Linear Regression Project: Queensland Sugarcane

Iris Kim (6757827), Ted Tinker (3223468), Tianzhi Zhai (9843723), and Ray Cai (9482068)

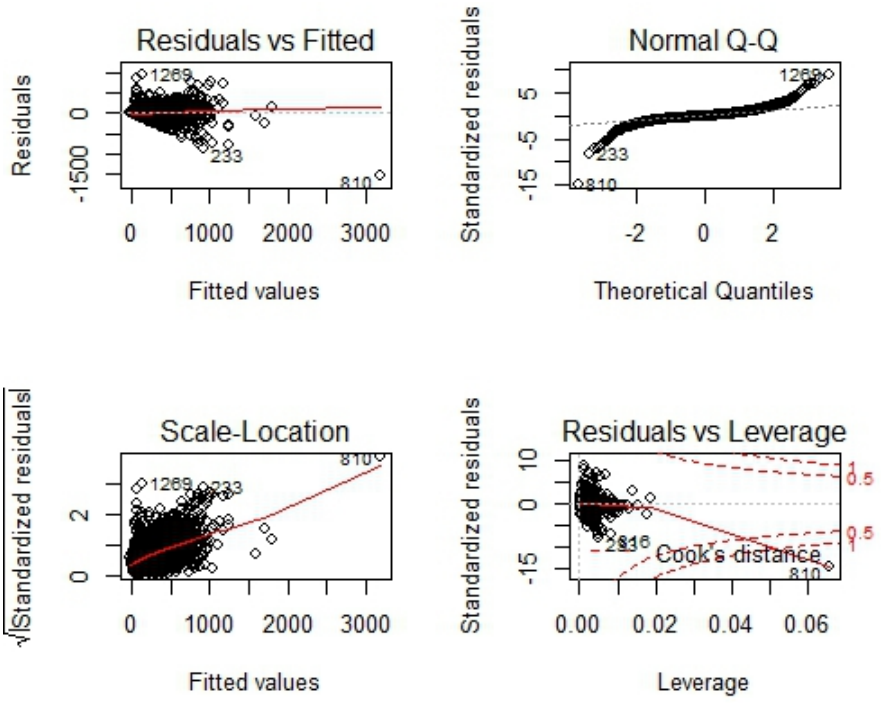
June 10, 2017

Abstract

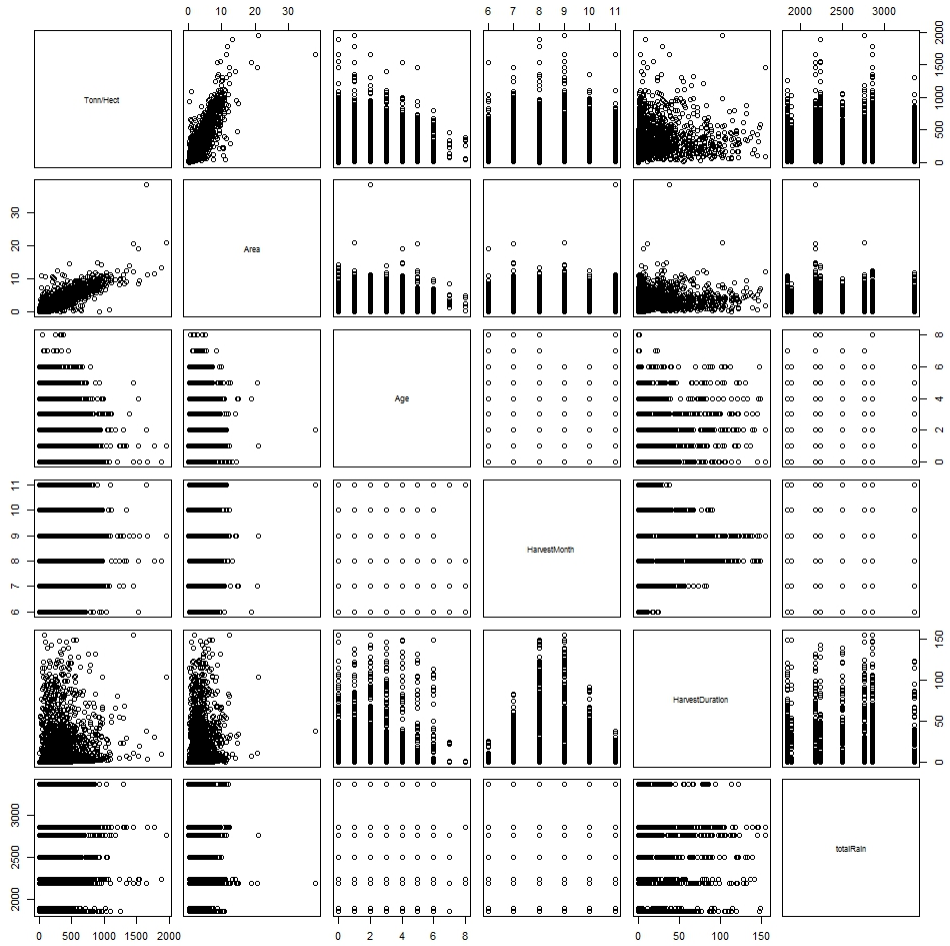
In 1998, the University of Queensland obtained a dataset of statistics related the 1997 sugarcane season in the Mulgrave area of North Queensland, Australia. We seek to model the quantity of sugarcane produced (in terms of tonnes per hectare), the fibre per rake, and the sugar content per rake (two measures of quality). Predictor variables include quantifiable terms like the size of the paddock, the duration of the harvest, and rainfall data for the preceding year and a half, and categorical variables like the location of the paddock in Queensland and the type of sugarcane grown. In the Method section, we determine which predictor variables influence the response variables and use this data to generate models. In the Analysis, we address these models, account for the different types of sugarcane and soil, and determine their effect on sugar production. Finally, we review our models and draw conclusions.

