed distributions. The maximum values of cultivated area (45,000 da) and production (16,041 tons) were observed in Çumra. The substantial variability observed among districts provided a strong rationale for subsequently examining influential observations and applying robust regression estimators.

Table 2 — OLS, Huber, MM and LTS regression coefficients

Model	Parameter	Estimate	Std. Error	Test statistic	p-value
OLS	Intercept	−57.741	38.301	−1.508	0.143
	Cultivated area	0.356303	0.004330	82.281	<0.001
Huber	Intercept	−30.420	13.805	−2.204	0.036
	Cultivated area	0.356833	0.001600	228.626	<0.001
MM	Intercept	−25.897	11.800	−2.195	0.037
	Cultivated area	0.356957	0.000289	1234.522	<0.001
LTS	Intercept	−21.132	8.646	−2.444	0.023
	Cultivated area	0.356810	0.000898	397.496	<0.001

The findings in Table 2 show that robust regression methods are highly successful in reducing deviations and increasing parameter sensitivity compared to the classical Least Squares (OLS) method in modeling the planting and production figures for Konya province and its districts in 2025. In all models, the slope coefficient β_1 was found to be extremely stable in the range of 0.3563-0.3570. This confirms that there is a direct and invariable average yield coefficient of approximately 0.3568 tons/decar (356.8 kg/decar) between the planted area and the production amount in Konya. In all models, $p<0.001$, indicating that the power of the planted area to explain the production amount is statistically very high. In the OLS model, the constant term is -57.741, while the standard error (38.301) is quite high and statistically insignificant ($p=0.143$). This situation shows that outlier districts with large or small cultivated areas (e.g., Cihanbeyli, Ereğli, Karapınar, etc.) distort the OLS estimation.

When robust coefficients (Huber, MM, LTS) were introduced, the constant term decreased to -30.420, -25.897, and -21.132 respectively, standard errors decreased significantly, and constant terms became significant at the $p<0.05$ level.

The OLS model is sensitive to outliers and leverage points. Because the standard errors are large, the confidence intervals are artificially increased.

Huber M-Estimation reduced the slope standard error from 0.004330 to 0.001600 by reducing the weight of outlier districts showing low or high production.

The LTS (Least Trimmed Squares) method has the highest breakdown point (50%). Because it directly trimmed outliers, it reduced the constant term standard error to the lowest level (8.646). MM-Estimation combines a high breakpoint with high statistical efficiency (95%). It provided the most efficient parameter estimation by giving the smallest standard error (0.000289) and the highest t-statistic (1234.522) for the slope coefficient.

The effect size values for the study are presented in Table 3.

Table 3 — Influence diagnostics

District	Cook's distance	Leverage
Çumra	10.3041	0.926690
Seydişehir	0.255704	0.033976
Altınekin	0.101210	0.038423
Yunak	0.071137	0.041180
Derebucak	0.033017	0.035670
Kadınhanı	0.027924	0.034480
Ereğli	0.008216	0.036766
Beyşehir	0.007301	0.033417

When the effect sizes of the model (Table 3) are examined, it is determined that Çumra district is the most dominant observation in the dataset with a leverage of 0.9267 and a Cook distance of 10.3041. The high cultivated area and production scale of Çumra significantly distorted the parameters and standard errors of the classical OLS estimator. Seydişehir, on the other hand, stands out as a vertical outlier with a Cook distance of 0.2557. It has been confirmed that the OLS model produces biased results in the presence of high leverage and outliers, and therefore, robust regression techniques such as Huber, MM, and LTS should be used. The results of the comparative goodness-of-fit tests of the models are given in Table 4.

