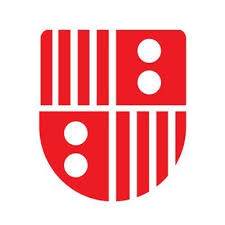
Andreas Frey, Andres Sanchez, Eduardo Fernandez, Emma Lotz, Enrico Fausti, Jose Conde, Veronica Pereira, Yuyan Lai



Jayalaxmi analysis

Team 2

**Executive summary**

# First, we will discuss the regression model we built to predict most precisely the variation in income per acre as dependent upon the variables available in “jayalaximi.xls”. Using our final model, which we finetuned by eliminating variables with a p-value greater than 0.05, we are then able to see through our regression model how the variables may impact the income per acre. In this we may also observe that it is likely there are factors which we do not have the data for, as our model only explains 34.61% of the variability in income per acre. Then, considering which of the variables that when removed had the largest impact on R-squared, we see the variable with the largest impact on variability to be Chawki\_bivol.

Using this information as a foundation, we will then consider whether an investment of 15,000 Rs per acre will be beneficial for increasing profit. We will show the benefits of this investment and take into consideration time and how much temperature management affects income in comparison to other variables.

Furthermore, we will conclude by discussing the effectiveness of training programs and consider whether they have a substantial impact on income. Considering training with the other variables we will show that it doesn’t seem to have a positive impact although this is minimized with the inclusion of more variables and their individual impacts on income per acre.

***Explain what you have done and why in a brief and clear way, reporting the regression equation and a measure of the goodness of fit***

Initially, we ran a regression model considering all possible variables. As it can be seen in **Exhibit 2** this model would explain 33.97%[[1]](#footnote-1) of the variability of income per acre, but there were a lot of values that were very likely to add no information to the model (very high p values) and that if put there might add some noise to the model.

Using a simple elimination process[[2]](#footnote-2) of trial and error of variables (using low correlations with income per acre and high p-values as our basis of eliminating variables) a model was built having all variables being relevant (p values smaller than 0.05).

As it can be seen in **Exhibit 2b** the variables that were used for our model were loan amount, crop insured, training on sericulture, own vermi compost, bio fertilizers, mechanization, mulberry diseases, affected by pest, rearing cost, temperature management and chawki bivol. This model manages to explain 34.61% of the variability of income per acre while giving more valuable variables which have a smaller risk of being affected by collinearity.

***How much variation in income\_per\_acre are you able to explain with the variables included in your model?***

Our model returns a parameter of adjusted R-squared equal to 0.3461, implying that approximately 34.6% of the variation in income per acre can be explained with the variables used in our model, using the best-fit regression line. These variables, as mentioned above, are loan amount, rearing cost, insurance of the crop, training on sericulture, vermi compost, biofertilizers, mechanization, mulberry disease, affected by pest, rearing cost, temperature management of the rearing house and chawki bivol.

**Exhibit 4** visually demonstrates how this R-squared value represents the difference between the actual and predicted dependent variable, income\_per\_acre, in this case. As we can see below, a certain number of data points, represented by dots, scatter around the identified best-fit regression line.

Such diversion from the linear regression line is not unusual in practice, as a regression model is obtained by merely observing the data set and variables chosen – the variables given don’t explain everything that might affect income per acre. More specifically, there might be other factors that are not addressed by the collected data influencing the actual dependant variables, in return affecting the R-squared parameter. The limitation on number of samples observed can also be an influencing element.

In this case, other relevant factors to the farmer’s income per acre may include but are not limited to:

* District where the farmers are located
* Soil fertility
* Number of silkworms reared
* Other outstanding costs, such as transportation

***According to the sample used and your model, what is the most important single factor explaining income\_per\_acre***

In order to determine which is the most significant variable affecting the income generated per acre, we must first define certain criteria to measure the significance of each.

The first approach would be to observe and compare the p-values obtained in our regression model. However, although the p-values do indicate the probability of a variable having a weight of zero, they do not indicate the importance of the variable in the model, meaning that this option is not suitable  
The second approach would be to use the regression coefficients obtained from our model. These coefficients show the average change of the dependant variable with respect to the independent variable.

If we want to use this approach, we must consider the values that each variable can take. However, we cannot make use of this approach given the fact that in our model we have made use of dummy variables and variables with different units, making it impossible to compare one to another.

Our final decision was to find the variable that, when added or taken out of the regression model, would have the most drastic change in R-squared. When an independent variable is added to a regression model, the change in the R-squared value explains the upgrade or downgrade in the model in question, in this case income per acre.