Method

Having loaded the data into R-Studio, we visually assess which predictor variables are explanatory for the sugar quantity and quality, and what transformations should be applied so that linear models appear valid. In the case of sugar quantity (in terms of tonnes per hectare), we begin by naively making a regression model using all predictor variables, summing the rainfall totals into one "Total Rain" variable. This produces the following diagnostic plots:



These diagnostic plots show clear patterns which must be accounted for before using a purely linear regression. The Normal Q-Q plot in particular shows a strong sigmoidal trend in the residuals. There are three noticeable outliers, points 233, 810, and 1269, which we will reevaluate after adjusting the model. By making the matrix plots for the variables used, we find which need to be transformed to make a linear model:



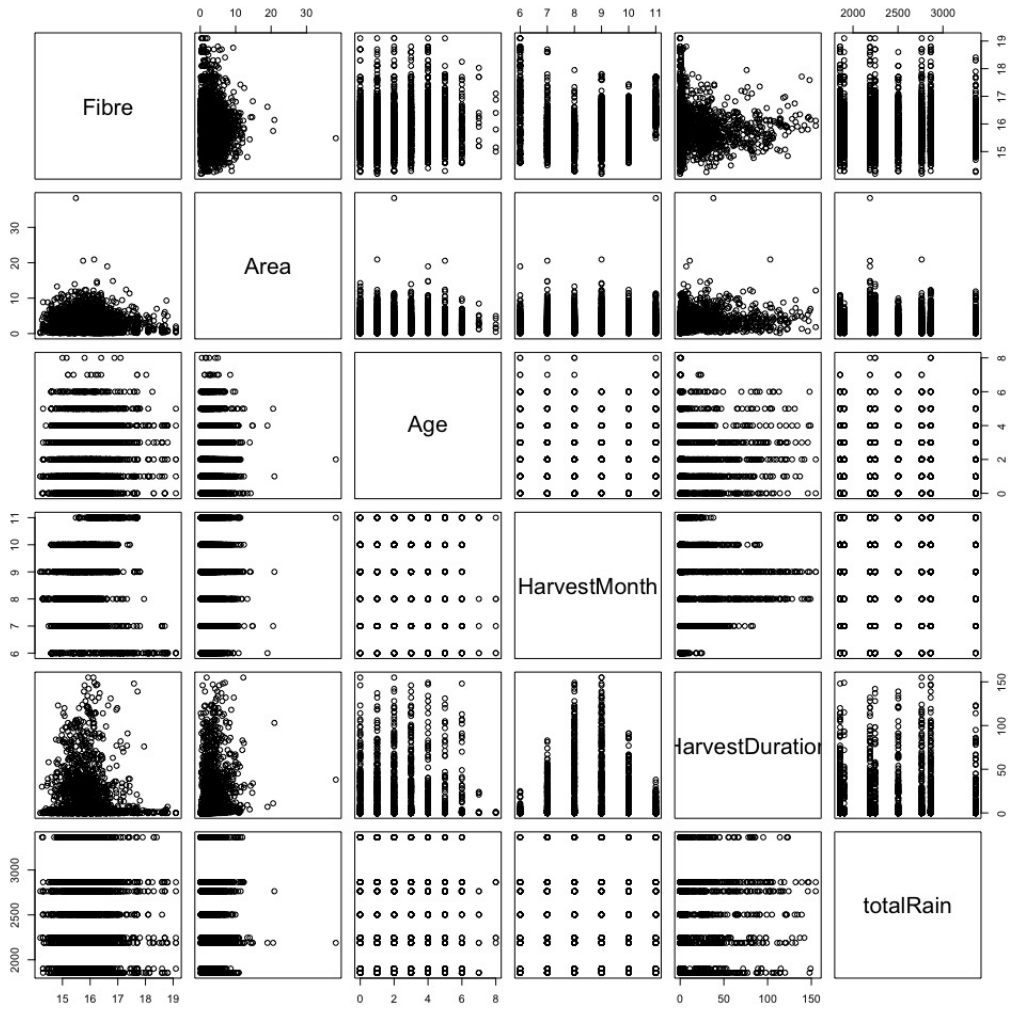
Tonnage per hectare is linearly related to farm size (larger farms produce more per hectare), and inversely related to Age. The impact of Harvest Month is not obvious (even if the R-Summary gives it three stars). Harvest tonnage decreases as harvest duration increases, except for two outliers in the top right, which we discount. "Total Rain" has a bimodal distribution which may be explained by different sugarcane strains.

