Regression Analysis of Sugarcane Yields & Potential Interventions

Kara Jia 3031907464

10/18/2021

1 Introduction

Agricultural yields are influenced by many factors which include environmental conditions and farming practices. To maintain a sustainable business, it is important to plan, understand, and adjust these practices over time. This analysis focuses on using linear models to predict sugarcane yields of a sugar company, specifically sugar production and sugar content. We evaluate the models' performance and use them to determine possible interventions to improve yields.

2 Data Description

The data contains information for 3,775 paddocks in different regions around northern Australia in 1997. The specific variables in the dataset are described below. `Region` and `Position` are categorical variables, and `HarvestMonth` is ordinal. `Area`, `Age`, `HarvestDuration`, `Tonn.Hect`, `Sugar`, and `Rainfall.96` are continuous. In this analysis, the goal is to build two models, one to predict sugarcane production `Tonn.Hect` and the other to predict sugar content `Sugar`. All other variables are potential explanatory variables, which will be explored in the next section.

1. **Region**: Region in which each paddock is located (defined by physical position and average rainfall)
2. **Position**: Geographic position of each paddock in the general area according to the compass directions (E = east, W = west, N = north, S = south, C = central)
3. **Area**: Size of the paddock in hectares
4. **Age**: Years elapsed since the paddock was plowed out and planted with new sugarcane seeds
5. **HarvestMonth**: Month of the year in which the harvest took place, enumerated starting from January
6. **HarvestDuration**: Time taken to harvest the sugarcane in days
7. **Tonn.Hect**: Tons per hectare of sugarcane produced by this paddock
8. **Sugar**: Commercial sugar content produced by this paddock (in tons per batch processed)
9. **Rainfall.96**: Total rainfall for the district from July 1996 through December 1996 (millimeters)

3 Explanatory Data Analysis

To gain a better understanding of the variables and data, explanatory data analysis (EDA) will be conducted. First, distributions of the two predictor variables are plotted in Figure 1. The distribution of sugarcane produced is right skewed, while that of sugarcane content is approximately normal. It is desirable to symmetrize data, not only to make it easier to model, but also because normality is an assumption of statistical inference. Thus, a log transformation is applied to sugarcane production, which results in an approximately normal distribution that is used as the predictor.

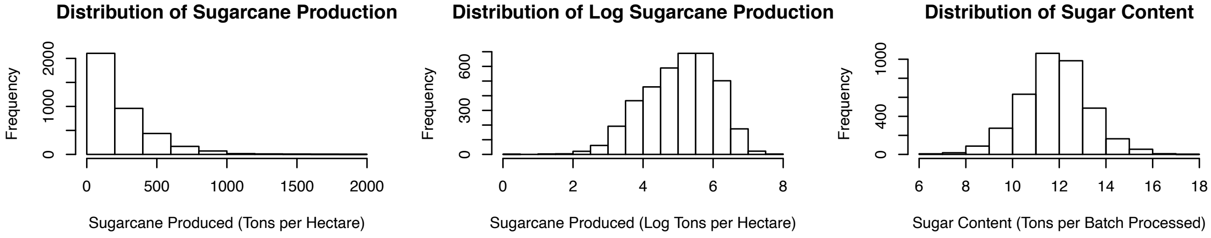

Figure 1: Distribution of predictor variables

The univariate distributions of numerical explanatory variables are visualized in Figure 2. `Area` is skewed to the right, so a log transform is applied to symmetrize it. `HarvestDuration` is much more heavily skewed and bimodal with a small peak around 25 days, which makes the transformation less helpful in spreading out the values. While `Rainfall.96` is continuous, a closer look reveals that there are 8 unique values, so it can be treated as discrete. There appears to be no outliers or unusual values.

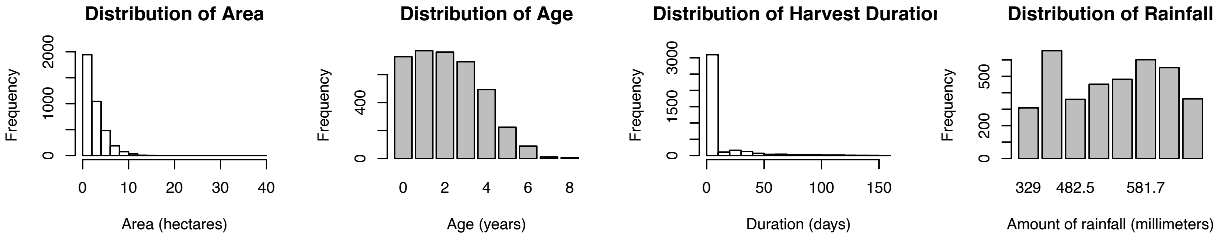

Figure 2: Univariate distributions of explanatory variables

In Figure 3, the correlations between each pair of numerical variables show that most explanatory variables are uncorrelated or weakly correlated with each other. The highest correlation of 0.877 is between explanatory variable `LogArea` and predictor `LogTonn.Hect`. The correlations suggest that `LogArea` and `HarvestDuration` could be useful for predicting sugar production, and `Age` could be useful for predicting sugar content. However, since there is some correlation between `LogArea` and `HarvestDuration`, it is important to check for possible collinearity.

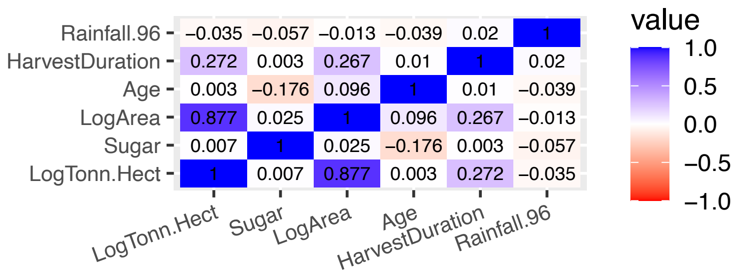

Figure 3: Correlation matrix

4 Sugar Production

Plotting the relationship of explanatory variables to the predictor `LogTonn.Hect` in Figure 4, the positive linear relationship with `LogArea` visualizes the strong correlation. The boxplots for categorical and discrete numerical variables give some information about which terms to include in the model. From Figure 5, it is observed that the median log sugar production does not vary much between `Region`, but does for `Position` where E has the highest and S has the lowest median. Similarly, there is variation across `Age`.