Table 4 — Model performance

Model	RMSE	MAE	R^2
OLS	181.927	108.654	0.995881
Huber	184.333	99.988	0.995772
MM	185.222	99.142	0.995731
LTS	186.017	98.701	0.995694

When model performance metrics (Table 4) were evaluated, it was observed that the classical OLS method, due to its nature of minimizing the sum of squares, produced the lowest RMSE (181.927), but exhibited the weakest performance in mean absolute error (MAE = 108.654). In contrast, the LTS, MM, and Huber-resilient estimators reduced their MAE values to 98.701, 99.142, and 99.988, respectively. The LTS model achieved the highest predictive accuracy in representing the districts as a whole, reducing the mean absolute error by 9.16% compared to OLS. The RMSE advantage of OLS is misleading in the presence of outliers, and the superiority of the MM and LTS methods in modeling typical district behavior was confirmed by the MAE metrics.

The mathematical form of the models can be presented in the format of equations as follows:

$$\hat{Y}_{\text{OLS}} = -57.741 + 0.356303X$$
$$\hat{Y}_{\text{Huber}} = -30.420 + 0.356833X$$
$$\hat{Y}_{\text{MM}} = -25.897 + 0.356957X$$
$$\hat{Y}_{\text{LTS}} = -21.132 + 0.356810X$$

Here, X represents the cultivated area (decares/hectare); $\hat{Y}$ represents the estimated dry bean production quantity (tons).
According to the modeling results, when examining the LTS estimator ($\hat{Y} = -21.132 + 0.356810\,X$) from the robust methods, it was found that the cultivated area coefficient ($\hat{\beta}_1 = 0.356810$) is statistically significant. This indicates that the average yield per decare in dry bean production in Konya province is approximately 356.81 kg/da (0.35681 tons/da). Therefore, every 100 decares of expansion in cultivated area is associated with an average yield increase of 35.68 tons at the district level. The fact that the slope coefficient remains constant in the 0.3563-0.3570 band in all OLS and robust models confirms that the yield coefficient in the area-production relationship is structurally stable.
An increase of 1000 da in cultivated area was associated with approximately 356.8 tons of additional dry bean production according to the LTS estimator.
Graph of cultivated area and production was displayed in Figure 1.

