

COMPARING YIELD BETWEEN CORN SEED VARIETALS

RELEVANT JMP PLATFORMS AND STATISTICAL TECHNIQUES

Distribution :	Histogram ; Summary Statistics
Graph Builder :	Comparative Dotplots ; Summary Statistics ; Confidence Intervals for Means ; Scatterplot, Linear Regression
Multivariate :	Correlations and Matrix Scatterplot
Fit Y by X :	One Factor ANOVA ; One Variable Regression
Fit Model :	Multiple Variable Regression ; Stepwise (Forward Selection and Backward Elimination)

PROBLEM STATEMENT

Agricultural companies continuously work on developing seed varieties that improve the performance of a crop such as yield, plant health, resistance to disease and pests, temperature tolerance, among others.



A seed breeding research team for one such company is interested in comparing the yield between three corn seed varieties they produce (Standard, High Yield, and High Yield Plus).

Data from 161 fields throughout the midwestern United States was gathered. 47 fields used the Standard seed, 70 used the High Yield variety, and 44 used the High Yield Plus variety.

The primary objectives of the analysis are:

1. Determine if sufficient statistical evidence exists demonstrating differences in Yield between the Seed Varietals, and to quantify those differences.
2. Determine if other growing conditions variables (Soil Quality, Fertilizer, % Sun, Rainfall, and Number of Irrigations) have a significant effect on Yield and can be accounted for in the estimates comparing Yield between the Seed Varietals.
3. Determine if the effect those other variables have on Yield is different based upon the Seed Varietals.

DATA SET

Comparing Yield Between Corn Seed Varietals.jmp

Yield	Amount of corn harvested at location (kg per hectare)
Seed Varietal	One of the three seed varieties used (Standard, High Yield, High Yield Plus)
Soil Quality	Index quantifying the quality of the soil based on a variety of physical, chemical, and biological characteristics (50-100)
Fertilizer	Amount of fertilizer used (kg per hectare)
% Sun	Percentage of time sunlight reached ground during daylight hours throughout growing season
Rainfall	Amount of rain during growing season (mm)
Irrigations	Number of irrigations applied

EXERCISES

The exercises consist of building four different statistical models used to compare and quantify the differences in Yield between the three Seed Varietals. This will be done first considering only the factor Seed Varietal via a One-Factor Analysis of Variance (ANOVA) model. You'll then create three other models that account for the impact on Yield that growing conditions could also have via Multiple Variable Regression. Different model building methodologies will be used and each will result in a different set of model terms selected. You will compare these models with regard to their performance in describing the data, their interpretability, and in their usefulness in addressing the fundamental objectives of the analysis.


There are two concepts that are very important to keep in mind when building statistical models. One was aptly summarized in a quote from the famous statistician George Box: "*All models are wrong but some are useful*". That is, models should not be thought of as an exact mathematical description of the true relationships between a set of variables, but merely useful approximations to those relationships. Usefulness in one application may prompt selecting one particular model that describes the data but a different model might be chosen for another application.

The other important concept in model building is the idea of model parsimony. A highly effective approach is to not use a model that is more complex than is necessary for a particular application. This concept is expressed in the quote from Albert Einstein: "*Everything should be made as simple as possible, but not simpler*". A parsimonious model is one that accomplishes a desired level of usefulness for a particular application with as few variables as possible and without unnecessary mathematical complexity.

1. Summarize each of the numeric variables with histograms and summary statistics to develop some initial understanding of the data. Are there any anomalies seen in the data such as extreme outliers?

Instructions: Analyze > Distribution. Select the variables 'Yield', 'Soil Quality', 'Fertilizer', '% Sun', 'Rainfall', and 'Irrigations' for the Y role.


2. Create comparative dotplots of the Yield for each Seed Varietal. Add the sample means and standard deviations for each. Add 95% Confidence Intervals for the means. What is your initial assessment of the differences in Yield between the three Seed Varietals?