In addition to the main effect terms, interaction terms are visualized using boxplots and coplots. The boxplot in Figure 6 shows the interaction between two variables `Age` and `Position`. It can be noticed that the difference in medians for position C and E changes across age, which motivates using an interaction term. Figure 7 plots the interaction of `Position` (categorical) and `HarvestDuration` (continuous). Due to the tail of `HarvestDuration` at high values, the scatter plots seem to differ by `Position`, but it is unclear whether it is important to include it in the linear model. This will be further investigated during model selection.

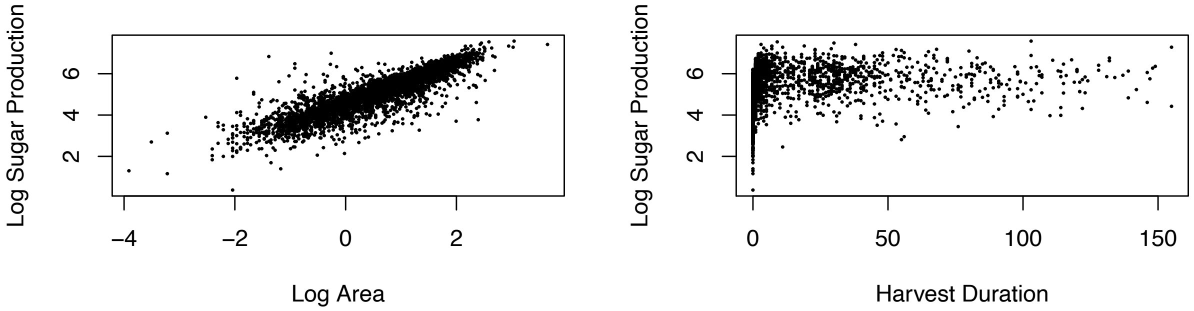

Figure 4: Scatter plots of numerical explanatory variables against log sugar production

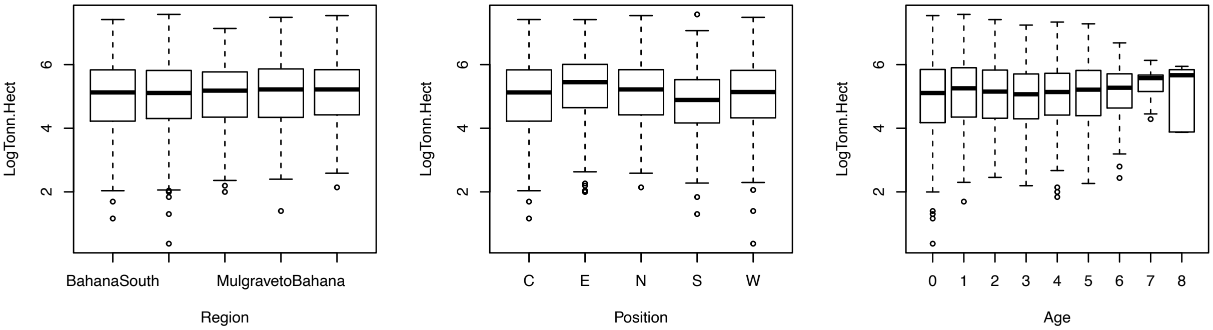

Figure 5: Boxplots of categorical or discrete variables against log sugar production

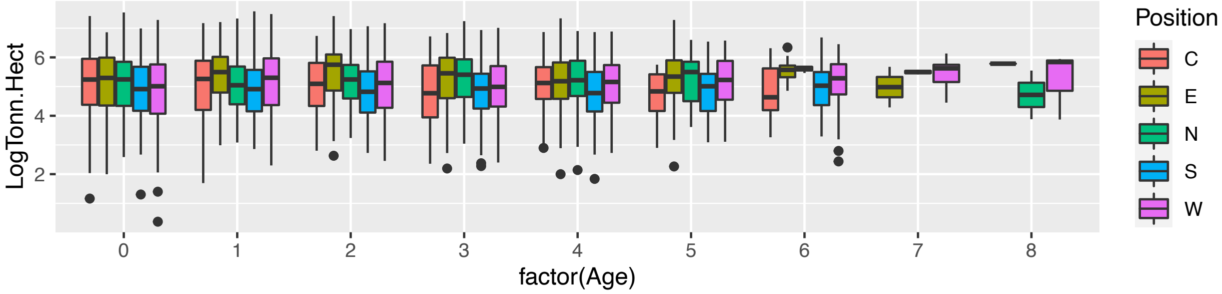

Figure 6: Interaction boxplot for log sugar production, grouped by Age and Position

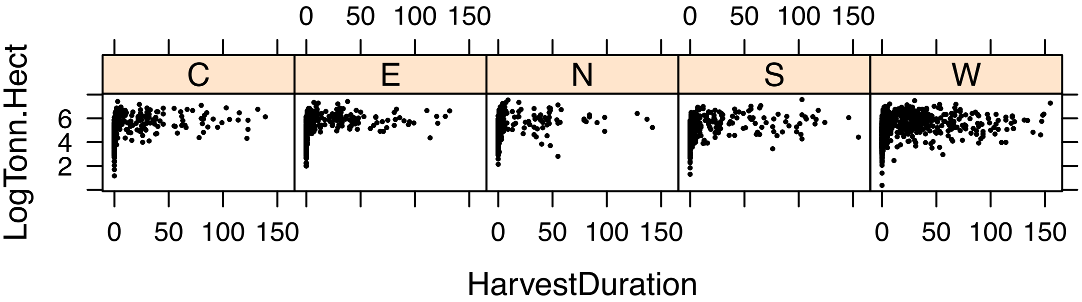

Figure 7: Interaction plot of HarvestDuration for log sugar production, grouped by Position

4.1 Model Selection

From EDA, several variables and interactions were identified as potentially useful for predicting sugar production. The full model is: $log(Tonn.Hect) \sim log(Area) + HarvestDuration + Position + Age + Position * Age + Position * HarvestDuration$. 80% of the data is randomly selected to fit the model. The remaining 20% is set aside to use later for evaluating how well the model predicts on unseen data.

The R^2 values and F-test results will be used to select the best model. The full linear model gives a decent R^2 value of 0.7789, with only a slight reduction in the adjusted R^2 of 0.7778. Adjusted R^2 discourages using too many explanatory variables. From the `lm` output in the Appendix, the F-test has very low p-value < 2.2e-16, so we can reject the null that all coefficients are 0. The coefficients with p-values less than 0.05 are `LogArea`, `Age`, and `PositionW:Age`, for which we reject the null that each of those coefficients is 0.