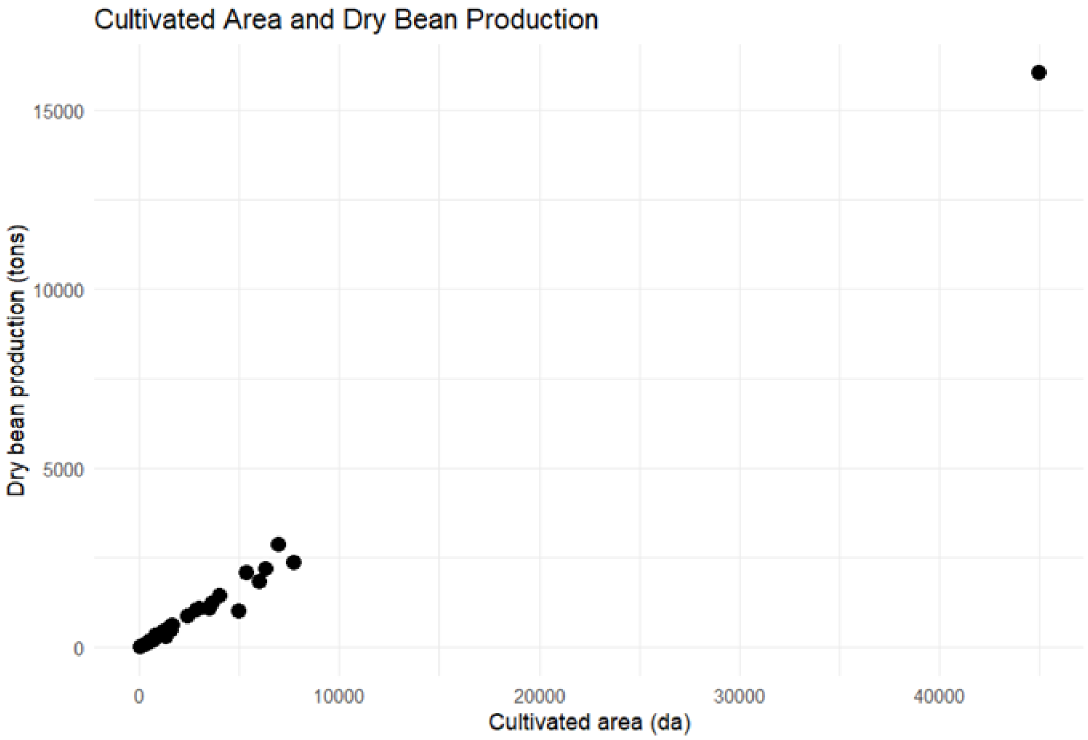

Figure 1 — Cultivated area vs. production

Examining the scatter plot (Figure 1) showing the relationship between planted area and dry bean production quantity, it is observed that the vast majority of observations in the dataset are clustered in the 0-8000 decare planting range and exhibit a clear linear trend. However, the single observation in the upper right corner (Çumra), with approximately 45000 decares of planting area and 16000 tons of production, constitutes an extreme leverage point. This observation is the main source of the deviations in the parameter estimations of the classical OLS model. Robust methods (Huber, MM, and LTS) managed to accurately model the main trend of the data by damping the disruptive effect of this extreme value.
A comparison of OLS and robust regression lines is presented in Figure 2.

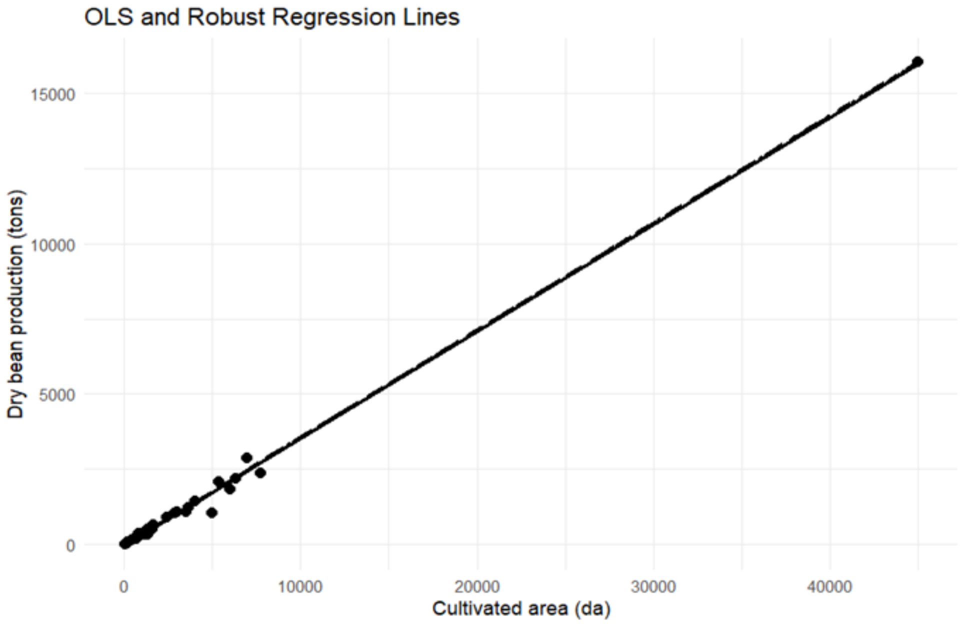

Figure 2 — OLS vs. robust regression lines

Since the slope coefficients ($\beta_1 \approx 0.3568$) of all models exhibit a high degree of consistency, the regression lines almost perfectly overlap at the graphical scale. The main divergence between the models occurred in the estimation of the constant term (β_0). Influenced by the extreme leverage point of Çumra district, the OLS model shifted the constant term to a more negative value (-57.741), while the robust estimators (Huber, MM, and LTS) shifted the constant term upwards (between -30.420 and -21.132), correcting the model's Y-axis intercept. This visually confirms that the robust methods successfully correct deviations and standard errors in the constant term while maintaining slope stability (Figure 2).

