

An assignment on

final project on

**Prediction of Camelina Seed Yield Using Statistical Regression Analysis Techniques**

(Applied Regression Analysis)

STAT 757, Spring 2018

**Submitted to:**

Grant Schissler, PhD

Assistant Professor,

Department of Mathematics and Statistics

University of Nevada, Reno

**Submitted by:**

DhurbaNeupane

BiplavTimalsina

Graduate Student,

University of Nevada, Reno

**May 15, 2018**

**Prediction of Camelina Seed Yield Using Statistical Regression Analysis Techniques**

**1. Introduction**

In water-deficit environments like the US state of Nevada, identifying and selection of crop types suitable for such environments can be very challenging. Alfalfa (*Medicago sativa* L.) is the number one crop grown in Nevada both in land area cultivated (101,171 hectares) and value in dollars from tonnage produced (an estimated two hundred million dollars) on an annual basis [1]. However, alfalfa is a high-water use crop and hydrological instability is a common and direct hindrance to continued production of alfalfa and thus Nevada’s agricultural production and sustainability. Therefore, implementing a strategy of crop diversification [2] can reduce producer’s vulnerability and accelerates agricultural productivity in water-limited agroecosystems like Nevada.

Camelina (*Camelina sativa* [L.] Crantz) is an emerging oilseed crop of the *Brassicaceae* family and which has the unique attribute of multiple uses. Among the many uses of camelina, the most frequent focus is on its value as a high potential biofuel feedstock mainly biodiesel and jet fuel [3-6], a high protein animal feed for different classes of livestock [7-10] are among the major focus of its cultivation globally. Camelina positive net energy balance and overall net lower emissions of greenhouse gasses [11-13], its low water demand [14-16], low fertilizer requirements [17], and resistance to many agricultural insect pests and diseases [14] makes camelina an ideal alternative candidate crop for diversification and the potential for a permanent niche in the crop production cycle of water-scarce environments like Nevada.

From an agronomic management perspective, grain yield of non-leguminous crops is limited by soil nitrogen (N) availability, and crop genotype Soils in Nevada are typically low in organic matter and therefore, low total soil N concentration is a widespread occurrence. Though camelina has been considered as a low fertilizer input crop, for obtaining optimum yields, N fertilizer must be applied in accordance with the yield potential of camelina, water use, source, availability, and climatic conditions of the area cultivated. Further, inorganic N fertilizer is a major variable input cost in crop production agriculture and overuse of N will substantially increase the overall production cost and hence reduce the cost-competitiveness of camelina as a biofuel crop.

Further, the influence of controlled-release N fertilizer on camelina grain production in Nevada’s semiarid climate is not known. The quantity and source of N fertilizer applied are often timesdictated by genotype of the crop, existing soil nutrient supply capabilities, water supply, type of water used, and the intended use of the crop. Different statistical techniques have been applied for predicting yield including multiple linear regression and correlation matrix.

The objective of our study was to evaluate the effects of nitrogen application rate, nitrogen source, and genotypes on grain yield of camelina and, to estimate stepwise regression analysis for seed yield to develop suitable model.

**Materials and methods**

To evaluate the nitrogen rates, sources and camelina genotype effects on seed yield and to perform stepwise regression analysis, a randomized complete block design(RCBD) experiment with four replications, a two-year field experiment wascarried out at the University of Nevada, Reno Main Station Field Laboratory, Reno NV (39º30’ N, 119º44’ W, and altitude of 1339 m)during the spring to early-summer (March-July) growing seasons of 2016and 2017.The predominant soil at the experimental siteis classified as a Truckee silt loam (a fine-loamy, mixed, superactive, mesic Fluvaquentic Haploxerolls). Soil chemical characteristics was analyzed at a commercial laboratory. There were some variations in soil chemical characteristics between the two site-year (Table 1). Weather data were collected from the Western Regional Climate Center, Desert Research Institute Weather Station located at the experimentalsite approximately 1000 m away from research plots(Table 2). Monthly evapotranspiration (ET) did not deviate widely over the two years of the study and total supplemental irrigation of 470.1 and 481.3 mm was provided during the 2016 and 2017 growing seasons respectively (Fig. 1). Treatments were four N application rates (0, 40, 80 and 120 kg N ha-1), two N sources [conventional urea (CU) and polymer-coated urea(PCU)], and two camelina genotypes (‘Blaine Creek’ and ‘Pronghorn’). Crop management factors including land preparation, other fertilizer application including phosphorus and potash, and weeds controlled was performed based on the recommendation.