Since an interaction term involving `Position` has a statisitically significant p-value, by the principle of marginality, the main effect `Position` must be kept in the model although there is no evidence any of its coefficients are nonzero. None of the coefficients for `HarvestDuration` and its interaction terms have low p-value. We can further test whether these variables should be kept using `Anova`.

The `Anova` table below shows the p-values from conducting incremental F-tests on each variable independently, where the null hypothesis is that the coefficient is zero. All coefficients are significant, except for interaction `HarvestDuration:Position`. Thus, we can reject the null that the coefficient of `HarvestDuration` is 0, determined by comparing the fullest possible model with and without `HarvestDuration` while obeying the principle of marginality. We fail to reject the null that the coefficient of `HarvestDuration:Position` is 0, so the interaction term can be excluded from the model. The new model has R^2 of 0.7787 and adjusted R^2 of 0.7779, both similar to that of the full model.

```
## Anova Table (Type II tests)
##
## Response: LogTonn.Hect
##                           Sum Sq   Df   F value    Pr(>F)
## LogArea                  2078.89    1 9392.1139 < 2.2e-16 ***
## HarvestDuration             5.38    1   24.3031 8.678e-07 ***
## Position                   12.96    4   14.6360 7.631e-12 ***
## Age                        19.97    1   90.2365 < 2.2e-16 ***
## Position:Age                5.50    4    6.2118 5.639e-05 ***
## HarvestDuration:Position    0.51    4    0.5772    0.6792
## Residuals                 664.92 3004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To evaluate the model, the residual plot is produced in Figure 8, which ideally is homoscedastic. We observe that the points are approximately clustered around 0. However, the residuals tend to be positive for smaller fitted values and somewhat negative for larger fitted values, so there is some heteroscedasticity. Though Figure 1 shows the transformed predictor is approximately normal, it is a little left-skewed, so it may be necessary to reevaluate assumptions or to consider more complex models.

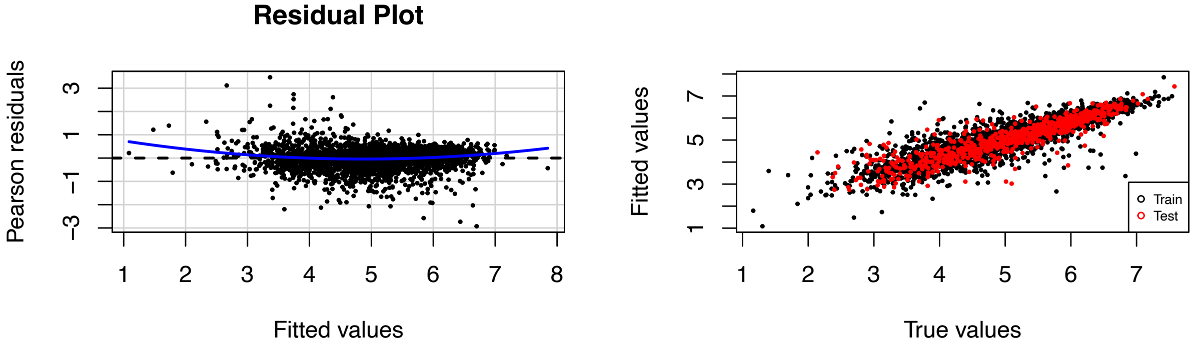

Figure 8: Residual plot and fitted values for log sugar production

4.2 Model Interpretation

Interpretation of the model will be examined both overall and for each variable. The model overall performs decently well, and the explanatory variables are able to explain about 77.87% of the variation in log sugar production, based on R^2. We can also interpret how each variable is associated with the predictor by looking at the partial coefficients from lm, which isolates the relationship of variables. Since different variables are on different scales, we will first standardize them, so the coefficients are comparable. This is equivalent to coefficients from the model fit on standardized explanatory variables, shown in the Appendix. However, it only makes sense to standardize some numerical variables. Specifically, HarvestMonth is ordinal and should not be standardized, but Age can be because it is reasonable to think about fractions of years.

LogArea has the standardized coefficient with largest magnitude (0.8731) and relatively small standard error so the variable is important. Holding all other variables constant, a one standard deviation increase in logArea is associated with a 0.8731 standard deviation increase in logTonn.Hect. It is also interesting to note that the coefficient for logArea is very similar to the correlation of around 0.877 from Figure 3, which is the standardized coefficient from simple regression. Also, both logArea and HarvestDuration have small p-values in multiple regression, so they are each important even in the presence of the other variable. These demonstrate that there is no collinearity, addressing the previous concern.

After standardizing, it is interesting to observe that coefficients for PositionS, previously with a large p-value, now has p-value less than 0.05. The opposite holds for the intercept, which had small p-value for non-standardized model and larger p-value for standardized model. This likely occurs because using the same scale prevents some variables from masking changes in smaller-scale variables.

Other variables with low p-values are HarvestDuration, Age, and PositionW:Age. HarvestDuration has a positive association and small standard error relative to the coefficient. One standard deviation increase in Age is associated with about 0.1004 standard deviation decrease in logTonn.Hect, holding all other variables constant. However, its standard error of 0.025 is also quite large, so there is more uncertainty in the estimate. For paddocks in PositionW, Age has positive association, though it is hard to interpret this without additional context of the west area.

4.3 Prediction

The fitted model is now applied to unseen test data to evaluate how well the model generalizes and whether it is reasonable to trust the model. Comparing the training and test set root mean squared error (RMSE), 0.469 and 0.468 respectively, they are similar and suggest the test set performance is as good as that of training, though they do not capture the residual plot behavior. Figure 8 confirms that train and test perform similarly. Also, after undoing the log transformation, there is larger uncertainty at higher values of `Tonn.Hect`. The model may be able to predict future sugar production for new paddocks, but we should be cautious about the predictions, especially if they are smaller or larger than most fitted values.

5 Sugar Content

The other prediction task to inform investment in new paddocks is commercial sugar content. There are no strongly correlated variables from the correlation matrix in Figure 3, but the boxplots show several variables could have main effects due to changes in median across categories. Two interaction terms are also considered, motivated by changes in category mean differences.

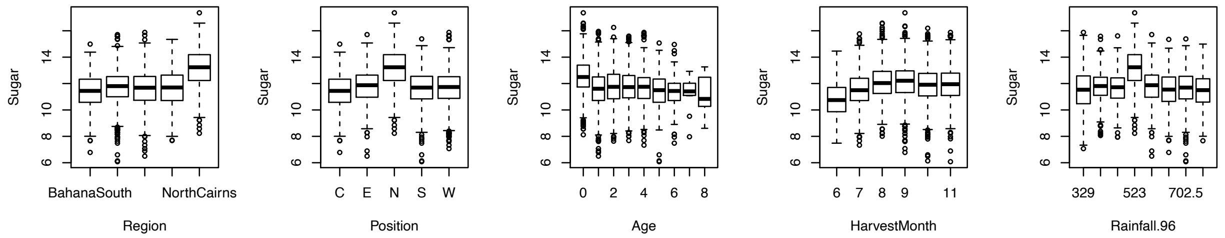

Figure 9: Bivariate distributions of explanatory variables against sugar content

5.1 Model Selection

The full model is: $Sugar \sim Region + Age + HarvestMonthFactor + Rainfall.96 + Region * Age + Region * HarvestMonthFactor$. The initial model used `HarvestMonth`, which had $R^2 = 0.1963$, but since month is better described as ordinal over continuous, the variable is factorized to define `HarvestMonthFactor`. This increases R^2 to 0.2586 and adjusted R^2 to 0.2499, so the increase in number of variables is not of concern.

Each variable except for `Rainfall.96` from `Anova` on the full model has statistically significant results, so we can reject the null that each coefficient is zero, using the incremental F-test. `Rainfall.96` will be excluded from the model. It has no interaction terms, so this obeys the principle of marginality. Comparing the F-test outputs from using `HarvestMonth` vs `HarvestMonthFactor`, there is very little difference, so we will use `HarvestMonthFactor` to both improve R^2 and interpretability.

