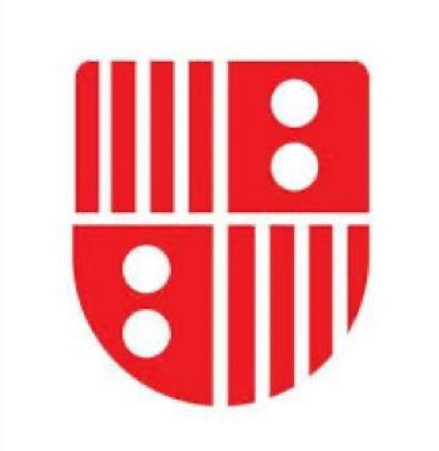

JAYALAXMI ANALYSIS

Team 2

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

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] 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] 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.

[1] We will use adjusted R as our percentage of variation explain as simple R squared might increase by chance just adding new variables

[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

- ***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 34.61% 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 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.

On the other side, it would be interesting to consider that sericulture allows for practice in low hand landing while maintaining its green cover and mitigating land erosion. It is a technique that has been used to uplift the rural economy in India, so the training and introduction of this technique allows farmers to work in lands where it was not possible before. For lands that do not require that type of training; however, based on the regression, it would not be a needed as it would negatively impact income. Knowing this, it is hard to estimate the effectiveness of training or the value that it generates in the long run by avoiding soil erosion.

Exhibit 1

The Code

```
#Libraries used
library(ggplot2)
library(tidyverse)
```

```
#Linear model and the summary
jaya <- readxl::read_excel("jayalaxmi.xlsx")
mod_ln <- lm(income_per_acre ~ loan_amount + crop_insured +
               training_on_sericulture + own_vermi_compost +
               bio_fertilizers+ mechanization + mulberry_diseases +
               affected_by_pest + rearing_cost + temp_mgmt + chawki_bivol
             ,data=jaya)
summary(mod_ln)
```

```
#Predict vs actual plot
predict <- data.frame(prediction=predict(mod_ln),actual=jaya$income_per_acre)
predict$error <- predict$actual-predict$prediction
ggplot(predict,aes(prediction,actual)) + geom_point(size=0.3,alpha=0.4) +
  geom_smooth(method="lm",formula="y~poly(x,1)",se=F,color="blue",size=0.4,)
```

Exhibit 2a

Coefficients

```
Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.707e+04  1.041e+04   3.562 0.000406 ***
loan_amount                -6.927e-02  2.755e-02  -2.515 0.012250 *
loan_repaid                 1.835e+03  4.390e+03   0.418 0.676164
crop_insured               -1.409e+04  5.410e+03  -2.605 0.009492 **
years_of_exp_in_sericulture 9.773e+01  1.557e+02   0.628 0.530555
training_on_sericulture    -6.841e+03  3.619e+03  -1.890 0.059348 .
krishi_pond                -4.472e+03  5.044e+03  -0.887 0.375737
borewell_recharge          -1.997e+04  1.232e+04  -1.621 0.105651
rain_harvesting            -4.752e+03  1.870e+04  -0.254 0.799534
own_compost_manure          1.530e+03  3.292e+03   0.465 0.642389
own_vermi_compost           1.476e+04  4.302e+03   3.432 0.000653 ***
trenching_mulching          1.395e+03  4.077e+03   0.342 0.732345
bio_fertilizers             7.028e+03  3.610e+03   1.947 0.052160 .
mechanization               1.364e+04  3.848e+03   3.546 0.000431 ***
mulberry_diseases          -1.392e+04  3.462e+03  -4.019 6.81e-05 ***
affected_by_pest           -9.307e+03  3.242e+03  -2.871 0.004276 **
rearing_cost                2.021e-01  7.217e-02   2.800 0.005317 **
instrument_mgmt_cost        2.628e-01  3.243e-01   0.810 0.418147
temp_mgmt                   1.850e+04  1.143e+04   1.618 0.106292
humidity_mgmt              -5.092e+03  1.131e+04  -0.450 0.652907
airvent_temp_mgmt          -3.890e+03  8.729e+03  -0.446 0.656091
rotary_mounting            -2.752e+03  4.847e+03  -0.568 0.570401
seri_total_subsidy         -1.263e-02  2.331e-02  -0.542 0.588236
chawki_bivol               -1.344e+04  4.903e+03  -2.741 0.006363 **
rearing_cost_missing               NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31580 on 464 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.3709,    Adjusted R-squared:  0.3397
```