Grain yield was quantified at physiological maturity (when >90% of the pods had changed color to brown) by harvesting an area of 5.6 m2 in each experimental unit after border rows were removed using a Kincaid plot combine (Kincaid Equipment Manufacturing, Haven, KS, USA). Thereafter, the harvested seeds from each plot was cleaned separately using a Clipper Office Tester Cleaner (Clipper, Bluffton, IN, USA) and weighed to compute grain yield. Seed samples of 5 g from each plot was then oven dried at 60ºC using an Isotemp forced-air oven (Fisher Scientific, Hampton, NH, USA) to determine seed dry matter. Grain yield was adjusted to 92% dry matter.

**Statistical analysis**

Data were analyzed using R programing. In this study, the general methods for developing the regression models for crop yield prediction is discussed. Regression analysis is generally used for the prediction purposes as it predicts the dependent (response) variable as a function of independent variables. A scatter plot matrix of response and predictor variables were plotted to see if there were any obvious pattern of our data. Box plot of seed yield against each of the five predictors (Nitrogen rate, Block, Year, nitrogen sources and camelina genotypes)was plotted to see the distribution of the predictor variables. Stepwise regression analysis was performed using all possible subsets, and we were able to develop a model using maximum value R2adj, and minimum values of AIC, AICC, and BIC. Standardized residuals of significant predictors were generated to ensure the assumption of constant variance. Again, the diagnostic plots from R were also plotted to see if there is any non-random pattern in the data. Furthermore, marginal model plots were also generated to see the model fit. Fake data was also generated and fit the regression model, fit model on real data, then plot regression line and finally plot diagnostic to check the model fit.

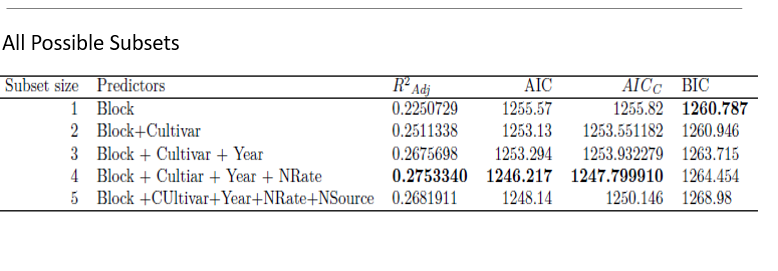
Maximum log likelihood approach was also performed for model selection using backward selection approach (Using SAS programming, codes not shown here). In this approach, at each step, the variables with highest P-value were removed from the full model one at a time. The process was repeated until all the p-values for the variables in the model were less than the prespecified cutoff. Values of p < 0.05(cutoff value) were considered statistically significant.

**Results and Discussion**

A scatter plot matrix of seed yield and predictor variables are shown in Figure 1(Appendices). Because our data are mostly categorical, we cannot see anything obviously distinctive pattern from the scatter plot. Thus, we moved forward to plot the box plot of the y-variable for various categorized predictor variables as shown in Figure 2 (Appendices). From the first figure of Seed yield and nitrogen rates, we see the obvious linear pattern of increase of seed yield as nitrogen rates increase from N00, N40, N80 and N120 kg N ha-1. This confirms our notion that seed yieldshould increase with the increase in nitrogen rate. The figure of seed yield vs year shows that there is greater seed yield in 2016 compared to 2017. The variation in seedyield was attributed to late planting of camelina, delayed by flood and snow in early January in 2017. The figure of seedyield vs cultivar shows that cultivar Blaine creek had greater seed yield compared to Pronghorn. We observe no difference in seed yield between two nitrogen sources. We also see some variation in seedyield in different blocks, which might be due to uneven distribution of water, variation in slope and soil texture.

After preliminary analysis, we move forward with performing model selection using various regression techniques. One of the most favored technique called all possible subset is preferred and is used here because it check for all possible subset of the given predictor variables to predict the seed yield. The results of analysis for model selection using possible subsets with the values of R2 Adj, AIC, AICC and BIC is shown in table 3

Table 3 Values of R2 Adj, AIC, AICC and BIC for the best subset of each size



Various information criteria used to model selection suggests that the best model is the one with four predictors. Highlighted bold letter This subset had the highest R2Adj, and lowest AIC and AICc, and comparable BIC.

The model observed is shown below,

**Yield= 480353.88 + 185.47\*Block -265.64\*CultivarProghorn -238.1\*Year**

**+ 407.64\* Nrate120 + 60.48\*NRate40 + 376.31\* NRate80.**

We also observe that the variance inflation factor (VIF:Block: 1.01, cultivar: 1.0, year: 1.0, NRate: 1.01), and none of them exceeds 5, that means there is no multicollinearity effect.