Using backwards AIC testing to select the months during which rainfall most influences sugarcane growth, July, August, September, November, and December of 1996 are found to be most influential. These months correspond to the Australian Winter and Summer of the year prior to harvest. This result is interesting, but has no significant, consistent effect on the data. Therefore, our suggested model is

$$Tonnes\ of\ Sugar\ per\ Hectare = 51.4 + 83.1(Area) - 13.5(Age) + .0345(HarvestDuration)$$

The diagnostic plots for this model still show a strong sigmoid trend in the residuals, and points 233, 810, and 1269 are still outliers. Point 810 has the highest leverage of all points, but does not seem to pull the regression line. Interestingly, Harvest Duration's impact is positive, suggesting that most farms find themselves on the lower edge of the triangle of data-points.

Now, we perform the same operations to find models for the fibre content and sugar content of the harvests. Here is the matrix plot for the predictors with relation to sugarcane fibre content, showing a strange trend for Area, and little trend with respect to Age, but a quadratic relation to Harvest Month and Harvest Duration. TotalRain is again negligible:

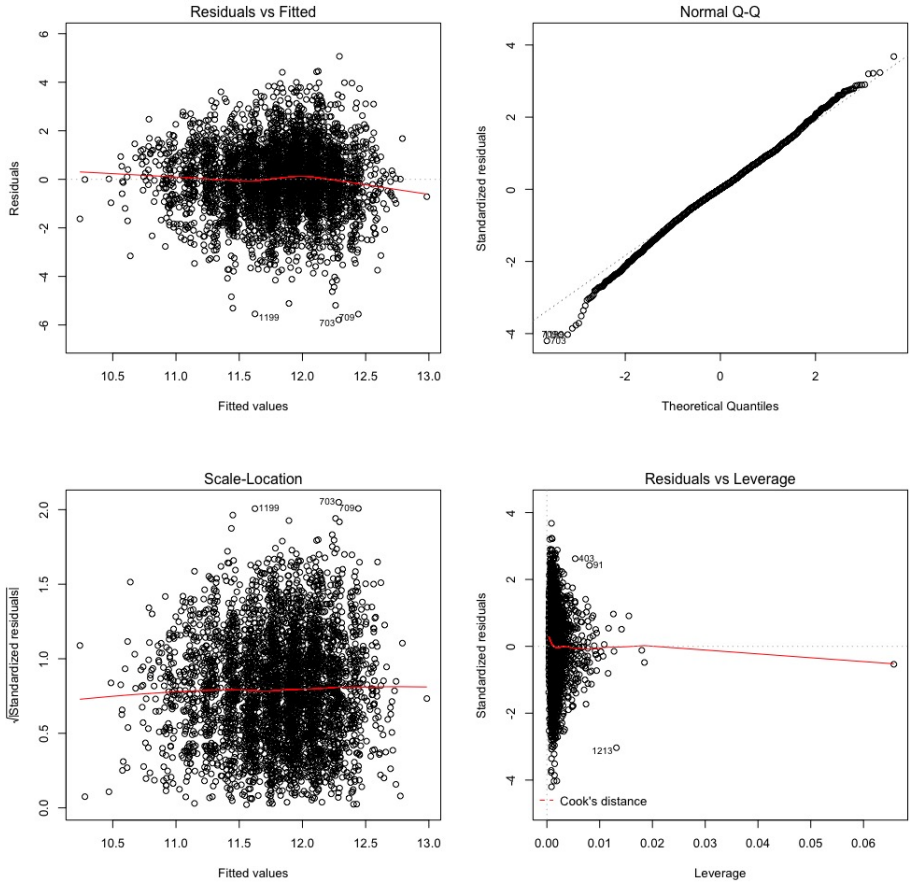


Taking the square roots of Harvest Month and Harvest Duration does not significantly improve their linear applicability to the data, so unfortunately, these quadratic trends are not helpful in interpretation.

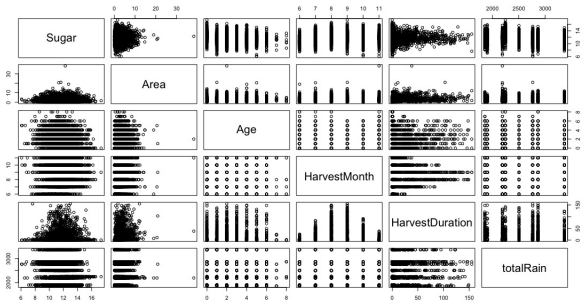
This time, performing backwards AIC steps to determine which months' rainfall amounts are most influential, there is not much pattern to the most significant months. Therefore, we suggest the following model:

$$Fibre\ per\ Rake = 15.1 - .0078(Area) + .089(HarvestMonth) - .001(HarvestDuration)$$

Finally, the diagnostic plots for sugar content look well-behaved, in the sense that the residuals have consistent, normal variance, the Q-Q plot shows a strong one-to-one trend, and the outliers have low leverage:



Looking to the matrix plot to see which variables are significant, we see well-behaved, linear trends:



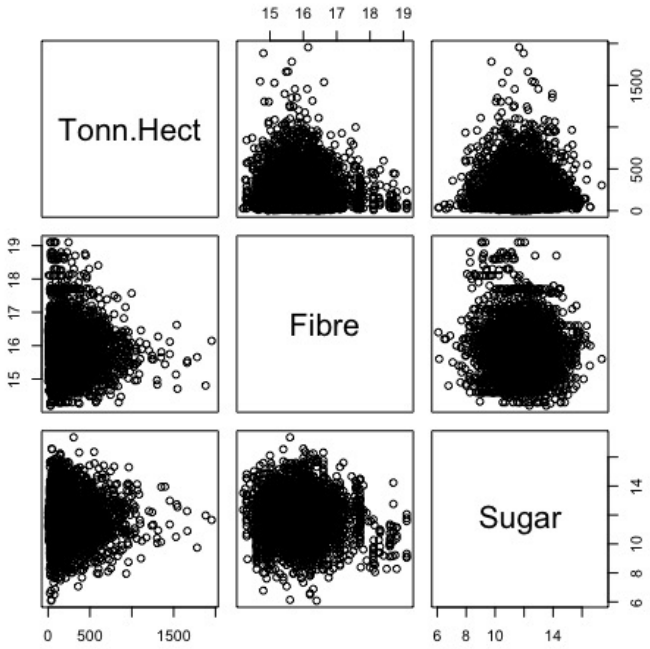
Backwards AIC tests suggest the most influential months for rainfall are July, August, September, November, and December of the previous year, and January and February of 97, the Australian Winter and Summer. However, they are not as influential as Age and Harvest Month. Our final model selection is

$$Sugar\ per\ Rake = 7.63 - .137(Age) + .176(HarvestMonth)$$

Analysis

Under the models we propose, a farmer wishing to grow as much sugarcane as possible (in terms of tonnage) should operate a large farm (maximize area), a young farm (minimize age), and harvest quickly. However, the most efficient farms might be large farms because those farms are owned by experienced farmers; the most efficient farms might be young farms because those farms have the newest, most efficient technology. This would explain the distribution of harvest tonnage with respect to harvest duration: these experienced farmers, using the most efficient technology, would harvest a large, high-efficiency crop within months, while other farmers take home smaller harvests over longer durations, with a few outliers who dedicate years to collecting larger harvests.