Exhibit 2b

Coefficients continued

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.788e+04  6.420e+03   5.901 6.71e-09 ***
loan_amount               -7.623e-02  2.243e-02  -3.399 0.000730 ***
crop_insured              -1.127e+04  5.079e+03  -2.218 0.026988 *
training_on_sericulture   -7.593e+03  3.394e+03  -2.237 0.025714 *
own_vermi_compost          1.410e+04  4.143e+03   3.403 0.000720 ***
bio_fertilizers            6.914e+03  3.318e+03   2.084 0.037681 *
mechanization              1.451e+04  3.660e+03   3.965 8.42e-05 ***
mulberry_diseases         -1.330e+04  3.190e+03  -4.169 3.62e-05 ***
affected_by_pest          -8.512e+03  3.063e+03  -2.779 0.005654 **
rearing_cost               1.996e-01  6.802e-02   2.935 0.003489 **
temp_mgmt                  1.194e+04  4.684e+03   2.549 0.011107 *
chawki_bivol              -1.505e+04  4.041e+03  -3.724 0.000218 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31320 on 496 degrees of freedom
Multiple R-squared:  0.3603,    Adjusted R-squared:  0.3461
F-statistic: 25.39 on 11 and 496 DF,  p-value: < 2.2e-16
```

Exhibit 3

Correlation Table

income_per_acre	loan_amount	crop_insured	training_on_sericulture	own_vermi_compost	bio_fertilizers	mechanization	mulberry_diseases	affected_by_pest	rearing_cost	temp_mgmt	chawki_bivol
1											
-0.27	1										
-0.2	0.09	1									
-0.29	0.14	0.28	1								
0.14	-0.12	0.05	-0.05	1							
0.21	-0.06	-0.04	-0.04	-0.29	1						
0.4	-0.19	-0.04	-0.21	0.14	0.26	1					
-0.26	0.16	0.14	0.13	-0.08	0.05	0.03	1				
-0.09	0.02	0.04	0.07	0.3	-0.11	0.09	0.16	1			
0.17	0.02	0.01	-0.01	-0.13	0.11	0.14	-0.05	-0.08	1		
0.3	-0.11	0.03	-0.16	0.09	0.18	0.41	-0.11	0.01	0.03	1	
-0.42	0.17	0.28	0.35	-0.04	-0.36	-0.48	0.09	-0.12	-0.13	-0.29	1

Exhibit 4

Regression Line

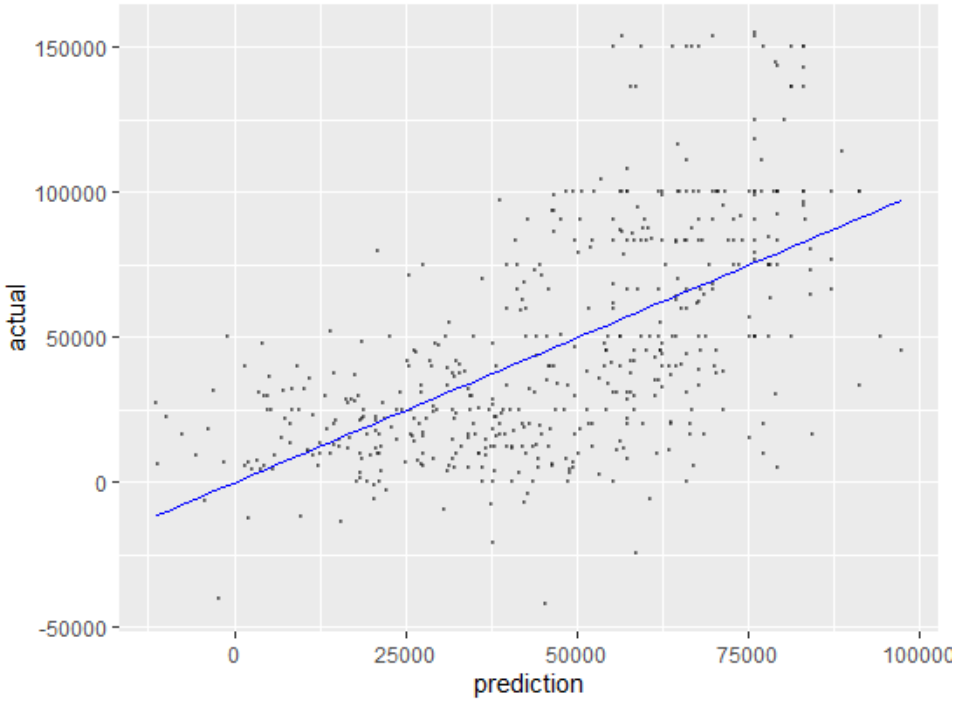

Exhibit 5

Temperature Management

```
#temperature management and mechanization
jaya_no_tmp <- jaya %>% filter(temp_mgmt<1)
mean(jaya_no_tmp$mechanization)
```

```
## [1] 0.1384615
```

```
mean(jaya$mechanization)
```

```
## [1] 0.6515748
```