Hence, we can accept this model as our final model 1.

Again, we perform Maximum Log Likelihood test using backward selection approach. The results from this analysis is shown in Table 4 below.

Table 4. Log Likelihood Parameters for various model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Terms | -2log-Lik | Chi-square | df | p-value |
| 1 | Block2+Block3+Block4+BlainCreek+N40+N80+N120+N2+2017 | -962.0409 |  |  |  |
|  | Model1-Block2-Block3-Block4 | -974.52 | 24.96 | 1 | 5.8e-07 |
|  | Model1-BlaineCreek | -967.51 | 10.94 | 1 | 0.0009 |
|  | **Model1-Year** | **-962.041** | **0.002** | **1** | **0.964** |
|  | Model1-NRate | -971.59 | 19.1 | 1 | 1.23e-05 |
|  | Model1-N2 | -962.042 | 0.0022 | 1 | 0.962 |
|  |  |  |  |  |  |
| 2 | Block2+Block3+Block4+BlainCreek+N40+N80+N120+N2 | -962.041 |  |  |  |
|  | -Model 2 - Block | -978.19 | 31.95 | 1 | 1.5e-08 |
|  | **-Model 2- BlaineCreek** | **-971.51** | **18.94** | **1** | **1.34e-05** |
|  | -Model 2- N40-N80-N120 | -975.46 | 26.84 | 1 | 2.2e-07 |
|  | -Model 2- N2 | -966.47 | 8.86 | 1 | 0.0029 |

Observing the p-values for different models, we conclude that a model with 4 predictors including Block, Blaine Creek, NRate and NSource are the good predictor of seed yield.

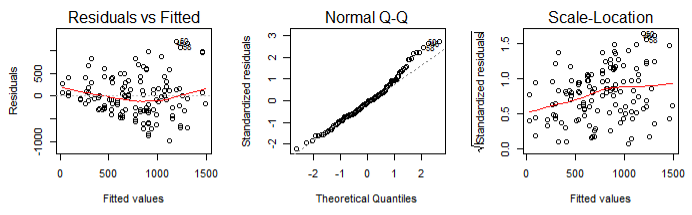
Model2:

**Yield = 134.28 + 265.63\*BlaineCreek + 406.24 \*N120 +61.54\*N40 +380.9\* N80-8.02\*N2 + 205.1685 \*B2 + 417.38\*B3+546.41\*B4**

The Variance Inflation factor (VIF: Block: 1.13, Cultivar:1.0, NRate: 1.08, Nsource: 1.04) are all less than 5, hence it exhibits no multicollinearity effect.Thus, we can also use this model as our second final Model. The standardized residuals of seed yield against each of the significant predictors from this model shows the standardized residuals are random in nature and assumption of constant variance is verified (figures are not shown here).

After model development, we have performed model validation. Figure 3 shows the standardized residuals of final Model 1. The plot of residuals vs the fitted values is somewhat non-random, but it could be accepted as it is not too obvious. There are some outliers (46,50,58) but none of them are bad leverage points. Marginal model plots of the same good alignment of lines suggesting that the model is a great fit.

Figure 3 Standardized residuals of significant predictors



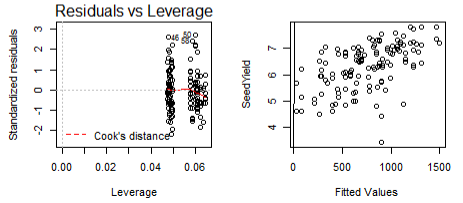
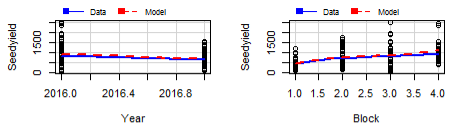


Figure 4 Diagnostic Plots: Marginal model plots(mmp) of seed yield against year and block



We also generated fake data using the model1. The diagnostic plots area shown in Figure5, that shows pretty good fit and the intercept generated also contain the actual intercept of the model.

Regression output from R is shown below.