```
## Anova Table (Type II tests)
##
## Response: Sugar
##                            Sum Sq   Df  F value    Pr(>F)
## Region                      725.0    4 117.5981 < 2.2e-16 ***
## Age                          43.7    1  28.3285   1.1e-07 ***
## HarvestMonthFactor          421.7    5  54.7196 < 2.2e-16 ***
## Rainfall.96                   0.0    1   0.0006  0.980258
## Region:Age                   22.9    4   3.7114  0.005103 **
## Region:HarvestMonthFactor  296.0   20   9.6009 < 2.2e-16 ***
## Residuals                 4599.2 2984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual plot in Figure 10 displays unusual patterns. First, there are no fitted values around 12.5 to 13 tons per batch processed. Second, the residuals tend to be clustered at certain fitted values, causing points to create vertical lines in the plot. To inspect the first behavior, the points corresponding to fitted values at least 12.5 are identified, and we find that all of them come from the NorthCairns region. As seen in Figure 9, NorthCairns has significantly higher `Sugar` distribution than the other regions, specifically with median around 13. Thus, while other regions also have some paddocks that produce high sugar content, most data points are from NorthCairns. Including this variable in the model causes it to predict that paddocks with high `Sugar` are from NorthCairns.

While we could exclude `Region`, the plots from EDA suggest it is important to include it as a main effect and in interaction terms. While investigating the residual plot patterns, it is also noted that all `Positions` N and `Rainfall.96` values of 523 mm are from paddocks in NorthCairns. This supports our decision to remove these redundant variables from the model.

The vertical banding of the residual plot is likely traced back to the use of many categorical variables in the model, including `HarvestMonthFactor`. Dummy variables allow the model to fit different linear relationships across categories. When categories are too finely divided, the model could have difficulty fitting linear relationships and instead just predict similar values for points within a group, like with NorthCairns. In this case, the continuous variables like `Area` and `HarvestDuration` do not help to predict `Sugar`. We either have these behaviors in predictions and residuals or have very low R^2, neither of which is desirable. For the remainder of the report, the model with many categorical variables is used, but this challenge remains a limitation of the model.

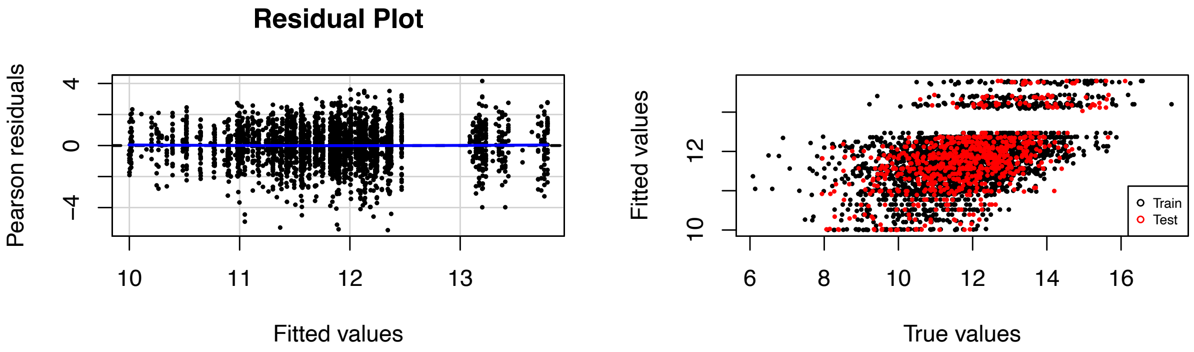

Figure 10: Residual plot and fitted values for sugar content produced

5.2 Model Interpretation

This model has an R^2 of 0.2586 and adjusted R^2 of 0.2501, so while the model is able to explain about 25.9% of the variation in sugar content, it does not perform well overall. Similar to the previous model, we consider the standardized partial coefficients so all values are on the same scale. Most variables used in the model are categorical, so only the response variable `Sugar` and explanatory variable `Age` are standardized. This has little impact on the coefficients. Aligning with the previous discussion, the p-value for main effect of Region NorthCairns is very low.

By using factored harvest months, there is a fitted coefficient for each dummy variable in the main effect and interaction terms with `Region`. An observation is that the Region Cairns/Mulgrave (Med-wet) has significant interaction terms with most harvest months except November. Harvesting in July seems to be important through interactions with most region types except MulgravetoBahana. Additionally, paddocks harvested in November within the MulgravetoBahana region are important for predicting sugar content. `Age` also matters, as one standard deviation increase corresponds to 0.1440 standard deviation decrease in sugar content produced. Another cautionary note is that all coefficients' standard errors are quite large, around 1/3 to 1/2 the value of the coefficient, so there is uncertainty in the estimates.

