Effect of Phenotypic Plant Growth Characteristic on Seed Yield in Soybeans

By Grace Gnronfoun, Kate Messitte, Jack Freimann, and Kayleen Kabba

Introduction

It has been estimated that soybean production contributes hundreds of billions of dollars to the global economy annually. The production and processing of this versatile crop has created thousands of jobs in the food and livestock sectors as well as the industrial product and biofuels industry. On a global scale, soybeans are vital in livestock feed, providing high levels of protein and nutrients to animal diets. Seed yield is a critical indicator of agricultural productivity, as it directly influences economic value of the crop, food supply, and the efficiency of land use in meeting global demand. There is also a vast number of soybean genotypes; some studies have estimated that upwards of 283 exist, with each genotype giving rise to different phenotypic qualities among plants of the same species. This is important because different phenotypic features may give certain soybean crops an advantage in ability to grow and thrive, affecting the amount of nutrients they contain and how many seeds are yielded during every growth cycle. Thus we raise the question: how do plant growth characteristics—such as plant height, number of pods, chlorophyll content, and biological weight– affect seed yield in soybean crops, and how do these relationships differ between the two genotypes C2S1G3 and C2S1G6?

Data and Exploratory Data Analysis

Data was accessed on Kaggle, a public platform hosting datasets, data science competitions and machine learning training, owned by Google. Our dataset was compiled as research for the College of Agriculture, University of Tikrit, Iraq, for facilitating diverse analytical and predictive modeling applications in precision agriculture, yield prediction, and crop health assessment.

Variables relevant to the research question are, plant height, number of pods, biological weight, and chlorophyll content; these four variables are quantitative predictors, they help inform our outcome variable, seed yield. Due to a sizable amount of categorical predictor variables (6), we decided to limit our research to just two, that differ only in terms of base genotype of the soybean, in terms of salicylic acid levels and water stress level they are the same. Potential confounders include regions of growth (sunlight, rainfall), environmental pollution, interference by animals.