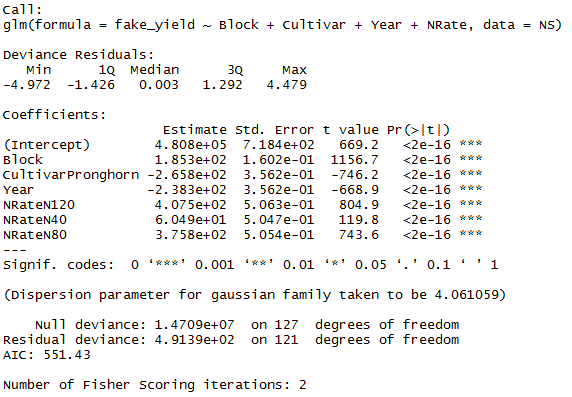
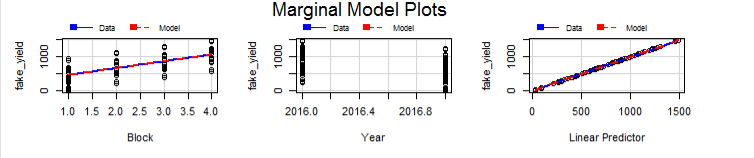


Figure 5 Diagnostic plots from the fake data generated



**Conclusion**

Using the regression analysis techniques, we were able to predict the seed yield using various predictors. We have proposed two different regression models to predict seed yield using different predictor variables. In the model with all possible subsets, four predictors including nitrogen rates, block, year, and cultivars are the best predictors. However, in the model developed using the maximum log likelihood ratio from backward selection approach, our significant predictors are nitrogen rates, block, nitrogen sources and cultivar. This analysis is limited to only five predictor variables to predict seed yield, this might mislead our findings. Because of limited observations, our result might be overfitted. Further research is recommended at other sites to see spatial autocorrelation effect. Further analysis with mixed effect model is also recommended if we will be able to generate another model.

**Acknowledgement**

We would like to thank our Prof Grant Schissler who assisted in every step of our project.

REFERENCES

[1] 2017 State agriculture overview Nevada, crops - planted, harvested, yield, production, price (MYA), value of production sorted by value of production in dollars, accessed online: https://www.nass.usda.gov/Quick\_Stats/Ag\_Overview/stateOverview.php?state=NEVADA (2018).

[2] P.F. McCord, M. Cox, M. Schmitt-Harsh, T. Evans, Crop diversification as a smallholder livelihood strategy within semi-arid agricultural systems near Mount Kenya, Land Use Policy 42 (2015) 738-750.

[3] H.M. Paulsen, V. Wichmann, U. Schuemann, B. Richter, Use of straight vegetable oil mixtures of rape and camelina as on farm fuels in agriculture, Biomass and Bioenergy 35 (2011) 4015-4024.

[4] C. Ciubota-Rosie, J.R. Ruiz, M.J. Ramos, Á. Pérez, Biodiesel from *Camelina sativa*: A comprehensive characterization, Fuel 105 (2013) 572–577.

[5] C.M.H. Keske, D.L. Hoag, A. Brandess, J.J. Johnson, Is it economically feasible for farmers to grow their own fuel? A study of Camelina sativa produced in the western United States as an on-farm biofuel, Biomass and Bioenergy 54 (2013) 89-99.

[6] Y.C. Liu, A.J. Savas, C. T. Avedisian, The spherically symmetric droplet burning characteristics of Jet-A and biofuels derived from camelina and tallow, Fuel 108 (2013) 824–832.

[7] T. Jaśkiewicz, A. Sagan, I. Puzio, Effect of the Camelina sativa oil on the performance, essential fatty acid level in tissues and fat-soluble vitamins content in the livers of broiler chickens, Livestock Sci. 165 (2014) 74-79.

[8] J. Pikul, J. Wójtowski, R. Dankówa, J. Teichert, G. Czyżak-Runowska, D. Cais-Sokolińska, A. Cieślak, M. Szumacher-Strabel, E. Bagnicka, The effect of false flax (*Camelina sativa*) cake dietary supplementation in dairy goats on fatty acid profile of kefir, Small Ruminant Research 122 (2014) 44-49.

[9] P.A. Adhikari, J.M. Heo, C.M. Nyachoti, standardized total tract digestibility of phosphorus in camelina (*Camelina sativa*) meal fed to growing pigs without or phytase supplementation, Animal Feed Sci. Technol. 214 (2016) 104-109.

[10] C.N. Bullerwell, S.A. Collins, S.P. Lall, D.M. Anderson, Growth performance, proximate and histological analysis of rainbow trout fed diets containing *Camelina sativa* seeds, meal (high-oil and solvent-extracted) and oil, Aquaculture 452 (2016) 342-350.

[11] A.C. Drenth, D.B. Olsen, P.E. Cabot, J.J. Johnson, Compression ignition engine performance and emission evaluation of industrial oilseed biofuel feedstocks camelina, carinata, and pennycress across three fuel pathways, Fuel 136 (2014) 143-155.

[12] J.A. Dyer, X.P.C. Vergé, R.L. Desjardins, D.E. Worth, B.G. McConkey, The impact of increased biodiesel production on the greenhouse gas emissions from field crops in Canada, Energy for Sustainable Development 14 (2010) 73-82.