5.3 Prediction

The same training set was used to fit this model as for sugar production prediction, so we will use the same test set to evaluate the model. Given the behavior from the residual plot, it is not surprising that predictions on the test set resemble the patterns from the training set, as seen in Figure 10. The RMSE for training and test are 1.234 and 1.175 respectively.

6 Assessing Interventions

The sugar company wants to improve yield of the sugarcane fields it already owns. From the model predicting sugar production, two variables provide useful suggestions based on standardized coefficients. The first is to replant new sugarcane seeds more regularly, decreasing `Age`. Decreasing `Age` by one standard deviation can increase production by roughly 0.1 standard deviations. Increasing how many days it takes to harvest sugarcane could also lead to increased production, though this effect is smaller than changing `Age`. The model for sugar content reaches similar conclusions on decreasing `Age` being beneficial overall for improving yield, but for the Cairns/Mulgrave (Med-wet) region, the opposite holds. For that particular region, increasing `Age` by one standard deviation could lead to about 0.1499 standard deviation increase in sugar content. Since the coefficients of other regions' interactions terms with `Age` are not significant enough, it is difficult to tell how this compares across regions. Based on this, there does not seem to be trade-offs between improving production and sugar content overall, except for the Cairns/Mulgrave (Med-wet) region. Because of the limited information from the models and uncertainty of coefficients, the models can give some guidance but will probably not accurately describe causal impacts of these interventions.

7 Conclusion

This analysis studies two related prediction tasks: to predict sugar production and to predict sugar content. Both quantities are informative for the sugar company to evaluate yields and make decisions on how to improve it. The model for predicting sugar production is able to predict for test values well, so it can likely predict future production as well, assuming future data from paddocks are similar to the training set. Future data may be different if, for example, production is impacted by climate change or other shifts in environmental conditions. The predictions are less accurate near the lower and upper range of log sugar production, as shown by the residual plot in Figure 8, so some caution needs to be taken when interpreting results.

The model for predicting sugar content produced, on the other hand, is not precise at all and should not be used to reliably predict future sugar content. From EDA, categorical variables are most important and improve R^2, but they lead to undesirable behaviors in the fitted values and residuals. As a next step, it would be beneficial to add relevant continuous variables and interactions as an attempt to mitigate this problem. Given how (1) numerical variables like `Rainfall.96` are limited and discrete-like and (2) variables `Region`, `Position`, and `Rainfall.96` capture similar information, additional relevant variables not directly related to current variables in the dataset can potentially improve the model.

8 Appendix

```
library(car)
library(lattice)
library(ggplot2)
library(reshape2)
library(gridExtra)

# Import data
sugarcane <- read.csv("sugarcane.csv", header = TRUE)
# Log transform
sugarcane$LogTonn.Hect <- log(sugarcane$Tonn.Hect)
# Log transform to adjust right skewness
sugarcane$LogArea <- log(sugarcane$Area)
```

8.1 EDA

```
# Figure 1: Predictor variable histograms
par(mfrow=c(1, 3))
hist(sugarcane$Tonn.Hect,
     xlab = "Sugarcane Produced (Tons per Hectare)",
     main = "Distribution of Sugarcane Production")
hist(sugarcane$LogTonn.Hect,
     xlab = "Sugarcane Produced (Log Tons per Hectare)",
     main = "Distribution of Log Sugarcane Production")
hist(sugarcane$Sugar,
     xlab = "Sugar Content (Tons per Batch Processed)",
     main = "Distribution of Sugar Content")
```

```
# Figure 2: Univariate distributions of explanatory variables
par(mfrow=c(1, 4))
hist(sugarcane$Area, breaks = 20,
     xlab = "Area (hectares)", main = "Distribution of Area")
# Age - discrete
barplot(table(sugarcane$Age),
        xlab = "Age (years)", ylab = "Frequency", main = "Distribution of Age")
# Transform does not help with skewness
hist(sugarcane$HarvestDuration,
     xlab = "Duration (days)", main = "Distribution of Harvest Duration")
# Rainfall - continuous, but only 8 unique values
barplot(table(sugarcane$Rainfall.96),
        xlab = "Amount of rainfall (millimeters)", ylab = "Frequency",
        main = "Distribution of Rainfall")
```

```
# Figure 3: Correlation matrix for numerical variables
cor_matrix <- round(cor(sugarcane[c("LogTonn.Hect", "Sugar", "LogArea",
                                    "Age", "HarvestDuration", "Rainfall.96")]), 3)

ggplot(data = melt(cor_matrix), aes(x = Var1, y = Var2, fill = value)) +
      scale_fill_gradient2(low = "red", high = "blue", mid = "white",
                           midpoint = 0, limit = c(-1,1)) +
      geom_tile() +
      geom_text(aes(label = value), size = 2.5) +
      theme(axis.title.x = element_blank(),
            axis.title.y = element_blank(),
            axis.text.x = element_text(angle = 20, vjust = 1, size = 8, hjust = 1),
            axis.text.y = element_text(size = 8)) +
      guides(fill = guide_colorbar(barwidth = 1, barheight = 4,
                    title.position = "top"))
```