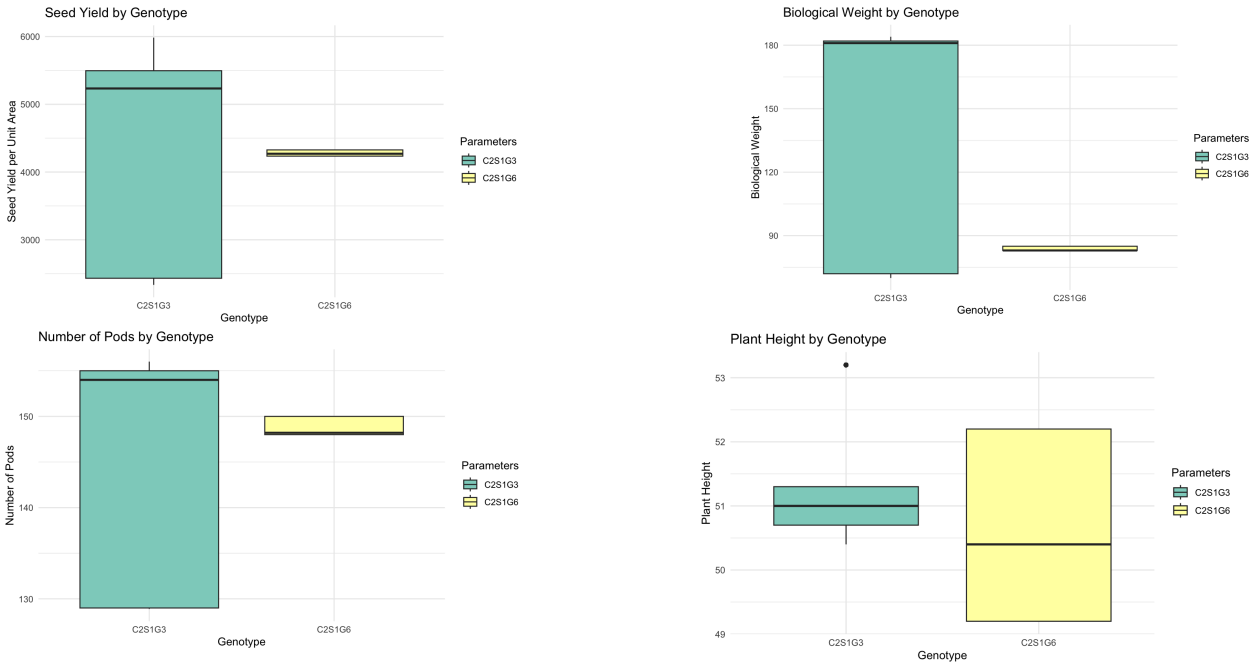
Genotype C2S1G3

Variable	Mean	Standard Deviation
Plant Height (PH)	51.30013	0.9024174
Number of Pods (NP)	142.1544	12.87026
Biological Weight (BW)	127.0188	55.35637

ChlorophyllA663	2.716943	1.162629
Protein Percentage (PPE)	34.53207	2.763506
Number of Seeds per Pod (NSP)	1.896728	0.1416881
Seed Yield per Unit Area (SYUA)	4008.396	1580.273

Genotype C2S1G6

Variable	**Mean**	**Standard Deviation**
Plant Height (PH)	50.60091	1.233167
Number of Pods (NP)	148.7338	0.8997722
Biological Weight (BW)	83.6671	0.943268
ChlorophyllA663	1.566775	0.1700246
Protein Percentage (PPE)	33.66768	1.307484
Number of Seeds per Pod (NSP)	2.146632	0.07546144
Seed Yield per Unit Area (SYUA)	4277.028	37.99246



Figures 1 - 4, Box plot charts comparing genotypes by quantitative variables (From top left to bottom right: Seed Yield, Biological weight, Number of Pods, Plant height)

Our linear regression model:

E[Seed Yield | Genotype, Plant Height, Number of Pods, Biological Weight, Chlorophyll Content] =
β0 + β1(Genotype) + β2(Plant Height) + β3(Number of Pods) + β4(Biological Weight) + β5(Chlorophyll Content)

The genotype of the soybean plant is a confounding variable because it influences all of the other variables and can confuse the relationship between other traits and seed yield if not correctly accounted for. The coefficients of plant height and number of pods are mediating variables because they are not directly affecting seed yield but they help explain how genotypes end up affecting our end result. Biological weight, chlorophyll content, and protein percentage are precision variables because they help to predict seed yield more accurately, without messing with the relationship between other variables.

There was no other real consideration to use another model because the linear regression model was optimal in addressing our research question. It allowed for us to best assess how our variables affected our outcome.

Our ultimate alternative hypothesis that we are testing is:
$B1 \neq 0$
→ Genotype does have an effect on seed yield, even when we control for other variables.

Our null hypothesis that we are testing is:
$B1 = 0$
→ Genotype does not have an effect on seed yield, when we control for other variables.

Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2287.169798	218.0891978	-10.487314	0.0000000
Plant Height (PH)	-96.023713	1.5001581	-64.009064	0.0000000
Number of Pods (NP)	68.090164	1.9463267	34.983934	0.0000000
Biological Weight (BW)	1.191516	0.4426377	2.691854	0.0071312
ChlorophyllA663	511.948039	5.0725616	100.924953	0.0000000
ParametersC2S1G6	393.977025	33.2568075	11.846508	0.0000000

After cleaning up our data to focus on the specific two genotypes, our adjusted r-squared value was 0.995. This indicates that about 99.5% of our variance in data is explained

by our model. The intercept of roughly -2287.17 represents the predicted seed yield when all other variables are 0, which is not meaningful in this context because it is not possible to yield a negative number of seeds. Plant height has a negative coefficient of roughly 96, which suggests that for every additional unit of plant height, seed yield decreases by about 96 units, holding all else constant. The number of pods has a strong positive effect on seed yield of about 68 units, revealing that more pods leads to higher seed yield, which is biologically understandable. Biological weight also has a positive effect on seed yield of 1.19 units, suggesting that heavier plants produce slightly more seeds. The chlorophyll coefficient has a strong positive correlation with seed yield, implying that plants with higher chlorophyll levels are more productive. Finally, the C2S1G6 parameter highlights that an additional C2S1G6 genotype in plants is associated with an increase in about 394 units of seed yield.

