Tuijin Jishu /Journal of Propulsion Technology
ISSN:1001-4055
Vol. 44 No.04 (2023)

Statistical Modelling of the Agriculture Variables using Various Regression Techniques

[1] Dr. SanjithBharatharajan Nair, [2] Prof. ZuhairAbdulamir Al-Hemyari,

[3] Dr. Naresh Kumar

[1] [2] [3] Department of Mathematical and Physical Sciences,

College of Arts and Sciences, University of Nizwa, Nizwa, Oman

Abstract

A response variable's association with one or more explanatory variables examined and modelled using the strong statistical technique, regression analysis. Regression models used in a variety of ways to estimate agricultural production and area, to assist farmers, academics, and policymakers in making well-informed choices. By revealing information about crop productivity, resource optimization, and risk management, regression models in agriculture help decision-makers. These models support sustainable and effective agricultural operations by exploring connections between different elements and results. The analysis conducted to assess the growth pattern in terms of total cultivated area, production, temperature and humidity. In this study, numerous regression procedures were thoroughly analysed, and the goodness of fit examined. Simple and multiple regression models designed for the study purpose. All simple regression models in this investigation found to match the data satisfactorily, likewise all multiple regression models were in line with the data well. Bootstrap technique also performed to build confidence intervals for regression models' estimations in order to check the validity of the estimates. All simple and multiple regression estimates found to be valid. All simple and multiple regression models' estimates found bounded within the 95% bootstrapping confidence limits.

Keywords: *Ordinary least squares regression, Weighted least squares regression, Mean absolute error,Watanabe-Akaike Information Criterion,Breusch-Pagan test.*

1. Introduction

A fundamental statistical method used to model relationships between variables and generate projections is regression analysis. These kinds of models frequently used in many different domains, such as machine learning, social sciences, finance, and economics. Despite the fact that Ordinary Least Squares regression(see Ii et al., 2019) is frequently used, there are a number of advanced regression approaches that provide greater flexibility when modelling complex relationships and managing data that has specific properties.Understanding these methods will give academics and professionals a wider toolbox to address practical issues and improve the precision and dependability of their regression studies.

Three regression techniques applied for the study purpose. They were the Ordinary Least Squares regression, Weighted Least Squaresregression(see Koenker et al., 1993a) and Generalized Least Squares regression(see Morgenthaler, 1992). Data between year 1999 and 2020 for the study acquired from "The National Centre for Statistics and Information Oman (NCSI)".

The study was on the agriculture production of the Sultanate of Oman. In this country, we have not seen many statistical studies related to this subject. Being the foundation of an economy and having an impact on many facets of society, agriculture production is crucial to the growth and prosperity of a nation(see Sulewski et al.,

2017). There are several important factors like food security, economic expansion, foreign exchange earnings, poverty reduction, development of rural areas, supply chain integration, conservation of the environment, social stability, and innovation in technology that highlight how important agriculture output is for any country(Njegomir et al., 2017).

To maintain a successful and sustainable country, the Oman Government developed several policies and investments that support the expansion and sustainability of the agricultural industry.

Exploring the connection between yields from agriculture and the associated land area involves regression analysis in a big way. Regression approaches (see Dielman et al., 2005)can be used to shed light on the variables affecting agricultural productivity and help governments and populations to allocate resources and manage land more effectively.

In this paper, we go through how regression analysis used in relation to agricultural productivity and area.The value of such modelling efforts is that, because different agricultural mechanisms follow different patterns, an overview of the model selection criteria will reveal the underlying mechanism. Furthermore, knowing predicted future production would aid agriculture researchers in refocusing their study on the goal of long-term development. The regression models chosen to assess the trend of harvesting area, and production.

Thus the justification and the needs of the paper is coming from the above reason and also to fill the gap of lack of research into the country's current agricultural output trends has hampered effective planning and program development in this area(see Sumberg, 2022). It is crucial to keep a close eye on the agricultural sector's growth pattern over time to see if it is making the best use of its resources to attain maximum/optimum output. Because the sector is so important to the country's economy, it is critical to pinpoint the periods of insufficient growth and do deeper research into the relevant government policies and initiatives. This study's objective was to look at the trends in growthof agricultural production and cultivating area and forecasting the food security.

1.1. Our Contribution

1) In this study, the measures of the goodness of fit are examined to efficiently model agricultural statistics of land area, temperature and humidity with various class of regression techniques.

2) It was observed that all the regression models fit well for the data. All simple models showed a very high significance (p-value < 0.001) connection between production and area with very high R-squared values which concluded that the association between the variables was significantly high.

3) All multiple models fit the data well but humidity was not appeared significantly for all the models. In multiple models, the WLS model had the largest residual standard error. The adjusted R-squared value was significant and very similar in all multiple models. On computing OLS and WLS have the same AIC (-37.426) and BIC (-32.204), which are lower than the GLS AIC and BIC values. OLS and WLS have the highest log-likelihood (23.713), indicating a better fit to the data compared to GLS.

4) Within the 95% bootstrapping confidence intervals, observed regression model estimates were found except for GLS model, which satisfied the validity of the regression estimates. A 95% prediction interval of production was observed valid for the models.

There are six sections in this paper. Section 2 included the literature review, which comprises a thorough investigation and analysis of current scholarly sources and academic literature that are pertinent to the research issue. In Section 3, the methodologyis discussed methodology, which described the methodical steps followed to answer the research questions and accomplish the study's goals. This section provided a clear and thorough explanation of the procedures utilized which were able to evaluate the validity and dependability of the findings. Sections 4 and 5 included the results and conclusion. Limitations of the study were included in Section 6.