8.2 Sugar Production

```
# Figure 4: Scatterplot against log sugar production
par(mfrow=c(1, 2))
plot(sugarcane$LogArea, sugarcane$LogTonn.Hect,
     pch = 19, cex = 0.2,
     xlab = "Log Area", ylab = "Log Sugar Production")
plot(sugarcane$HarvestDuration, sugarcane$LogTonn.Hect,
     pch = 19, cex = 0.2,
     xlab = "Harvest Duration", ylab = "Log Sugar Production")

# Figure 5: Boxplots of categorical or discrete variables against log sugar production
par(mfrow=c(1, 3))
boxplot(LogTonn.Hect ~ Region, data = sugarcane)
boxplot(LogTonn.Hect ~ Position, data = sugarcane)
boxplot(LogTonn.Hect ~ Age, data = sugarcane)

# Figure 6: Interaction boxplot
ggplot(sugarcane, aes(x=factor(Age), y=LogTonn.Hect, fill=Position)) + geom_boxplot()

# Figure 7: Interaction coplot
xyplot(LogTonn.Hect ~ HarvestDuration | Position, data = sugarcane,
       show.given = FALSE, pch = 19, cex = 0.2, col = "black")
```

```
# Train-test split: 80-20%
set.seed(1)
n <- nrow(sugarcane)
sugarcane_shuffle <- sugarcane[sample(1:n),]
test <- sugarcane_shuffle[1:(n*0.2),]
train <- sugarcane_shuffle[(n*0.2 + 1):n,]
```

```
# Full model for sugar production
lm.full.prod <- lm(LogTonn.Hect ~ LogArea + HarvestDuration + Position + Age +
                     Position*Age + Position*HarvestDuration, data = train)
summary(lm.full.prod)
```

```
##
## Call:
## lm(formula = LogTonn.Hect ~ LogArea + HarvestDuration + Position +
##     Age + Position * Age + Position * HarvestDuration, data = train)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9287 -0.2073  0.0497  0.2577  3.4652
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.678e+00  3.782e-02 123.670  < 2e-16 ***
## LogArea                    8.959e-01  9.244e-03  96.913  < 2e-16 ***
## HarvestDuration            1.688e-03  1.027e-03   1.643   0.1004
## PositionE                 -1.194e-02  5.560e-02  -0.215   0.8299
## PositionN                 -2.151e-02  5.364e-02  -0.401   0.6884
## PositionS                 -8.155e-02  5.378e-02  -1.516   0.1295
## PositionW                 -3.269e-02  4.363e-02  -0.749   0.4538
## Age                       -6.101e-02  1.522e-02  -4.008 6.28e-05 ***
## PositionE:Age             -2.430e-02  2.068e-02  -1.175   0.2400
## PositionN:Age              1.208e-02  2.249e-02   0.537   0.5913
## PositionS:Age             -2.182e-02  1.962e-02  -1.112   0.2662
## PositionW:Age              3.684e-02  1.711e-02   2.153   0.0314 *
## HarvestDuration:PositionE  1.857e-03  1.540e-03   1.206   0.2279
## HarvestDuration:PositionN -8.874e-05  1.648e-03  -0.054   0.9571
## HarvestDuration:PositionS -1.958e-04  1.362e-03  -0.144   0.8857
## HarvestDuration:PositionW  2.242e-04  1.159e-03   0.193   0.8467
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4705 on 3004 degrees of freedom
## Multiple R-squared:  0.7789, Adjusted R-squared:  0.7778
## F-statistic: 705.5 on 15 and 3004 DF,  p-value: < 2.2e-16
```

```
# Shown in report
Anova(lm.full.prod)
```

```
# Final model
lm.final.prod <- lm(LogTonn.Hect ~ LogArea + HarvestDuration + Position + Age +
                      Position*Age, data = train)
summary(lm.final.prod)
```

```
##
## Call:
## lm(formula = LogTonn.Hect ~ LogArea + HarvestDuration + Position +
##     Age + Position * Age, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9263 -0.2049  0.0509  0.2582  3.4667
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.6752692  0.0368875 126.744  < 2e-16 ***
## LogArea         0.8959142  0.0092343  97.020  < 2e-16 ***
## HarvestDuration 0.0019569  0.0003968   4.931 8.62e-07 ***
## PositionE       0.0034047  0.0540212   0.063   0.9498
## PositionN       -0.0215299  0.0521270  -0.413   0.6796
## PositionS       -0.0843257  0.0520478  -1.620   0.1053
## PositionW       -0.0307350  0.0423266  -0.726   0.4678
## Age             -0.0610821  0.0152162  -4.014 6.11e-05 ***
## PositionE:Age   -0.0234912  0.0206636  -1.137   0.2557
## PositionN:Age    0.0121205  0.0224832   0.539   0.5899
## PositionS:Age   -0.0216069  0.0196135  -1.102   0.2707
## PositionW:Age    0.0369290  0.0171035   2.159   0.0309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4703 on 3008 degrees of freedom
## Multiple R-squared:  0.7787, Adjusted R-squared:  0.7779
## F-statistic: 962.3 on 11 and 3008 DF,  p-value: < 2.2e-16
```

```
# Figure 8: Residual plot and fitted values for log sugar production
par(mfrow=c(1, 2))
residualPlot(lm.final.prod, pch = 19, cex = 0.3,
              main = "Residual Plot")

train_pred <- predict(lm.final.prod, newdata = train)
test_pred <- predict(lm.final.prod, newdata = test)

# RMSE
sqrt(mean((train_pred - train$LogTonn.Hect)^2))
sqrt(mean((test_pred - test$LogTonn.Hect)^2))

plot(train$LogTonn.Hect, train_pred, pch = 19, cex = 0.3,
      xlab = "True values", ylab = "Fitted values")
points(test$LogTonn.Hect, test_pred, pch = 19, cex = 0.3, col=2)
legend(x = "bottomright", legend = c("Train", "Test"),
       col = c("black", "red"), pch=1, cex = 0.6)
```

```
# Final model, standardized coefficients
LogTonn.Hect <- train$LogTonn.Hect
LogArea <- train$LogArea
HarvestDuration <- train$HarvestDuration
Age <- train$Age

LogTonn.Hect_std <- (LogTonn.Hect - mean(LogTonn.Hect)) / sd(LogTonn.Hect)
LogArea_std <- (LogArea - mean(LogArea)) / sd(LogArea)
HarvestDuration_std <- (HarvestDuration - mean(HarvestDuration)) / sd(HarvestDuration)
Age_std <- (Age - mean(Age)) / sd(Age)

lm.final.std.prod <- lm(LogTonn.Hect_std ~ LogArea_std + HarvestDuration_std +
                          train$Position + Age_std + train$Position*Age_std)
summary(lm.final.std.prod)
```

```
##
## Call:
## lm(formula = LogTonn.Hect_std ~ LogArea_std + HarvestDuration_std +
##     train$Position + Age_std + train$Position * Age_std)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9320 -0.2053  0.0510  0.2587  3.4735
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.002600   0.023862   0.109   0.9132
## LogArea_std              0.873088   0.008999  97.020  < 2e-16 ***
## HarvestDuration_std      0.043868   0.008896   4.931 8.62e-07 ***
## train$PositionE         -0.046648   0.033847  -1.378   0.1682
## train$PositionN          0.004257   0.035359   0.120   0.9042
## train$PositionS         -0.130535   0.032921  -3.965 7.51e-05 ***
## train$PositionW          0.047900   0.026933   1.778   0.0754 .
## Age_std                 -0.100387   0.025007  -4.014 6.11e-05 ***
## train$PositionE:Age_std -0.038607   0.033960  -1.137   0.2557
## train$PositionN:Age_std  0.019920   0.036951   0.539   0.5899
## train$PositionS:Age_std -0.035510   0.032234  -1.102   0.2707
## train$PositionW:Age_std  0.060692   0.028109   2.159   0.0309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4713 on 3008 degrees of freedom
## Multiple R-squared:  0.7787, Adjusted R-squared:  0.7779
## F-statistic: 962.3 on 11 and 3008 DF,  p-value: < 2.2e-16
```

```
# Figure: Test fitted values and prediction intervals for sugar production
# Not included in report due to lack of space
new_pi <- exp(predict(lm.final.prod, newdata=test,
                      interval = "prediction", level = 0.95))
new_pi <- data.frame(new_pi)
new_pi$true <- test$Tonn.Hect

ggplot(new_pi, aes(x = true, y = fit)) + geom_point(size = 0.7) +
      geom_line(aes(y = true), col = "blue") +
      geom_line(aes(y = lwr), col = "coral2", linetype = "dashed") + # lower pred interval
      geom_line(aes(y = upr), col = "coral2", linetype = "dashed") + # upper pred interval
      labs(x = "True values", y = "Fitted values", title = "Test Set: Sugar Production")
```