Instructions: Graph > Graph Builder. Place 'Yield' on the Y axis and 'Seed Varietal' on the X. Select Line of Fit from the graph palette.  95% Confidence Intervals for the mean will be added. Select to display the Means and Std Devs from the Line of Fit controls on the left.

3. **(Model 1)** Fit a One-Factor ANOVA Model estimating the average Yield for each Seed Varietal. Examine the F-Test and p-value testing the equality of the average Yield. Is there statistical evidence demonstrating non-equality of the means? Provide an interpretation of each of the 95% Confidence Intervals. Interpret the Root Mean Square Error (RMSE) and R^2 value. Conduct a multiple-comparison analysis for each pairwise difference (Standard vs High Yield, Standard vs High Yield Plus, and High Yield vs High Yield Plus). Is there significant statistical evidence (at 0.05 level) indicating that all the means are different from each other? Is there significant statistical evidence at the 0.10 level? Provide an interpretation of each of the 95% Confidence Intervals for the pairwise differences.

Instructions: Analyze > Fit Y by X. Choose 'Yield' for Y and 'Seed Varietal' for X. Click OK. Under the top red triangle, select Means/ANOVA. Then select Compare Means > Each Pair, Student's t.

4. Create scatterplots of Yield vs each of the growing conditions variables. Which variables appear to have the strongest relationship to Yield? Briefly describe those relationships.

Instructions: Graph > Graph Builder. Place 'Yield' on the Y axis and 'Soil Quality' on the X. Select Line of Fit from the graph palette.  A 95% Confidence Interval will be added. Select Redo > Column Switcher from the top red triangle. Select 'Soil Quality' from the Initial Column list, and select 'Soil Quality', 'Fertilizer', '% Sun', 'Rainfall', 'Irrigations' for the Replacement Columns. Use the Column Switcher to view each growing condition variable on the X axis.

5. Calculate the correlations and create scatterplots for all combinations of the growing condition variables. Which pairs of variables are most correlated? Provide a possible explanation of why this might be. Examine the variables 'Rainfall' and 'Irrigations'. What information is missing from the 'Irrigations' variable that would be more useful to have than number of irrigations, possibly allowing it to be combined with 'Rainfall' to form a single new variable?

Instructions: Analyze > Multivariate Methods > Multivariate . Select 'Soil Quality', 'Fertilizer', '% Sun', 'Rainfall', 'Irrigations' as the Y variables. Select Shaded Ellipses under the red triangle next to the Scatterplot Matrix title.

6. Perform a simple linear regression analysis of Yield vs each of the growing condition variables. Which variables have a statistically significant relationship to Yield at the 0.05 significant level? Are there any that do not have a significant relationship at the 0.05 but do at the 0.10 level? For those variables that do not have a significant relationship at either level, why is it not correct to conclude that these variables don't have any impact at all to Yield? For those considered significant at the 0.10 level, provide an interpretation of the equation, RMSE, and R-Squared values.

Instructions: Analyze > Fit Y by X. Choose the variable 'Yield' for the Y and 'Soil Quality', 'Fertilizer', '% Sun', 'Rainfall', 'Irrigations' for the X. Under the top red triangle next to each graph, select Fit Line.

Note: Holding the Command Key (Mac) or Ctrl Key (Windows) while selecting an option under the red triangle for the first variable will execute that operation for all the others.

7. **Model 2.** Create a multiple regression model for Yield using the variable Seed Varietal and only those variables you identified in Exercise 6 as significant for the explanatory variables. The final model will be the result of removing any terms not considered to be significant in a multiple variable regression that is accounting for the other variables.

- a. Initial model.

Instructions: Analyze > Fit Model. Choose the variable 'Yield' for the Y. Choose Minimal Report under Emphasis. Highlight the variables 'Seed Varietal' and those you identified in Exercise 6 as having a significant effect at the 0.10 level. Click Run.

- b. Reducing the model.