2. Literature review

When input variable predicts a continuous variable, regression analysis is the method of choice. With an error estimate provided by an optimization method, it frequently gives explicit estimates of measure for the cause-effect relationship between the various inputs and the outcome.Ramsey(1969)looked at the consequences of various model mis-specifications on the distribution of least-squares residuals. In order to assess the linearity of calibration curves when using the ordinary least squares method (OLSM), Souza and Junqueira(2005)established a rigorous technique that included experimental design, parameter estimates, handling outliers, and assumption evaluation.Huang (2018) compared accuracy of estimates in Multilevel Modelling and Ordinary Least Squares Regression.Lechene et al.,(2022) investigated the intrahousehold distribution of expenditure using ordinary least squares estimation.

Kim and White (2001) compared weighted least squares regression analysis to polynomial regression for confirming the analytical measurement range. Funes et al., (2019) used weighted least squares regression to define carbon sequestration techniques at the regional level in relation to climate change impacts on agricultural land usage and agricultural management techniques. Zhang and Mei (2011)discovered proof from a spatially weighted regression model on the variables influencing CO2 emissions in the agricultural sector.Romano and Wolf (2017)showed how asymptotically valid inference in regression models based on the weighted least squares estimator may still be obtained even when the model for reweighting the data is incorrect.Schermelleh-Engel et al. (2003)used weighted least squares to evaluate the fit of structural equation models, including tests for significance and descriptive goodness-of-fit indices.

Xu et al., (2021)used extended regression neural network and vis-nir spectroscopy with fractional-order derivative to estimate the presence of heavy metals in agricultural soils. Linnet (1993)evaluated regression procedures for methods comparison studies. Kim et al. (2022)looked at the interactions between environmental, social, and governance (ESG) initiatives and the results of multinational companies' (MNCs') subsidiaries. They looked into how market-oriented organizational culture affected the link between environmental factors and performance using generalized least squares regression.

In order to compare the effectiveness of various estimators used in structural equation modelling, Zulkifli et al. (2022)worked on the following studies: maximum likelihood, generalized least square, scale-free least square, partial least square, and consistent partial least square.The finest random forest and M-Hampel models, they advised, are useful for demonstrating the fewest problems and effective validation for studying and contrasting vast data. Majewski et al., (2022)established a negative correlation between the production of renewable energy, the value contributed to agriculture, and per capita CO2 emissions using two-step generalized method of moments (GMM) regression.

3. The Methodology

3.1 Data of the study

Agriculture output is essential to a nation's success and prosperity. It meets the dietary demands of the populace, stimulates economic expansion, and supports social, environmental, and political stability. Data for the study gathered from "The National Centre for Statistics and Information Oman (NCSI)". Agriculture production has given much importance in the Sultanate of Oman to compensate the domestic needs.To satisfy the rising food demand of the country's population, it is vital to boost agriculture production growth through boosting land productivity. The data variables were agriculture production, cultivated area, temperature and humidity. Data analysesconducted to study the two models; the simple model and the multiple model. Production and area chosen as the response and predictor variables, respectively, for the simple model. Production chosen as the response variable for the multiple model, whereas area, temperature, and humidity chosen as the predictor variables. The unit of production was in tonnes, area was in feddan, temperature was in degree celsius and humidity was in terms of percentage. For reducing the effect of skewness and stabilizing variance of the agriculture data; natural log of agricultural production and cultivated area were taken for the analyses.

3.2 Proposed Testing Methods

Based on the literature review it observed that regression techniques are most suited to the data for analysing the progress on agriculture. For that, different regression models used for modelling and estimating the relationship between cultivated area and agriculture production. All analyses conducted using R software. Assumption testing conducted for regression models prior to the regression analysis. The study analysis included the assumption testing of regression models, model estimation, goodness of fit, bootstrapping approach of validity and forecasting outputs.

3.3 Assumption testing

Assumption testing is crucial in regression analysis, since it serves to ensure the validity and dependability of the findings derived from the regression models(see Flatt & Jacobs, 2019). It provided with the proper interpretation of the data and allowed us to verify if these assumptions are true or not. Regression analysis made use of a number of key assumptions(see Nau, 2015).

3.3.1 Normality of residuals

The first assumption testing was for normality of residuals. It was presumable that the residuals will follow a normal distribution(see Draper, N. R., & Smith, 1998). The hypothesis related to this assumption is;

Null Hypothesis (H_0): The regression model's residuals (or errors) are normally distributed,and

Alternative Hypothesis (Ha): The regression model's residuals (or errors) are not normally distributed,

to be tested using the Shapiro-Wilk test. We fail to reject the null hypothesis and conclude that the data are normally distributed if the p-value obtained from the Shapiro-Wilk test is higher than the selected significance threshold (often set at 0.05).

Also,Quantile-Quantile(Q-Q) plot tool used to determine whether dataset follows a normal distribution or not. The quantiles of the observed data are contrasted with the quantiles of a hypothetical normal distribution in the Q-Q graphic. If the data points roughly follow a straight line, the data are probably normally distributed. Deviations from normality are indicated by deviations from the straight line.

3.3.2 Linearity of residuals

The second test was for the linearity of residuals.It assumed that each independent variable and the dependent variable have a linear relationship. This means that the change in the dependent variable is proportional to the change in the independent variable(s). If the relationship is not linear, it may be necessary to transform the variables or consider alternative modelling techniques.Residual plot against fitted values created for assessing the linearity. The hypothesis supporting this assumption is the following:

Null Hypothesis (H_0): The residuals show a random distribution around zero with no discernible pattern, and the connection between the predictor variables and the response variable is linear,and

Alternative Hypothesis (Ha): The residuals display a non-random pattern around zero, indicating a lack of linearity, and the connection between the predictor variables and the response variable is not strictly linear.The Rainbow test used to test this assumption.