8.3 Sugar Content

```
# Figure 9: Bivariate distributions of explanatory variables against sugar content
par(mfrow=c(1, 5))
boxplot(Sugar ~ Region, data = sugarcane)
boxplot(Sugar ~ Position, data = sugarcane)
boxplot(Sugar ~ Age, data = sugarcane)
boxplot(Sugar ~ HarvestMonth, data = sugarcane)
boxplot(Sugar ~ Rainfall.96, data = sugarcane)
```

```
# Figure: Interaction plot for sugar content
# Not included in report due to lack of space
ggplot(sugarcane, aes(x=factor(Age), y=Sugar, fill=Region)) + geom_boxplot()
ggplot(sugarcane, aes(x=factor(HarvestMonth), y=Sugar, fill=Region)) + geom_boxplot()
```

```r
# Full model for sugar content
lm.full.content.month <- lm(Sugar ~ Region + Age + HarvestMonth + Rainfall.96 +
                              Region * Age + Region * HarvestMonth, data = train)
Anova(lm.full.content.month)
summary(lm.full.content.month)

# Full model using HarvestMonthFactor
train$HarvestMonthFactor <- factor(train$HarvestMonth)
lm.full.content <- lm(Sugar ~ Region + Age + HarvestMonthFactor + Rainfall.96 +
                        Region * Age + Region * HarvestMonthFactor, data = train)
Anova(lm.full.content) # shown in report
summary(lm.full.content)

# Final model
lm.final.content <- lm(Sugar ~ Region + Age + HarvestMonthFactor +
                         Region * Age + Region * HarvestMonthFactor, data = train)
summary(lm.final.content)
```

```r
# Figure 10: Residual plot and fitted values for sugar content produced
par(mfrow=c(1, 2))
residualPlot(lm.final.content, pch = 19, cex = 0.3,
              main = "Residual Plot")

train_pred <- predict(lm.final.content, newdata = train)
test$HarvestMonthFactor <- factor(test$HarvestMonth)
test_pred <- predict(lm.final.content, newdata = test)

# RMSE
sqrt(mean((train_pred - train$Sugar)^2))
sqrt(mean((test_pred - test$Sugar)^2))

plot(train$Sugar, train_pred, pch = 19, cex = 0.3,
      xlab = "True values", ylab = "Fitted values")
points(test$Sugar, test_pred, pch = 19, cex = 0.3, col=2)
legend(x = "bottomright", legend = c("Train", "Test"),
       col = c("black", "red"), pch=1, cex = 0.6)
```

```r
# Final model, standardized coefficients
Sugar <- train$Sugar
Age <- train$Age
Sugar_std <- (Sugar - mean(Sugar)) / sd(Sugar)
Age_std <- (Age - mean(Age)) / sd(Age)

lm.final.std.content <- lm(Sugar_std ~ train$Region + Age_std + train$HarvestMonthFactor +
                           train$Region * Age_std + train$Region * train$HarvestMonthFactor,
                           data = train)
summary(lm.final.std.content)
```

```
##
## Call:
## lm(formula = Sugar_std ~ train$Region + Age_std + train$HarvestMonthFactor +
##     train$Region * Age_std + train$Region * train$HarvestMonthFactor,
##     data = train)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8047 -0.4996  0.0293  0.5851  2.9020
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                      -0.761516
## train$RegionCairns/Mulgrave(dry)                                  0.025664
## train$RegionCairns/Mulgrave(Med-wet)                             -0.498164
## train$RegionMulgravetoBahana                                     -0.302367
## train$RegionNorthCairns                                           1.092007
## Age_std                                                          -0.144032
## train$HarvestMonthFactor7                                         0.021215
## train$HarvestMonthFactor8                                         0.506447
## train$HarvestMonthFactor9                                         0.711749
## train$HarvestMonthFactor10                                        0.712148
## train$HarvestMonthFactor11                                        0.453782
## train$RegionCairns/Mulgrave(dry):Age_std                         -0.001756
## train$RegionCairns/Mulgrave(Med-wet):Age_std                      0.149875
## train$RegionMulgravetoBahana:Age_std                              0.091268
## train$RegionNorthCairns:Age_std                                   0.111642
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor7        0.710264
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor7    0.663735
## train$RegionMulgravetoBahana:train$HarvestMonthFactor7            0.555009
## train$RegionNorthCairns:train$HarvestMonthFactor7                 0.596894
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor8        0.430962
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor8    0.777022
## train$RegionMulgravetoBahana:train$HarvestMonthFactor8            0.250050
## train$RegionNorthCairns:train$HarvestMonthFactor8                 0.501778
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor9        0.292887
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor9    0.378723
## train$RegionMulgravetoBahana:train$HarvestMonthFactor9            0.574418
## train$RegionNorthCairns:train$HarvestMonthFactor9                -0.116360
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor10     -0.255044
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor10  0.610485
## train$RegionMulgravetoBahana:train$HarvestMonthFactor10          0.577471
## train$RegionNorthCairns:train$HarvestMonthFactor10               0.048560
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor11      0.226930
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor11  1.180321
## train$RegionMulgravetoBahana:train$HarvestMonthFactor11          0.806522
## train$RegionNorthCairns:train$HarvestMonthFactor11              -0.525896
##                                                                 Std. Error
## (Intercept)                                                       0.210851
## train$RegionCairns/Mulgrave(dry)                                  0.226067
## train$RegionCairns/Mulgrave(Med-wet)                              0.248341
## train$RegionMulgravetoBahana                                      0.259210
## train$RegionNorthCairns                                           0.247232
## Age_std                                                           0.047986
## train$HarvestMonthFactor7                                         0.229437
## train$HarvestMonthFactor8                                         0.236293
## train$HarvestMonthFactor9                                         0.230617
## train$HarvestMonthFactor10                                        0.236332
## train$HarvestMonthFactor11                                        0.241981
## train$RegionCairns/Mulgrave(dry):Age_std                          0.054322
## train$RegionCairns/Mulgrave(Med-wet):Age_std                      0.059922
## train$RegionMulgravetoBahana:Age_std                              0.071094
## train$RegionNorthCairns:Age_std                                   0.072026
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor7        0.249400
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor7    0.276864