Therefore, using the option mentioned above we decided that the variable that increased R-squared the most when added to the model was “Chawki\_bivol”. This means that the single variable that explained the most change in the regression model of income generated per acre was whether the farm used only bivoltine hybrid or not.

***A farmer is considering a single investment of 15,000 Rs per acre in an effective temperature management system in rearing house. This farmer is not planning to make any other investments. Advise this farmer on this particular investment.***

Looking at our model, if a farmer invests 15,000 Rs per acre (or any investment as this variable is binary) in an effective temperature management system for the rearing house, and does not make any other investment in anything else, its income per acre per year, on average, will be 11,940 Rs (**Exhibit 2b**) higher than before (without a temperature management system), ceteris paribus.

However, if we take a deeper look at the data, we can find very interesting information. If we only take the data of the farmers that don’t have an effective temperature management system and we compare the mean of mechanization of this data set to the original data set, we would find out that only 14% of the farmers that don’t have temperature management systems have mechanized their processes (**Exhibit 5)**.As mechanization is a better overall investment, according to our model, the farmer should consider focusing his investment on mechanization if he hasn’t yet (which is likely, according to the data) even though a temperature management system would remain a good investment as well.

We would advise the farmer to invest in the temperature management system, only if its useful life is higher than 2 years. This is the case as with an initial investment of 15,000 Rs and an increase in annual income of 11,940 Rs thanks to the new system, it will take 1.25 years to recover the investment (15000/11940).

***How effective the training programs are? Does receiving a formal training on sericulture increase the income per acre?***

At first sight, the training programs look to be ineffective as the Beta of the variable training on sericulture was negative (meaning it has a negative impact on income\_per\_acre). As this was suspicious, we looked at the correlation table (**Exhibit 3**) and observed that training negatively correlated with variables, such as mechanization, that had a relatively strong positive impact on income per acre and positively correlated with variables, like chawki\_bivol, that had a relatively strong negative impact on income per acre.

To see if this was the cause of these results, we created simple linear models with the variables income per acre, chawki\_bivol, mechanization, years of experience in sericulture and training on sericulture. In all these tests the Beta for training on sericulture remained negative.

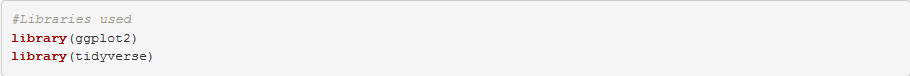
From this newly obtained data it can be deduced that training on sericulture does not significantly affect the income per acre; in fact, it makes it smaller. As the model only explains 34.6% of the total variability of income per acre it can’t be fully confirmed that training is a waste of money, even if the results strongly indicate it. Our model indicates that the training programs are ineffective and that they reduce income per acre on sericulture, however, its negative weight in the model reduces with each added variable.

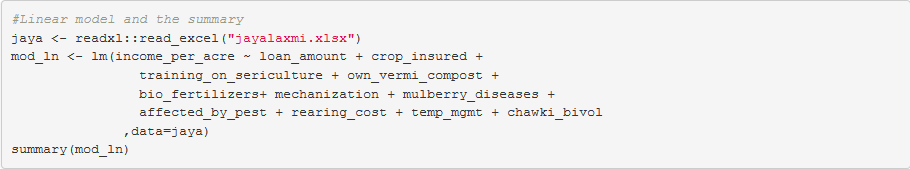
One explanation for this negative weight could be related to this line mentioned in the text:

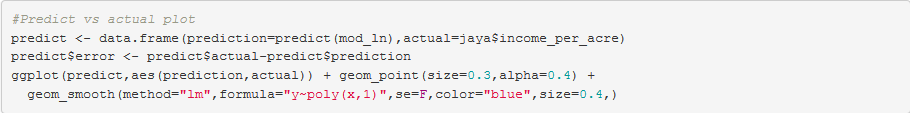
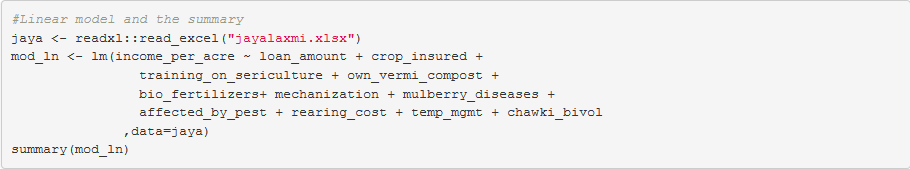
“The