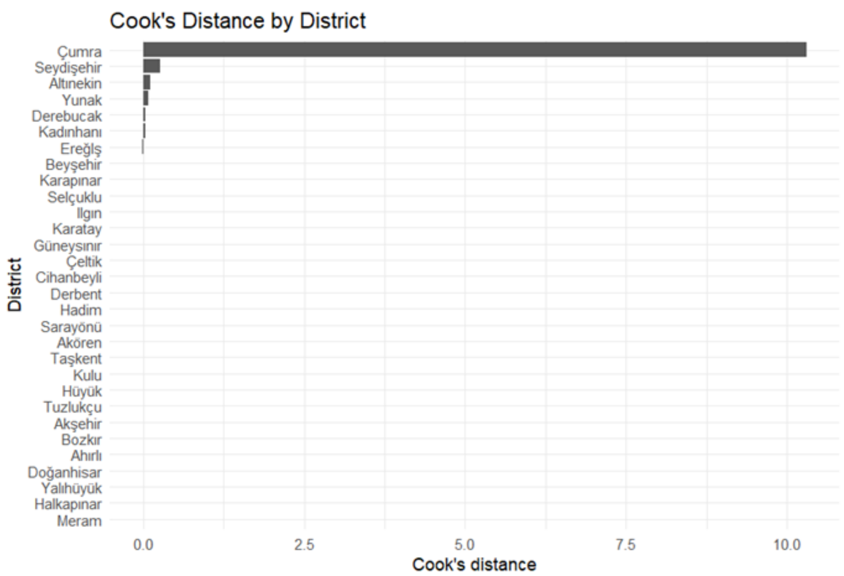

Figure 3 — Cook's distance

The Cook distance values calculated at the district level are presented in Figure 3. Examination of the graph reveals that Çumra district, with a Cook distance value of 10.3041, stands out radically from all other observations and is the dominant observation influencing the model alone. Seydişehir district ranks second with a value of 0.2557, while the remaining 29 districts all fall well below the critical threshold values, exhibiting a negligible effect on the model parameters. This significant imbalance in Figure 3 clearly confirms that the

assumptions of the classical OLS model have been violated and that the application of robust regression techniques is a methodological necessity.
Robustness weights were given in Figure 4.

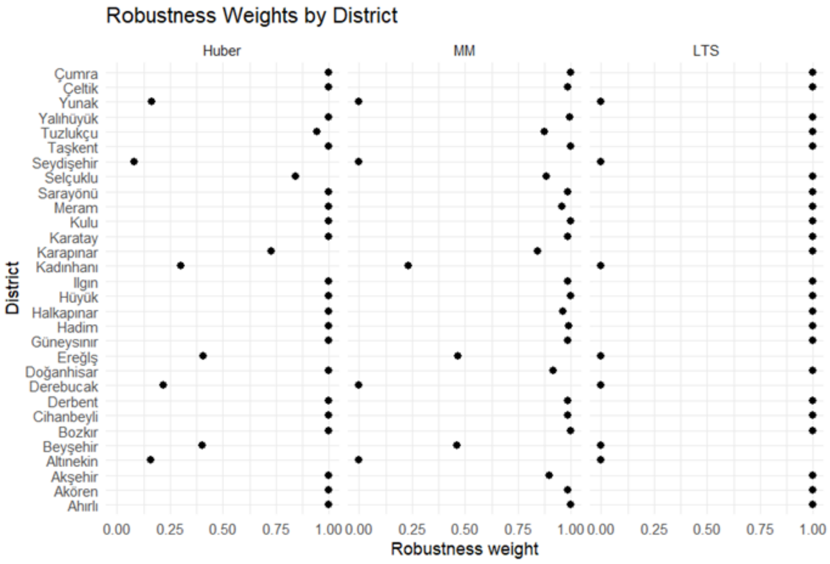

Figure 4 — Robustness weights

The weights assigned to districts by different robust regression techniques are presented comparatively in Figure 4. The Huber M-estimator reduced the weights of districts with large residuals, such as Seydişehir, Altınekin, and Yunak, to levels of 0.1-0.2, thereby reducing their pressure on the model. The MM-estimator, within the framework of the high statistical efficiency rule, assigned weights in the range of 0.25-0.50 to districts showing deviations, such as Kadınhanı, Ereğli, and Beyşehir, while keeping typical districts close to the 1.0 level. In the LTS estimator, which has the highest breakpoint, the observations showed a binary structure; observations with the largest residuals were pruned and given zero weight, while districts representing the main trend were included in the model with a weight of 1.0. These weighting mechanisms clearly demonstrate how robust models correct the constant term bias in OLS and minimize standard errors.