## train$RegionMulgravetoBahana:train$HarvestMonthFactor7            0.292301
## train$RegionNorthCairns:train$HarvestMonthFactor7                 0.281873
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor8        0.256508
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor8    0.282422
## train$RegionMulgravetoBahana:train$HarvestMonthFactor8            0.301951
## train$RegionNorthCairns:train$HarvestMonthFactor8                 0.289659
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor9        0.252076
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor9    0.275697
## train$RegionMulgravetoBahana:train$HarvestMonthFactor9            0.297827
## train$RegionNorthCairns:train$HarvestMonthFactor9                 0.287383
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor10       0.255894
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor10   0.278709
## train$RegionMulgravetoBahana:train$HarvestMonthFactor10           0.296306
## train$RegionNorthCairns:train$HarvestMonthFactor10                0.290576
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor11       0.265456
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor11   0.288479
## train$RegionMulgravetoBahana:train$HarvestMonthFactor11           0.310332
## train$RegionNorthCairns:train$HarvestMonthFactor11                0.309271
##                                                                 t value
## (Intercept)                                                      -3.612
## train$RegionCairns/Mulgrave(dry)                                  0.114
## train$RegionCairns/Mulgrave(Med-wet)                             -2.006
## train$RegionMulgravetoBahana                                     -1.166
## train$RegionNorthCairns                                           4.417
## Age_std                                                          -3.002
## train$HarvestMonthFactor7                                         0.092
## train$HarvestMonthFactor8                                         2.143
## train$HarvestMonthFactor9                                         3.086
## train$HarvestMonthFactor10                                        3.013
## train$HarvestMonthFactor11                                        1.875
## train$RegionCairns/Mulgrave(dry):Age_std                         -0.032
## train$RegionCairns/Mulgrave(Med-wet):Age_std                      2.501
## train$RegionMulgravetoBahana:Age_std                              1.284
## train$RegionNorthCairns:Age_std                                   1.550
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor7        2.848
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor7    2.397
## train$RegionMulgravetoBahana:train$HarvestMonthFactor7            1.899
## train$RegionNorthCairns:train$HarvestMonthFactor7                 2.118
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor8        1.680
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor8    2.751
## train$RegionMulgravetoBahana:train$HarvestMonthFactor8            0.828
## train$RegionNorthCairns:train$HarvestMonthFactor8                 1.732
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor9        1.162
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor9    1.374
## train$RegionMulgravetoBahana:train$HarvestMonthFactor9            1.929
## train$RegionNorthCairns:train$HarvestMonthFactor9                -0.405
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor10      -0.997
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor10   2.190
## train$RegionMulgravetoBahana:train$HarvestMonthFactor10           1.949
## train$RegionNorthCairns:train$HarvestMonthFactor10                0.167
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor11       0.855
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor11   4.092
## train$RegionMulgravetoBahana:train$HarvestMonthFactor11           2.599
## train$RegionNorthCairns:train$HarvestMonthFactor11               -1.700
##                                                                 Pr(>|t|)
## (Intercept)                                                     0.000309 ***
## train$RegionCairns/Mulgrave(dry)                                0.909623
## train$RegionCairns/Mulgrave(Med-wet)                            0.044950 *
## train$RegionMulgravetoBahana                                    0.243508
## train$RegionNorthCairns                                         1.04e-05 ***
## Age_std                                                         0.002708 **
## train$HarvestMonthFactor7                                       0.926335
## train$HarvestMonthFactor8                                       0.032170 *
## train$HarvestMonthFactor9                                       0.002045 **
## train$HarvestMonthFactor10                                      0.002606 **
## train$HarvestMonthFactor11                                      0.060852 .
## train$RegionCairns/Mulgrave(dry):Age_std                        0.974216
## train$RegionCairns/Mulgrave(Med-wet):Age_std                    0.012432 *
## train$RegionMulgravetoBahana:Age_std                            0.199323
## train$RegionNorthCairns:Age_std                                 0.121241
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor7      0.004431 **
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor7  0.016576 *
## train$RegionMulgravetoBahana:train$HarvestMonthFactor7          0.057692 .
## train$RegionNorthCairns:train$HarvestMonthFactor7               0.034292 *
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor8      0.093041 .
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor8  0.005972 **
## train$RegionMulgravetoBahana:train$HarvestMonthFactor8          0.407671
## train$RegionNorthCairns:train$HarvestMonthFactor8               0.083322 .
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor9      0.245369
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor9  0.169640
## train$RegionMulgravetoBahana:train$HarvestMonthFactor9          0.053863 .
## train$RegionNorthCairns:train$HarvestMonthFactor9               0.685583
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor10     0.319001
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor10 0.028572 *
## train$RegionMulgravetoBahana:train$HarvestMonthFactor10         0.051401 .
## train$RegionNorthCairns:train$HarvestMonthFactor10              0.867291
## train$RegionCairns/Mulgrave(dry):train$HarvestMonthFactor11     0.392693
## train$RegionCairns/Mulgrave(Med-wet):train$HarvestMonthFactor11 4.40e-05 ***
## train$RegionMulgravetoBahana:train$HarvestMonthFactor11         0.009398 **
## train$RegionNorthCairns:train$HarvestMonthFactor11              0.089153 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.866 on 2985 degrees of freedom
## Multiple R-squared:  0.2586, Adjusted R-squared:  0.2501
## F-statistic: 30.62 on 34 and 2985 DF,  p-value: < 2.2e-16
```