3.3.3 Homoscedasticity of residuals

The residuals' homoscedasticity served as the foundation for the third assumption. The concept of homoscedasticity is the idea that the variance of the residuals is constant at all levels of the independent variables. The distribution of the residuals, to put it another way, should be nearly the same over the range of the independent variable(s). This assumption's supporting hypothesis is the following:

Null Hypothesis (H_0): Homoscedasticity showed by the residuals of the regression model, which show constant variance at all levels of the predictor variables,and

Alternative Hypothesis (Ha): Heteroscedasticity is demonstrated by the fact that the regression model's residuals do not have constant variance at all levels of the predictor variables.

The Breusch-Pagan testor the Breusch-Pagan-Godfrey test(see Herwartz, 2006) is used to test this assumption(see Halunga et al., 2017).

3.3.4 Independence of observations

The fourth test was independence of observations, where it isassumed that each observation in the dataset is unrelated to the others. This assumption's supporting the null and alternative hypotheses are as follows:

Null Hypothesis (H_0): There is no autocorrelation or serial correlation, as shown by the independence of the regression model's residuals,and

Alternative Hypothesis (Ha): The residuals of the regression model are not independent, indicating the presence of autocorrelation or serial correlation.

The Durbin-Watson test performed to testthisassumption.

3.3.5Multicollinearity of observations

The fifth test was for multicollinearity, which occurred when two or more independent variables have a high degree of correlation. Multicollinearity emerged, making it challenging to separate out each independent variable's impact on the dependent variable(see Shrestha, 2020). The computed regression coefficients may become unstable as a result, making them challenging to understand. Multicollinearity recognized for getting reliable findings. Tolerance, conditional number, variance inflation factor, and correlation matrix employed in this process.

3.4 Model Estimation

3.4.1 Ordinary Least Squares regression(OLS) approach

OLS regression is a flexible and extensively applicable technique that is utilized in many disciplines, including data science, engineering, social sciences, and economics. In order to interpret the connections between the variables, it offers estimates of the regression coefficients, significance tests for the coefficients and measures of goodness-of-fit, allowing for statistical inference about the model. The OLS method, which minimizes error to estimate parameters, calculates the regression coefficient β. The parameters of the model estimated as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad (1)$$

where X is the matrix of explanatory variables of size $n \times (k + 1)$, $\hat{\beta}$ is the vector of estimated regression coefficients of size $(k + 1) \times 1$ and y is the $n \times 1$ vector of response variables(see D. C. Montgomery, P. E. A., 2021).

3.4.2 Weighted Least Squares regression (WLS)approach

The extension of the ordinary least squares (OLS) regression model that gives each data point a weight based on its relative significance or precision is called the WLS regression model. OLS may generate inaccurate and biased estimates in these circumstances. In WLS regression, a weight $w_i$is given to each observation to reflect its relative significance in the regression analysis. Because the weights are inversely correlated with the error term's variance, observations with smaller variances are given more weights, indicating they have a bigger impact on estimate. The weights are typically positive numbers. The WLS model is illustrative as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (2)$$

, where Y is the dependent variable (response variable), X_1, X_2, ..., X_i are the independent variables (predictors), β_0 is the intercept term, β_1, β_2, ..., β_i are the regression coefficients for each independent variable X_1, X_2, ..., X_i, ε_i represents the error term, σ^2 is the variance of the error term($\varepsilon_i \sim N(0, \sigma^{2*}W)$) and, W is a diagonal weight matrix with the inverse of the estimated variances of the residuals as its diagonal elements.

The most typical method for obtaining the WLS regression's best-fitting parameters is to minimize the weighted sum of squared residuals.

3.4.3 Generalized Least Squares(GLS) regression approach

The GLS regression model is an extension of the OLS regression model and the WLS regression model. When the OLS presumptions are not satisfied, GLS regression has a number of benefits over OLS regression. GLS can produce more accurate and efficient estimates of the regression coefficients by taking into account the covariance structure of the error term. It enables flexible modelling of serial correlation and heteroscedasticity, which improves inference and hypothesis testing. To verify the validity of the estimated model, thorough diagnostic checks carried out when using estimation methods that rely on assumptions about the covariance structure of the error term. The GLS model modelled mathematically as:

$$y = X\beta + \varepsilon \quad (3)$$

, where y is the n × 1 vector of observed values of the response variable, X is the matrix of explanatory variables of size n × (k + 1), β is the vector of estimated parameters,ε is the n × 1 vector of error terms.By reducing the generalized sum of squared residuals, the GLS technique calculates the regression coefficients$\hat{\beta}$:

$$\text{minimize}\, \varepsilon^{T}\Omega^{-1}\varepsilon\,, \quad (4)$$

where Ω^{-1} is the inverse of covariance matrix Ω(see Oksanen, 1991).

When dealing with complex data structures that may contain heteroscedasticity and correlation between observations, GLS is a potent tool. However, the covariance matrix must be specified precisely and is frequently estimated from the data using statistical methods or domain expertise. In the presence of correlated and heteroscedastic data, properly addressing these difficulties enables more accurate and effective parameter estimation.

3.5 Goodness of fit

In order to be assumed , the regression model goodness of fit was evaluated using the Akaike Information Criterion (AIC)(see Burnham & Anderson, 2004), the Bayesian Information Criterion (BIC)(see Hansen, 2007), the log-likelihood, the Mean Absolute Error (MAE), the Watanabe-Akaike Information Criterion (WAIC), and the Deviance Information Criterion (DIC). The conflict between the complexity of the model and its ability to fit the data is balanced by AIC, a metric of the model's goodness of fit. A model that fits the data better has a lower AIC. A lower BIC value, like AIC, denotes a better model fit. A measurement of how well a model fits the data is the log-likelihood, such that better model fit indicated by higher log-likelihood values. The average absolute difference between the expected and actual values measured by the MAE. Better, forecast accuracy indicated by a lower MAE. Model fit is measured using the WAIC and DIC, which considers both goodness of fit and model complexity. Better model fit is indicated by lower WAIC and DIC values.