**Exhibit 1**

The Code

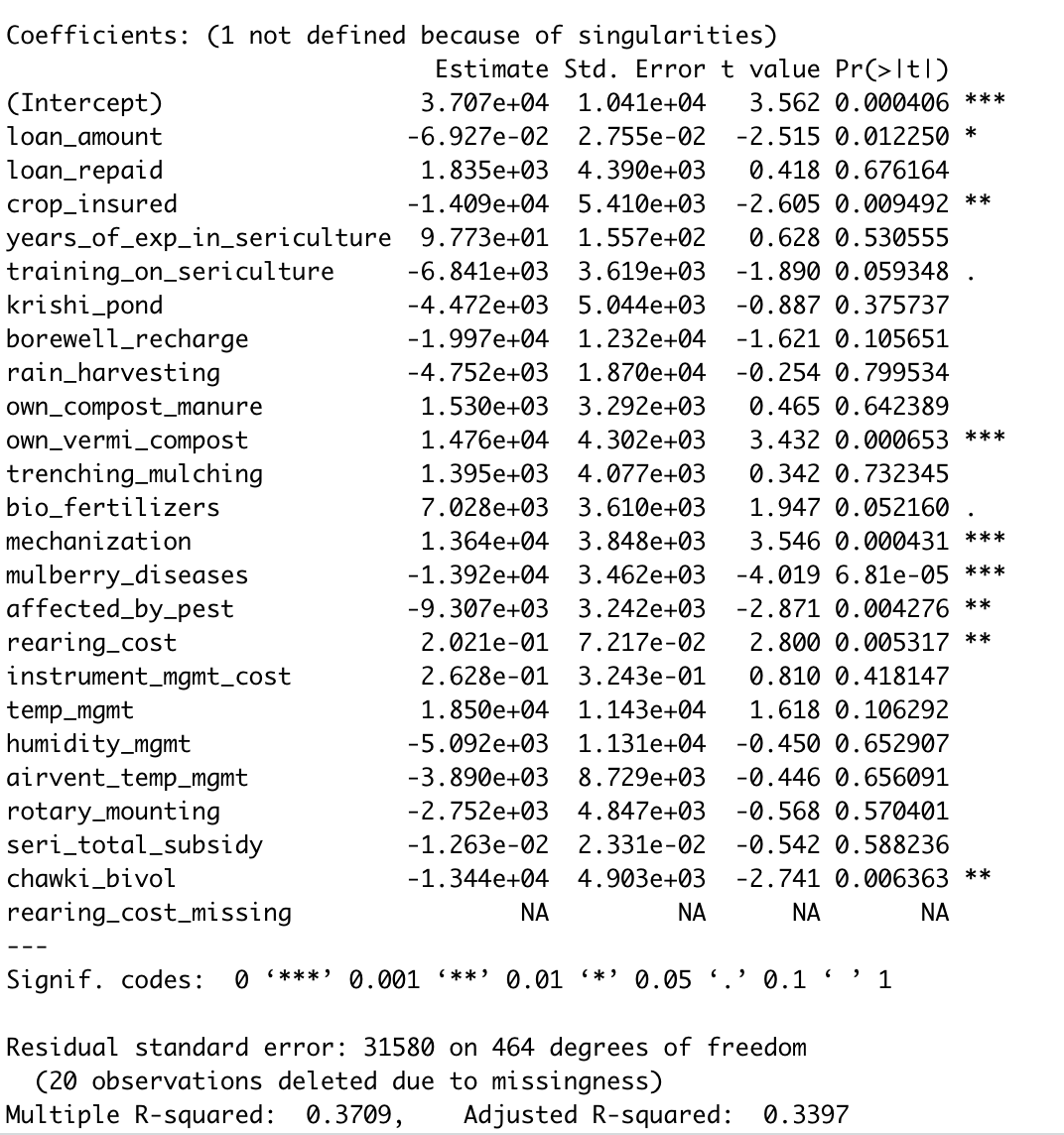






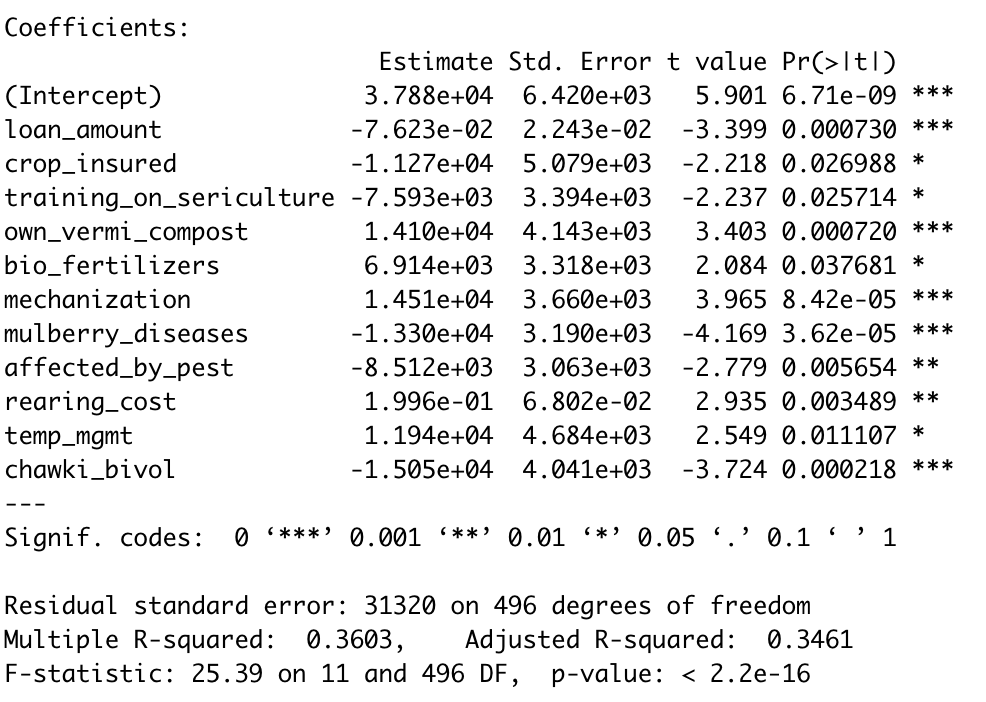
**Exhibit 2a**

Coefficients



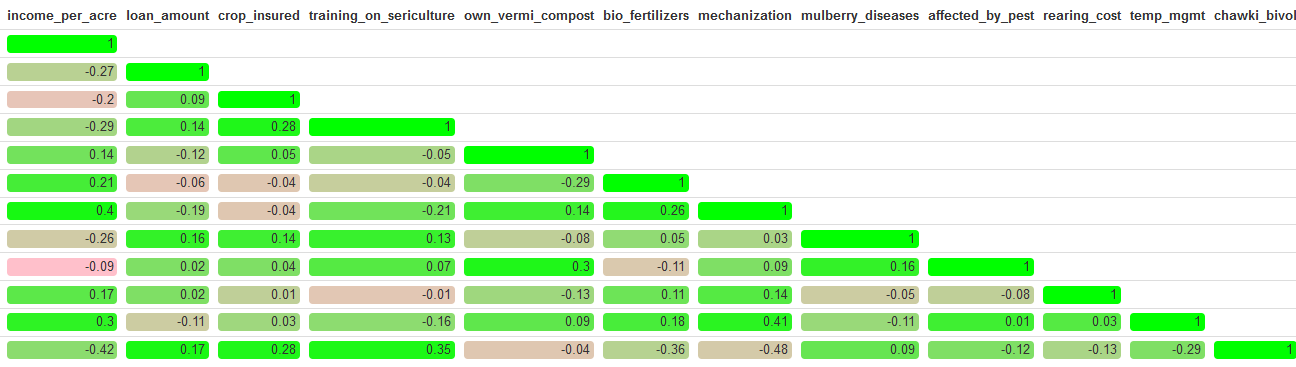
**Exhibit 2b**

Coefficients continued



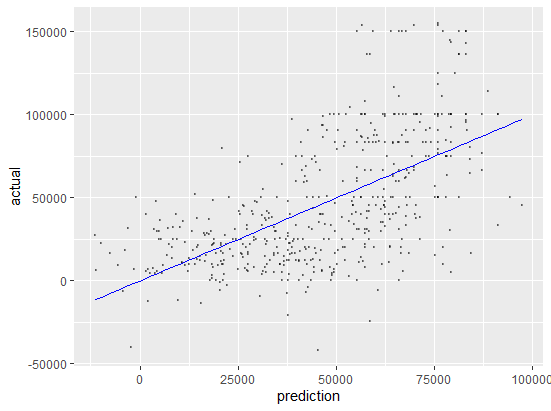
**Exhibit 3**

Correlation Table



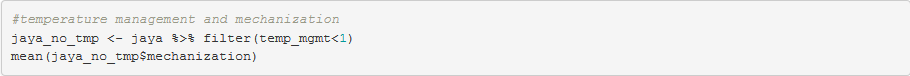
**Exhibit 4**

Regression Line



**Exhibit 5**

Temperature Management



Insertando imagen...





1. We will use adjusted R as our percentage of variation explain as simple R squared might increase by chance just adding new variables [↑](#footnote-ref-1)
2. We also used principal component regression and the correlation table to identify the most important variables but decided to focus only on the elimination approach in the report [↑](#footnote-ref-2)