Instructions: Examine the p-values in the Effect Summary table. Remove any terms with a non-significant p-values (at the 0.10 level).

- c. Display the equation for the model (numeric formula and visualizations). Can you provide a simple interpretation of the effect each variable has on Yield using this equation and/or the visualizations?

Instructions: Select Estimates > Show Predicted Expression under top red triangle. Select Factor Profiling > Profiler, and Factor Profiling > Surface Profiler under the top red triangle.

- d. Examine the model's fit and explanation of the variation in the data via the Actual vs Predicted plot, the RMSE, and R-Squared. Interpret these values and compare to the One-Factor ANOVA model you created in Exercise 3 (Model 1). How much additional variation in Yield is this multiple variable regression model able to describe?

Instructions: Select Row Diagnostics > Plot Actual by Predicted under top red triangle.

- g. Conduct a set of statistical tests comparing each pairwise difference of the three Seed Varietals. Compare the p-values to the results for the multiple comparison analysis using this model to Model 1? Have you been able to reach any new conclusions using a model that now accounts for the effect of the growing condition variables? Examine the Confidence Intervals for the means of each group as well as for each pairwise comparison. Explain why these intervals are narrower than those from Exercise 3.

Instructions: Select Model Comparisons under top red triangle. Ensure the variable chosen for the analysis is Seed Varietal. Select to Show Least Squares Means Plot and All Pairwise Comparisons – Student's t.

- h. Evaluate the assumptions of homogeneity of variance and normal distribution of the model's error term. Is there any cause for concern? Are there any unusual and/or highly influential data values?

Instructions: Under the top red triangle, select 'Plot Residuals by Predicted' and 'Plot Residual by Normal Quantile' in the Row Diagnostics sub menu.

- 8. **Model 3.** Create a multiple regression model starting off with all the main effects and two-factor interactions of Seed Varietal and all the growing conditions variables. The final model will be the result of performing Stepwise Backward Elimination using an automated tool in JMP that removes insignificant terms based upon a p-value criteria you'll specify.

- a. Launch Stepwise platform.

Instructions: Analyze > Fit Model. Choose the variable 'Yield' for the Y. Highlight 'Seed Varietal', 'Soil Quality', 'Fertilizer', '% Sun', 'Rainfall', and 'Irrigations' and select Factorial to Degree under the Macro drop down menu. This will specify a model will all the main effects and possible two-factor interactions. Under Personality, choose Stepwise. Click Run.

- b. Reduce the model.

Instructions: In the Stepwise report, select Enter All. All the possible model terms should be have a check mark. Choose p-value Threshold for the Stopping Rule. Set the Prob to Leave at 0.10. Select Backward for the Direction. Click the Step button. The two-factor interaction with the highest p-value will be removed first. Continue to click the Step button until it stops removing model terms. Note: You can also choose Go and it will automate the full reduction

- c. Create the final model.

Instructions: Select Make Model. You'll be brought to the Fit Model platform with only the selected variables in the Model Effect window. Choose Minimal Report under Emphasis drop down. Click Run.

- e. Display the equation for the model (numeric formula and visualizations). Can you provide a simple interpretation of the effect each variable has on Yield using the equation and/or visualizations?

Instructions: Select Estimates > Show Predicted Expression under top red triangle. Select Factor Profiling > Profiler, and Factor Profiling > Surface Profiler under the top red triangle.

- f. Examine the model's fit and explanation of the variation in the data via the Actual vs Predicted plot, RMSE, and R-Squared. Interpret these values and compare them to the model in Exercise 7 (Model 2). How much additional variation in Yield is this more complex model able to describe? Do you believe the additional variation this model is describing justifies the added complexity?

Instructions: Select Row Diagnostics > Plot Actual by Predicted under top red triangle.

- g. Examine the Confidence Intervals for the average Yield for each Seed Varietal and compare to those from Model 2. Are they much different? Conduct a set of statistical tests comparing each pairwise difference. Compare the p-values and Confidence Intervals for these comparisons to Model 2. Are they much different? Do you feel this model has had a significant impact on the ability to compare and estimate the Yield between the three Seed Varietals?