3.6Bootstrapping Technique

In order to get a more precise estimation results from small samples, regression analysis and other statistical techniques make use of the effective resampling procedure bootstrapping(see Altman, 1989). It enables the calculation of model parameter uncertainty, prediction accuracy, and other statistical measures without assuming a particular data distribution. The estimate of the parameters of regression models is simple and effective with bootstrap-based confidence intervals. When bootstrapping, random samples taken from the source dataset and replaced to produce several "bootstrap samples." Similar in size to the original dataset, these bootstrap samples are most likely to include duplicate observations. A distribution of model parameters and variability of estimates can be determined by fitting the regression model to each bootstrap sample(see Hall, 1992). A 95% confidence interval estimate of the parameters of regression models evaluated and examined whether the observed estimates from the models were within the confidence limits. One hundred bootstrap samples' confidence intervals examined for this purpose.

4. Results

For the objective of the study, which examined agricultural production, five regression approaches used. Regression model assumption testing, model estimation, quality of fit, bootstrapping method to validity, and output forecasting were all included in the study analysis.

4.1 General results

Prior to data analysis, a general understanding of the data' fundamental statistical characteristics is required. The mean and standard deviation of production, area, temperature and humidity were observed to be 14.25(0.393), 12.14(0.183),28.2093(0.0731) and 48.5872(0.2905) respectively. The median, range, skewness and kurtosis were (14.06, 1.12,0.83, -0.97) for production, (12.07, 0.55, 0.91, -0.583) for cultivated area, (28.2242, 1.24, -0.812, 0.349) for temperature and (48.8423, 4.48, -0.568, -0.695) respectively. The distribution of all the variables was platykurtic and just slightly asymmetric.Various regression techniques applied and modelled forsimple and multiple regression models forproduction based on cultivated area, temperature and humidity. Goodness of fit tests examined for the models.In addition, detailed results were given in Tables 1 to 8 in the appendix.

4.2 Assumption testing results

4.2.1 Simple regression model

For simple model, production and area taken as the response and predictor variables respectively. The data points' close alignment with the diagonal line in the Q-Q plot in Figure 1 demonstrated that the normality assumption is plausible and that parametric statistical techniques that make this assumption are appropriate. We lack sufficient evidence to challenge the normality assumption because the Shapiro-Wilk p-value (0.1177) was higher than the usual significance level of 0.05. Therefore, $H_0$for testing normality of residuals accepted. Consequently, we concluded that the residuals did not significantly deviate from a normal distribution.

Fig1:Normal Q -Q plot for simple model

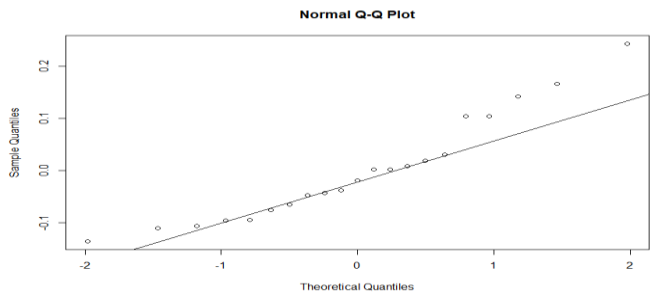
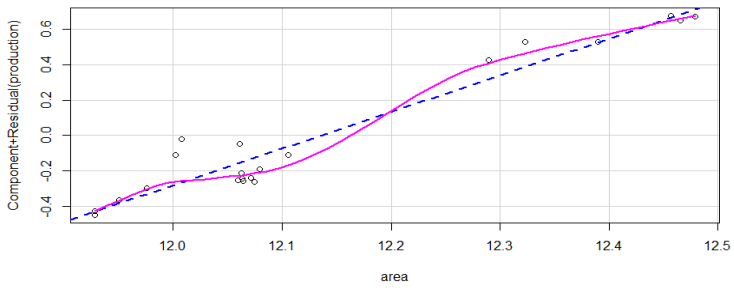


Fig. 2 illustrates the linearity of residuals using the regression plot to test the linearity assumption of the residuals. The plot showed that the data were slightly deviating from the linearity assumption. Rainbow test also showed a p-value of 0.0431, which confirmed this result. Therefore, the null hypothesis, H_0 that the residuals show a random distribution around zero with no discernible pattern, and the connection between the predictor variables and the response variable is linear, rejected at 5% level of significance.

Fig2: Regression plot for production residuals for simple model



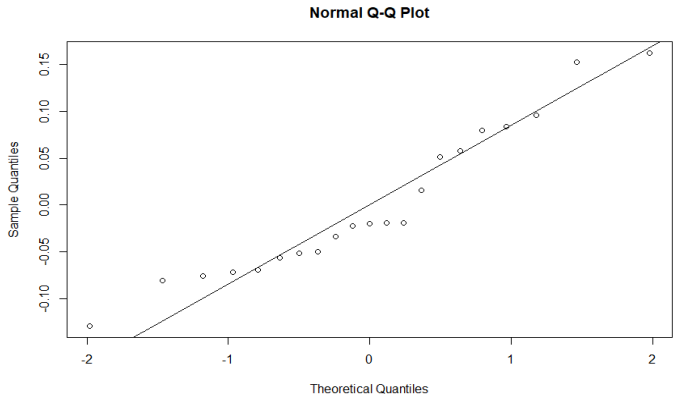
We lacked sufficient data to reject the null hypothesis$H_0$of homoscedasticity of residuals since the studentizedBreusch-Pagan test's p-value 0.08875, which was higher than the usual significance criterion of 0.05, did not show a difference from the nullhypothesis, H_0. As a result, according to the test, we were unable to locate any convincing evidence of heteroscedasticity in the multiple linear regression model. This showed that the model met the homoscedasticity assumption and that the assumption of constant variance of residuals not violated.

The p-value for the Durbin-Watson test was 0.0424, which showed that the observed autocorrelation differed marginally from the hypothesis.