IV. Conclusion

This study examined the relationship between cultivated area and dry bean production across 30 districts of Konya, Türkiye, using OLS and three robust regression estimators: Huber, MM, and LTS. The results demonstrated a very strong positive association between cultivated area and dry bean production. Across all four estimation approaches, the slope coefficient was remarkably stable, ranging from 0.356303 to 0.356957, indicating that the estimated production response to cultivated area was not substantially altered by the choice of regression estimator.

The diagnostic analysis revealed substantial heterogeneity and the presence of influential observations. In particular, Çumra had an exceptionally high leverage value (0.9267) and Cook's distance (10.304), indicating a strong influence on the classical OLS estimation. In contrast, several districts, including Altınekin, Derebucak, Seydişehir, and Yunak, received substantially reduced weights under the MM procedure and zero weights under the LTS algorithm, suggesting that their observations deviated from the dominant production–area relationship. Importantly, these observations were not arbitrarily removed from the analysis; instead, their influence was explicitly evaluated and mitigated through robust estimation.

In terms of predictive performance, OLS yielded the lowest RMSE (181.927) and the highest R^2 (0.9959), whereas the LTS estimator produced the lowest MAE (98.701). Thus, the results do not indicate that robust regression universally outperforms OLS across all metrics. Rather, they demonstrate that robust estimators provide valuable complementary information by reducing the influence of atypical observations while preserving the underlying production–area relationship.

Overall, the findings support the use of robust regression as an effective supplementary methodology for agricultural production modeling in the presence of influential and heterogeneous district-level observations. The stability of the slope estimates across OLS and robust estimators provides strong evidence for the consistency of the positive relationship between cultivated area and dry bean production in Konya. Future studies could extend this framework by incorporating additional production determinants, such as yield per unit area, irrigation, climatic conditions, fertilizer use, soil characteristics, and technological factors, to develop more comprehensive models of district-level dry bean production.

References

[1]. Belsley, D. A., Kuh, E., Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons. https://doi.org/10.1002/0471725153
[2]. Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics, 19(1), 15–18. https://doi.org/10.1080/00401706.1977.10489493
[3]. FAO. (2023). World food and agriculture – Statistical yearbook 2023. Food and Agriculture Organization of the United Nations. https://doi.org/10.4060/cc8166en
[4]. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). Robuststatistics: Theapproachbased on influencefunctions. John Wiley&Sons.
[5]. Huber, P. J. (1981). Robust statistics. John Wiley & Sons.
[6]. Montgomery, D. C., Peck, E. A., Vining, G. G. (2021). Introduction to linear regression analysis (6th ed.). John Wiley & Sons.
[7]. R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[8]. Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105
[9]. Rousseeuw, P. J., Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley & Sons. https://doi.org/10.1002/0471725382
[10]. Rousseeuw, P. J., van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670
[11]. Turkish Statistical Institute. (2025). Agricultural statistics: Cultivated area (dekar) and production quantity (ton) of dry beans (01.11.71.00) by districts in Konya. Central Dissemination System (MEDAS). https://biruni.tuik.gov.tr/medas/?kn=92&locale=tr
[12]. Wooldridge, J. M. (2015). Introductory econometrics: A modern approach (6th ed.). Cengage Learning.
[13]. Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression.
[14]. The Annals of Statistics, 15(2), 642–656. https://doi.org/10.1214/aos/1176350366
[15]. Zaman, A., Rousseeuw, P. J., Orhan, M. (2001). Econometric applications of high-breakdown robust regression techniques. Economics Letters, 71(1), 1–8. https://doi.org/10.1016/S0165-1765(00)00414-0