This linear regression model shows that seed yield per unit area is strongly predicted by plant growth traits and genotype. Higher chlorophyll content and number of pods are associated with significantly higher yields, while greater plant height is linked to lower yields. Genotype also plays a major role: plants with the C2S1G6 genotype produce, on average, 394 more units of seed yield than C2S1G3 plants, even after adjusting for plant traits. The model explains 99.53% of the variation in seed yield and is highly statistically significant.

All added variables—Number of Pods, Biological Weight, Chlorophyll Content, and Protein Percentage—significantly improve model fit for predicting seed yield. However, Biological Weight and Number of Pods provide the greatest improvements, suggesting that structural and reproductive growth traits are more critical drivers of yield than physiological or nutritional metrics. These findings validate including all five predictors in your final model.

With each addition of variable–Number of Pods, Biological Weight, Chlorophyll, and Genotype–the base model significantly improved the prediction of seed yield per unit area, all with a p-value of $p < 2.2e-16$.

Model Improvement Summary

Step	Model_Comparison	Variable_Added	F_statistic	p_value	Interpretation
1	SYUA ~ PH → SYUA ~ PH + NP	Number of Pods (NP)	3836.90	< 2.2e-16	Model Improved
2	SYUA ~ PH + NP → SYUA ~ PH + NP + BW	Biological Weight (BW)	26357.00	< 2.2e-16	Model Improved
3	SYUA ~ PH + NP + BW → SYUA ~ PH + NP + BW + ChlorophyllA663	ChlorophyllA663	4280.60	< 2.2e-16	Model Improved
4	SYUA ~ PH + NP + BW + ChlorophyllA663 → SYUA ~ PH + NP + BW + ChlorophyllA663 + Genotype	Genotype	398.89	< 2.2e-16	Model Improved

Conclusion

Through linear regression, a statistically significant relationship was found between plant height, number of pods per plant, biological weight, chlorophyll content, and seed yield of the C2S1G6 and C2S1G3 soybean genotypes. On average, we would expect that an increase in plant height typically leads to a decrease in seed yield and an increase in number of pods per plant, biological weight, and chlorophyll content leads to an increase in overall seed yield. We are also able to see through an F-test that there is statistical significance between genotype and seed yield with the C2S1G6 genotype typically having a higher seed yield than the C2S1G3 genotype, even after all other factors have been adjusted for. Given these results, we would advise soybean farmers and those in the biofuels industry that the C2S1G6 soybean genotype may produce a higher seed yield which would be economically beneficial. Additionally, soybean plants that are phenotypically shorter and have more bean pods with a genotype that produces a higher chlorophyll content and higher biological weight, may have a higher seed yield, which leads to a more sustainable soybean production both economically and nutritionally. In sampling soybeans, there is a potential sampling bias with the region that the sample was taken. The soybean data used in this analysis was taken in Iraq, which could have vastly different climate and environmental factors from other countries and regions where soybean production is high. It is important to keep this in mind because the relationship between plant height, chlorophyll content, number of pods, and seed yield, may be impacted depending on the climate and environmental factors of where the soybean is grown, thus we are unable to make any definitive causal claims for the entire soybean population. However, this analysis may still give some insight into how certain genotypic and phenotypic features may impact the overall seed yield of soybeans which has the broader implication of making the soybean production, livestock, and biofuel industry more sustainable.

References

“Advanced Soybean Agricultural Dataset (2025)”

- https://www.kaggle.com/datasets/wisam1985/advanced-soybean-agricultural-dataset-2025?resource=download

“The Economic Impact of U.S. Soybeans and End Products on the U.S. Economy - 2023 update”

- https://www.nopa.org/wp-content/uploads/2023/08/0LMC_SoyEconStudy_Aug2023.pdf

“Seeding Rates in Relation to Maximum Yield and Seed Costs”

- https://crops.extension.iastate.edu/encyclopedia/seeding-rates-relation-maximum-yield-and-seed-costs

“Seed yield, seed quality, profitability, and risk analysis among double crop soybean maturity groups and seeding rates”

- https://acsess.onlinelibrary.wiley.com/doi/10.1002/agj2.20626