4.2.2 Multiple regression model

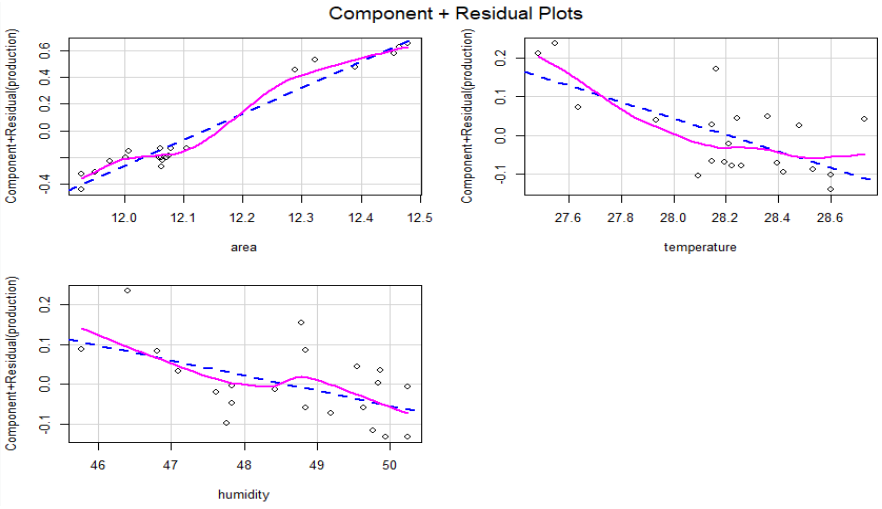
For the multiple model, production taken as response variable and area, temperature and humidity taken as the predictor variables. Since the data points in Figure 3's Q-Q plot closely matched the diagonal line, it was clear that the normality assumption was valid and that the parametric statistical methods used to support it were adequate. Because the Shapiro-Wilk p-value (0.1559) was higher than the customary significance level of 0.05, there is not enough data to refute the assumption of normality.Therefore, H_0 for testing normality of residuals accepted. The residuals did not significantly differ from a normal distribution, thus we got to that conclusion.

Fig3:Normal Q -Q plot for multiple model



The result was identical to the simple model, as indicated by the residual plot for multiple model in Figure 4. The Rainbow test p value was 0.0415, which indicated the marginal deviation from the linearity assumption. Therefore, the null hypothesis, H_0 for rainbow test rejected at 5% level of significance.

Fig4: Regression plot for production residuals for multiple model



Since the studentizedBreusch-Pagan test's p-value (0.2626) did not reveal a difference from the null hypothesis, we lacked the data to reject the null hypothesis,H_0,of homoscedasticity.

The Durbin-Watson test's p-value of 0.0445 indicated that the observed autocorrelation hardly deviated from the hypothesis. As a result, the Durbin-Watson test's null hypothesis, H_0, rejected at the 5% level of significance.

The variance inflation factor and tolerance for the predictor variables area, temperature and humidity were(1.250255, 1.361314, 1.571801) and (0.7998366, 0.7345844, 0.6362128) respectively. It revealed that multicollinearity is not a major issue. The condition number for the model was 1.254, which is quite close to one, indicating that multicollinearity is not severe.

Since some regression assumptions are marginally met the data, we have applied various regression techniques to finalize the conclusion for the study data.

4.3Model estimation results

4.3.1 Simple regression model

The results of the simple model's regression techniques summarized in Table1. Initially OLS regression analysis conducted for the simple regression model.The OLS regression showed an estimated intercept of -10.8493 with a standard error 1.5421 and the estimated coefficient for the 'cultivated area' variable is 2.0684 with a standard error 0.1271. There is significant evidence against the null hypothesis that there is no correlation between area and production, as seen by the extremely low p-value (1.29e-12). The average magnitude of the residuals indicated by the residual standard error, which is 0.1042.The dependent variable's variation explained by the model to a degree of 93.31%, as shown by the R-squared value of 0.9331. The OLS regression model equation written as:

Production = -10.8493 + 2.0684 ×Cultivated area + Error. (5)

WLS regression resulted similar to OLS as an intercept of -10.84 with a standard error of 1.54 and a coefficient of 2.07 with a standard error of 0.127. The estimated coefficient showed a substantial correlation between area and production with a high degree of confidence with a very small p value (8.199e-16). The R-squared value is 0.93. The average magnitude of the residuals indicated by the residual standard error, which is 0.5682.The WLS regression model equation is similar to the OLS model since they have the same coefficients:

Production = -10.84 + 2.07 × Cultivated area + Error. (6)

GLS regression produced results an intercept of -11.3389 with a standard error 1.6704 which was significant (p-value = 0.0092) and a coefficient estimate of 1.289217 with a standard error 0.2433 which is highly significant

(p-value<0.001) respectively. The R-squared value was 0.9330962. The residual standard error, which was 0. 6704.The GLS regression model equation written as:

Production = -11.3389 + 2.0992 × Cultivated area + Error. (7)

4.3.1 Multiple regression model

Table 5 provided an overview of the regression approaches' outcomes for the multiple regression model. OLS regression analysis initially performed for the multiple regression model. For the intercept, area coefficient, temperature coefficient, and humidity coefficient, the OLS regression indicated estimated values of -1.47059, 1.94494, -0.21424, and -0.03781, respectively, with significant values of 0.66139, 7.44e-12, 0.00567, and 0.05456. The findings indicated that area was significantly essential to production, whereas humidity was not. The residual standard error of the model was 0.08694. The adjusted R-squared value of 0.9583, demonstrated that the model explained the variation of the dependent variable to a degree of 95.83%. The OLS regression model equation written as:

Production = -1.471 + 1.945 × Cultivated area - 0.214 × Temperature - 0.038 × Humidity + Error. (8)

The WLS model showed all estimates same as OLS. The residual standard error of the model was 0.4743. The adjusted R-squared value was 0.9583, which was the same for OLS.The WLS regression model equation is similar to the OLS model since they have the same coefficients:

Production = -1.471 + 1.945 × Cultivated area - 0.214 × Temperature -0.038 × Humidity + Error. (9)

The GLS regression showed estimated values of -1.3419, 1.8467, -0.1879, and -0.03129 for the intercept, area coefficient, temperature coefficient, and humidity coefficient, respectively, with significant values of 0.6621, 0, 0.0114, and 0.0532. The results showed that, in contrast to humidity, area is highly significantly vital to production. The residual standard error of the model was 0.09959.Similar to OLS, the adjusted R-squared value was 0.9583.The GLS regression model equation written as:

Production = -1.342 + 1.847 × Cultivated area - 0.188 × Temperature -0.033 × Humidity + Error. (10)

4.4 Goodness of fit

4.4.1 Simple regression model

The OLS regression model seems to have a good fit to the data, with substantial coefficients and few significant assumptions not being satisfied. Table 2 presented the goodness of fit of all simple models. The values of AIC, BIC, MAE, Log-likelihood, WAIC and DIC for OLS regression model were observed as -31.49787, -28.3643, 18.74893, 0.07873778, 0.196414 and 0.2258248 respectively. In order to identify best regression model that fit the data well, other regression models were analysed and goodness of fits compared.

According to observations, the WLS model values for AIC, BIC, MAE, Log-likelihood, WAIC, and DIC were -31.498, -28.36, 18.75, 0.08223784, 0.196414, and 0.2258248 respectively.

Observations showed that the AIC, BIC, Log-likelihood WAIC, and DIC values for GLS modelwere respectively -30.95826, -28.1805, 18.47913, 0.3447875 and 0.4414273.

4.4.2 Multiple regression model

The OLS regression model seems to have a good fit to the data, with substantial coefficients and few significant assumptions being broken.Table 6 displayed the goodness of fit of all multiple models. The values of AIC, BIC, MAE, Log-likelihood, WAIC and DIC for OLS regression model were observed as -37.4263, -32.2037, 23.7132, 0.0666, 0.1224 and 0.01407 respectively. In order to identify best regression model that fit the data well, other regression models were analysed and goodness of fits compared.

WLS model showed all goodness of fit the same for GLS model.

Observations show that the GLS model values for AIC, BIC, MAE, Log-likelihood, WAIC, and DIC were, respectively, -22.5976, -17.5983, 17.2988, 0.0666, 0.13329, and 0.15325.

4.5 Bootstrapping

4.5.1 Simple regressionmodel

The bootstrapping results of simple models presented in Table 3. Observed estimates of the regression models found within the limits of 95% bootstrapping confidence intervals, which satisfied the validity of the regression estimates.

4.5.2 Multiple regressionmodel

Table 7 displayed the results of multiple models' bootstrapping. Itobserved that all models' estimates were within the 95% bootstrapping confidence interval, which satisfied the validity of all models' estimates.

4.6 Forecasting

4.6.1 Simple regressionmodel

Prediction intervals of agriculture production when area ranges between 12.5 and 12.7 by employing all the candidate simple models displayed in Table 4. The prediction interval appeared as valid, as the predicted limits were higher than observed as per the trend.

4.6.2 Multiple regression model

Prediction intervals of agriculture production when area ranges between 12.5 and 12.7, temperature ranges between 27 and 29 and humidity ranges between 46 and 50 using the multiple models were displayed in Table 8. Prediction intervals of all multiple models found to be appropriate for forecasting the data.

5. Conclusion

In this study for the simple models, it observed that all the regression models fit well for the data. All simple models showed a very high significance (p-value < 0.001) connection between production and area with very high R-squared values which concluded that the association between the variables was significantly high. In order to identify the most suited model, goodness of fit measures examined for each of the simple model and multiple model. Based on the goodness of fit values, OLS and WLS have the same AIC (-31.498) and very similar BIC values (-28.364 for OLS and -28.36 for WLS), which are higher than the GLS AIC and BIC values. GLS has the lowest log-likelihood (18.4791), indicating a lower fit to the data compared to OLS and WLS.OLS has the lowest MAE (0.07874), indicating that it has the smallest average absolute error in predicting the response variable. OLS and WLS have the same WAIC (0.19641), which is lower than the GLS WAIC. OLS and WLS have the same DIC (0.22583), which is lower than the GLS DIC.Overall, the OLS and WLS models appear to be doing better than the GLS model based on the several criteria taken into account. OLS has a little lower log-likelihood and MAE than WLS, which are crucial measures for evaluating the model fit. In light of these considerations, the OLS model seems to be the best suited of the three.

All multiple models fit the data well. Humidity was not significant for all the models. In multiple models, the WLS model had the largest residual standard error. The adjusted R-squared value was significant and very similar in all multiple models.OLS and WLS have the same AIC (-37.426) and BIC (-32.204), which are lower than the GLS AIC and BIC values. OLS and WLS have the highest log-likelihood (23.713), indicating a better fit to the data compared to GLS. All three models have the same MAE (0.067), indicating that they have the same average absolute error in predicting the response variable. OLS and WLS have the same WAIC (0.012), which is lower than the GLS WAIC (0.133). OLS and WLS have the same DIC (0.014), which is lower than the DIC (0.153) of GLS. Overall, based on the various criteria taken into account, the OLS and WLS models performing better the GLS model. The AIC, BIC, log-likelihood, MAE, WAIC, and DIC values for OLS and WLS are identical, demonstrating that both models are equally good fits for the data.

Bootstrapping method applied to all regression models to validate the estimates. Within the 95% bootstrapping confidence intervals, observed regression model estimates found except for GLS model, which satisfied the validity of the regression estimates.A 95% prediction interval of production observed valid for the models.