[13] S.R. Fore, W. Lazarus, P. Porter, N. Jordan, Economics of small-scale on-farm use of canola and soybean for biodiesel and straight vegetable oil biofuels, Biomass Bioenergy 35 (2011) 193-202.

[14] U. Iskandarov, H.J. Kim, E.B. Cahoon, Camelina: An emerging oilseed platform for advanced biofuels and bio-based materials, M.C. McCann, M. Buckeridge, N. Carpita (eds.), Plants and Bioenergy, Advances in Plant Biology 4 (2014) 131-140, Springer, New York, NY.

[15] A.K. Obour, H.Y. Sintim, E. Obeng, V.D. Jeliazkov (Zheljazkov), Oilseed camelina (Camelina sativa L. Crantz): Production Systems, Prospects and Challenges in the USA Great Plains, Adv. Plants Agric. Res. 2 (2) (2015) 00043.

[16] N. George, S.E. Thompson, J. Hollingsworth, S. Orloff, Measurement and simulation of water-use by canola and camelina under cool-season conditions in California, Agricultural Water Management 196 (2018) 15-23.

[17] D.H. Putnam, J. Budin, L.A. Field, W.M. Breene, Camelina: A promising low-input oilseed. In: J. Fanick, J.E. Simon (eds) New Crops, Wiley, New York, (1993) 314-322

[18] Sheather, S.J,2007. A Modern Approach to Regression with R. Textbook website: http://www.stat.tamu.edu/~Sheather /book/.

Appendices:

Table 1. Soil chemical characteristics at the research site prior to seeding camelina at the University of Nevada, Reno Main Station Field Laboratory Reno, NV during 2016 and 2017 growing seasons.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Organic Matter  g kg-1 soil | pH | CEC  Meq 100 g-1 | NO3-N | P | K | Ca | Mg | Na | SO4-S |
| mg kg-1 soil | | | | | | | |
| 2016 | 47.0 | 7.8 | 25.9 | 14 | 19 | 560 | 3539 | 598 | 437 | 95 |
| 2017 | 63.0 | 7.9 | 23.9 | 7 | 44 | 670 | 3401 | 538 | 180 | 9 |

Table 2. Monthly accumulated precipitationmean air temperature, and solar radiation, at the University of Nevada, Reno Main Station Field Laboratory Reno, NV during 2016 to 2017 growing seasons and 30-yr average (1987-2017).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Months | Monthly Precipitation  (mm) | | | Mean Air  Temperature (°C) | | | Mean  Solar Radiation Kw m-2 | |
|  | 2016 | 2017 | 30-yr average (1987-2017) | 2016 | 2017 | 30-yr average (1987-2017) | 2016 | 2017 |
| March | 17.0 | 20.0 | 24.15 | 7.4 | 7.5 | 13.6 | 362 | 351 |
| April | 30.7 | 23.3 | 18.14 | 10.6 | 9.2 | 17.3 | 483 | 456 |
| May | 28.5 | 17.5 | 22.31 | 13.7 | 14.8 | 21.8 | 536 | 568 |
| June | 0.00 | 3.56 | 18.02 | 20.0 | 19.6 | 25.9 | 619 | 605 |
| July | 0.00 | 0.76 | 12.33 | 22.0 | 23.2 | 28.0 | 616 | 589 |
| Total | 76.2 | 65.1 | 94.95 |  |  |  | 2616 | 2569 |

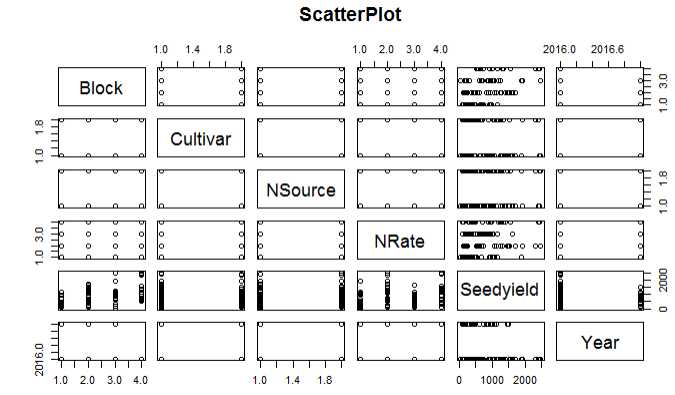
Figure 1. A scatter plot matrix of seed yield and predictor variables

Figure 2 Box plot of seed yield against each of the five predictors (Nitrogen rate, Block, Year, nitrogen sources and camelina genotypes)