Instructions: Select Model Comparisons under top red triangle. Ensure the variable chosen for the analysis is Seed Varietal. Select to Show Least Squares Means Plot and All Pairwise Comparisons – Student's t.

- h. Evaluate the assumptions of homogeneity of variance and normal distribution of the model's error term. Is there any cause for concern? Are there any unusual and/or highly influential data values?

Instructions: Under the top red triangle, select 'Plot Residuals by Predicted' and 'Plot Residual by Normal Quantile' in the Row Diagnostics sub menu.

9. **Model 4.** Create a multiple regression model using the method of Forward Selection. This method starts with no terms in the model. It automates the process of testing all possible main effects and two-factor interactions and adds the one that is the most significant provided it satisfy a p-value criteria. With that term in the model, the process then searches to find another term that is most significant and adds that. The process continues until there are no more terms that can be added to the model based upon a p-value criteria.

- a. Launch Stepwise platform.

Instructions: Analyze > Fit Model. Choose the variable 'Yield' for the Y. Highlight 'Seed Varietal', 'Soil Quality', 'Fertilizer', '% Sun', 'Rainfall', and 'Irrigations' and select Factorial to Degree under the Macro drop down menu. Under Personality, choose Stepwise. Click Run.

- b. Build the model.

Instructions: Choose p-value Threshold for the Stopping Rule. Set the Prob to Enter at 0.10. Select Forward for the Direction. Click the Step button. The term with the lowest p-value (provided <0.10) will be added. Continue to click the Step button until it stops adding model terms. Note: You can also choose Go and it will automate the full forward selection process.

- c. Create the final model.

Instructions: Select Make Model. You'll be brought to the Fit Model platform with only the selected variables in the Model Effect window. Choose Minimal Report under Emphasis drop down. Click Run.

- h. Display the equation for the model (numeric formula and visualizations). Can you provide a simple interpretation of the effect each variable has on Yield using the equation and/or visualizations? Is this model simpler or more complex than Models 2 and 3?

Instructions: Select Estimates > Show Predicted Expression under top red triangle. Select Factor Profiling > Profiler, and Factor Profiling > Surface Profiler under the top red triangle.

- i. Examine the model's fit and explanation of the variation in the data via the Actual vs Predicted plot, RMSE, and R-Squared. Interpret these values and compare to Models 2 and 3. Are they much different?

Instructions: Select Row Diagnostics > Plot Actual by Predicted under top red triangle.

- j. Examine the Confidence Intervals for the average Yield for each Seed Varietal and compare to those from Models 2 and 3. Are they much different? Conduct a set of statistical tests comparing each pairwise difference. Compare the p-values and Confidence Intervals for these comparisons to those from Models 2 and 3. Are they much different? Do you feel this model has had a significant impact on the ability to compare and estimate the Yield between the three Seed Varietals?

Instructions: Select Model Comparisons under top red triangle. Ensure the variable chosen for the analysis is Seed Varietal. Select to Show Least Squares Means Plot and All Pairwise Comparisons – Student's t.

- i. Evaluate the assumptions of homogeneity of variance and normal distribution of the model's error term. Is there any cause for concern? Are there any unusual and/or highly influential data values?

Instructions: Under the top red triangle, select 'Plot Actual by Predicted', 'Plot Residuals by Predicted', and 'Plot Residual by Normal Quantile' in the Row Diagnostics sub menu.

10. Compare the performance and usefulness of the four models (i.e., how well they address the objectives of the analysis, the amount of variation in Yield they describe, their interpretability, and overall information learned).
11. Create a brief report of 2-3 Power Point slides summarizing your analysis results explicitly addressing the primary objectives of the analysis. Note: Include no more than three graphs, a few tables of key numerical results, and a few bullet points on each slide. Also include some recommendations for further studies and other data that could be collected.