6. Limitations and future research directions

The three approaches (OLS, GLS, and WLS) each have strengths and weaknesses. Slight violation from model assumptions may affect the real output of the analysis. Outliers in the data can have a major impact on estimations of the regression coefficients and produce biased findings since OLS is sensitive to them.However, we have not seen any outliers in the dataset of our problem.

The selection of the right weights has a significant impact on WLS success. Selecting the wrong weights can result in estimates that are biased and ineffective.WLS's robustness is constrained when compared to some other regression methods, and it may not always adequately account for the impact of extreme outliers or other breaches of the assumptions. For this problem, we empirically obtain suitable weights for the model using R software.

The results may not be generalizable or applicable to a wider context if the sample selected for data collection is not representative of the intended population.The validity and dependability of the analysis may be affected by inaccurate or missing data.Accessing certain data sources can occasionally be difficult because of limitations, privacy issues, or property rights.Thus, partially we faced this problem, and we proposed the bootstrapping approach to assure the generalization of the results. Update the data basis of agriculture proposed to any country including the Sultanate of Oman. For future research direction, we hope to get the historical data, i.e., for long time, then the time series models are to be developed.

References

[1] Altman, D. G. (1989). Bootstrap investigation of the stability of a cox regression model. In *Statistics In Medicine* (Vol. 8).

[2] Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. In *Sociological Methods and Research* (Vol. 33, Issue 2). https://doi.org/10.1177/0049124104268644

[3] D. C. Montgomery, P. E. A., and G. G. V. (2021). Introduction To Linear Regression Analysis. *John Wiley & Sons, Inc New York, 6th Editio.*

[4] Dielman, T. E., Box, R. O., & Worth, F. (2005). *Least absolute value regression : recent contributions*. *75*(4).

[5] Draper, N. R., & Smith, H. (1998). *Applied regression analysis.* ((3rd ed.). (ed.)). USA: John Wiley & Sons, Inc.

[6] Flatt, C., & Jacobs, R. L. (2019). Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets. *Advances in Developing Human Resources*, *21*(4).

[7] Funes, I., Savé, R., Rovira, P., Molowny-Horas, R., Alcañiz, J. M., Ascaso, E., Herms, I., Herrero, C., Boixadera, J., & Vayreda, J. (2019). Agricultural soil organic carbon stocks in the north-eastern Iberian Peninsula: Drivers and spatial variability. *Science of the Total Environment*, *668*. https://doi.org/10.1016/j.scitotenv.2019.02.317

[8] Hall, P. (1992). On Bootstrap Confidence Intervals in Nonparametric Regression. In *Source: The Annals of Statistics* (Vol. 20, Issue 2).

[9] Halunga, A. G., Orme, C. D., & Yamagata, T. (2017). A heteroskedasticity robust Breusch–Pagan test for Contemporaneous correlation in dynamic panel data models. *Journal of Econometrics*, *198*(2). https://doi.org/10.1016/j.jeconom.2016.12.005

[10] Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4).

[11] Herwartz, H. (2006). Testing for random effects in panel data under cross sectional error correlation-A bootstrap approach to the Breusch Pagan test. *Computational Statistics and Data Analysis*, *50*(12). https://doi.org/10.1016/j.csda.2005.08.003

[12] Huang, F. L. (2018). Multilevel Modeling and Ordinary Least Squares Regression: How Comparable Are They? *Journal of Experimental Education*, *86*(2). https://doi.org/10.1080/00220973.2016.1277339

[13] Ii, E., Session, T. A., & Shimoshimizu, M. (2019). *Review of OLS Estimator*. 1–14.

[14] Kim, J., Cho, E., Okafor, C. E., & Choi, D. (2022). Does Environmental, Social, and Governance Drive the Sustainability of Multinational Corporation's Subsidiaries? Evidence From Korea. *Frontiers in Psychology*, *13*.

[15] Kim, T. H., & White, H. (2001). James-stein-type estimators in large samples with application to the least absolute deviations estimator. *Journal of the American Statistical Association*, *96*(454), 697–705. https://doi.org/10.1198/016214501753168352

[16] Koenker, R., Machado, J. A. F., Skeels, C. L., & Welsh, A. H. (1993a). Amemiya's form of the weighted least squares estimator. *Australian Journal of Statistics*, *35*(2). https://doi.org/10.1111/j.1467-842X.1993.tb01322.x

[17] Koenker, R., Machado, J. A. F., Skeels, C. L., & Welsh, A. H. (1993b). AMEMIYA'S FORM OF THE WEIGHTED LEAST SQUARES ESTIMATOR. *Australian Journal of Statistics*, *35*(2).

[18] Lechene, V., Pendakur, K., & Wolf, A. (2022). Ordinary Least Squares Estimation of the Intrahousehold Distribution of Expenditure. *Journal of Political Economy*, *130*(3). https://doi.org/10.1086/717892

[19] Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. *Clinical Chemistry*, *39*(3). https://doi.org/10.1093/clinchem/39.3.424

[20] Majewski, S., Mentel, G., Dylewski, M., & Salahodjaev, R. (2022). Renewable Energy, Agriculture and CO2 Emissions: Empirical Evidence From the Middle-Income Countries. *Frontiers in Energy Research*, *10*.

[21] Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika*, *79*(4), 747–754. https://doi.org/10.1093/biomet/79.4.747

[22] Nau, R. (2015). Regression diagnostics: testing the assumptions of linear regression. *Fuqua School of Business, Duke University*, *i*.

[23] Njegomir, V., Pejanovic, L., & Kekovic, Z. (2017). Agricultural entrepreneurship, environmental protection and insurance. *Ekonomika Poljoprivrede*, *64*(3). https://doi.org/10.5937/ekopolj1703035n

[24] Oksanen, E. H. (1991). A Simple Approach to Teaching Generalized Least Squares Theory. *The American Statistician*, *45*(3), 229–233. https://doi.org/10.2307/2684297

[25] Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*(2). https://doi.org/10.1111/j.2517-6161.1969.tb00796.x

[26] Romano, J. P., & Wolf, M. (2017). Resurrecting weighted least squares. *Journal of Econometrics*, *197*(1).

[27] Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *MPR-Online*, *8*.

[28] Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39–42. https://doi.org/10.12691/ajams-8-2-1

[29] Souza, S. V. C. De, & Junqueira, R. G. (2005). A procedure to assess linearity by ordinary least squares method. *Analytica Chimica Acta*, *552*(1–2). https://doi.org/10.1016/j.aca.2005.07.043

[30] Sulewski, P., Majewski, E., & Wąs, A. (2017). THE IMPORTANCE OF AGRICULTURE IN THE RENEWABLE ENERGY PRODUCTION IN POLAND AND THE EU. *Problems of Agricultural Economics*, *350*(1). https://doi.org/10.30858/zer/82999

[31] Sumberg, J. (2022). Future agricultures: The promise and pitfalls of a (re)turn to nature. *Outlook on Agriculture*, *51*(1). https://doi.org/10.1177/00307270221078027

[32] Xu, X., Chen, S., Ren, L., Han, C., Lv, D., Zhang, Y., & Ai, F. (2021). Estimation of heavy metals in agricultural soils using vis-nir spectroscopy with fractional-order derivative and generalized regression neural network. *Remote Sensing*, *13*(14). https://doi.org/10.3390/rs13142718

[33] Zhang, H., & Mei, C. (2011). Local least absolute deviation estimation of spatially varying coefficient models: Robust geographically weighted regression approaches. *International Journal of Geographical Information Science*, *25*(9), 1467–1489. https://doi.org/10.1080/13658816.2010.528420

[34] Zulkifli, R., Aimran, N., Deni, S. M., & Badarisam, F. N. (2022). A comparative study on the performance of maximum likelihood, generalized least square, scale-free least square, partial least square and consistent partial least square estimators in structural equation modeling. *International Journal of Data and Network Science*, *6*(2). https://doi.org/10.5267/j.ijdns.2021.12.015

Appendix

Table 1:Regression estimates of simple models

* and ** denote statistical significance at the 5% and 1% level, respectively

	Regression Estimates			
Regression Model	**Intercept (Standard Error)**	**Coefficient (Standard Error)**	**R-Squared**	**Residual Standard Error**
OLS	-10.8493(1.5421)**	2.0684 (0.1271)**	0.9331**	0.1042
WLS	-10.84 (1.54)**	2.07 (0.127)**	0.93**	0.5682
GLS	-11.3389 (1.67)**	2.0992 (0.143)**	0.9331**	0.6704

Table 2: Goodness of fit of simple models

	Goodness of fit					
Regression Model	**AIC**	**BIC**	**Log-likelihood**	**MAE**	**WAIC**	**DIC**
OLS	-31.498	-28.364	18.7489	0.07874	0.19641	0.22583
WLS	-31.498	-28.36	18.75	0.08224	0.19641	0.22583
GLS	-30.958	-28.181	18.4791	0.09345	0.34479	0.44143

Table 3: Bootstrap Confidence Interval of simple model regression estimates

	95% bootstrap Confidence Interval of Regression estimates	
Model	Intercept	Coefficient
OLS	(-13.0748 , -8.6617)	(1.31 , 49.76)
WLS	(-12.9209 , -8.4771)	(1.87 , 50.18)

GLS	(-13.1430 , -8.6749)	(1.92 , 68.47)

Table 4: Prediction intervals of agriculture production when area ranges between 12.5 and 12.7 by employing the candidate simple models

	95% Prediction Interval of regression models	
Model	**Lower bound**	**Upper bound**
OLS	14.76257	15.68863
WLS	13.81167	16.61936
GLS	14.81446	15.78954

Table 5: Regression estimates of multiple models

	Goodness of fit					
Regression Model	**AIC**	**BIC**	**Log-likelihood**	**MAE**	**WAIC**	**DIC**
OLS	-37.426	-32.204	23.713	0.067	0.012	0.014
WLS	-37.426	-32.204	23.713	0.067	0.012	0.014
GLS	-22.598	-17.598	17.299	0.067	0.133	0.153

* and ** denote statistical significance at the 5% and 1% level, respectively

Table 6: Goodness of fit of multiple models

	95% bootstrap Confidence Interval of Regression estimates			
Model	**Intercept**	**Area**	**Temperature**	**Humidity**
OLS	(-13.256, 5.153)	(1.675, 2.271)	(-0.346, 0.079)	(-0.068,-0.006)
WLS	(-15.084, 5.45)	(1.740,2.297)	(-0.325,0.056)	(-0.074,0.035)
GLS	(-12.515,6.518)	(1.682, 2.217)	(-0.344,-0.019)	(-0.069,-0.003)

Table 7: Bootstrap Confidence Interval of multiple model regression estimates

Regression Estimates						
Regression Model	**Intercept (Standard Error)**	**Area (Standard Error)**	**Temperature (Standard Error)**	**Humidity (Standard Error)**	**Multiple Adjusted R-Squared**	**Residual Standard Error**
OLS	-1.471 (3.29)	1.945(0.119)**	-0.214(0.068)**	-0.038(0.018)	0.958**	0.0869
WLS	-1.471 (3.29)	1.945(0.119)**	-0.214(0.068)**	-0.038(0.018)	0.958**	0.4743
GLS	-1.342(3.02)	1.847(0.181)**	-0.188(0.066)*	-0.033(0.015)	0.958**	0.0996

Table 8: Prediction intervals of agriculture production when Area rangesbetween 12.5 and 12.7, Temperature ranges between 27 and 29 and Humidity ranges between 46 and 50, by employing thecandidate multiple models

	95% Prediction Interval of regression models	
Model	**Lower bound**	**Upper bound**
OLS	14.762	16.002
WLS	14.001	16.733
GLS	14.852	15.557