With the same understanding, a farmer producing fibrous sugarcane typically owns a small farm, harvests later in the year, and harvests as quickly as possible. A farmer who maximizes sugar per rake probably operates a young farm and harvests late in the year.



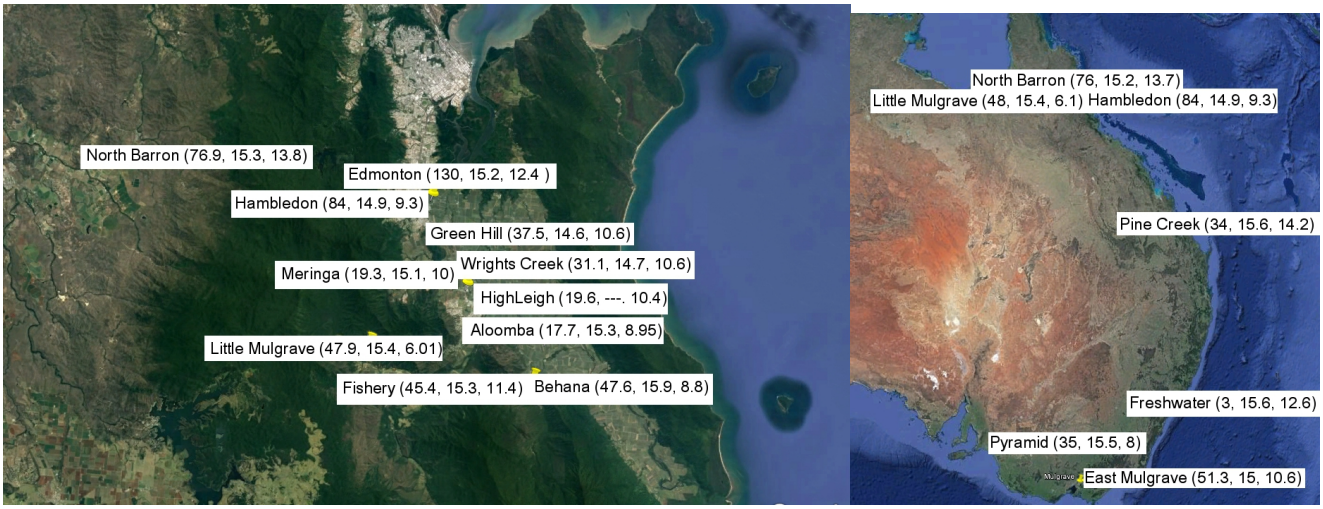
This matrix plot comparing the three response variables shows how prioritizing Fibre content decreases the expected weight of the harvest. Low-sugar rakes and high-sugar rakes come from smaller harvests, while the vast majority of harvests have middling sugar-content. Sugar and Fiber seem mostly uncorrelated, though some interesting banding occurs. Perhaps industry standards in fiber production encourage rounding to integer values, resulting in strata.

Having found these models, we can use their intercepts after categorizing by soil type and sugarcane species as a measure of how these variables influence production, fibre, and sugar content. Conditioning by soil, sugarcane grown in 'Liverpool' soil has higher sugar content by three units. 'Prior' soil makes for small harvests (an intercept of 19.4 Tonnes/Hectare, instead of 51.4) with enormous sugar content. 'Holloway' soil produces huge harvests, but expected production is less boosted by Area of the farm, and more penalized by Age of the farm (though there are only about 30 Holloway entries, so confounding variables might be the cause). 'Buchan' soil seems to produce high-sugar-content rakes. 'Clifton' soil produces smaller yields. 'Ramleh' produces high-fibre (18 at intercept), low-sugar (6.5 at intercept) yields. 'Galmara' produces low-fibre (13 at intercept), high-sugar (11.5 at intercept) yields. A farm using 'Inlet' soil expects 142 tonnes a hectare,

almost three times the intercept of the standard model, with high sugar (12) and standard fibre. With 'Silkwood' soil, that's 208 tonnes a hectare. With 'Coom,' 212. These supowerpowered 'Inlet,' 'Silkwood,' and 'Coom' soils account for only about 100 entries. 'Lugger' soil has low yield, but sugar content around 14 units.

Now, we check for similar trends in the type of sugar-cane grown. Variety 117 has high sugar content (11.6). Variety 152 appears to have low/no fibre at all. Variety 99 produces huge harvests of highly fibrous (20 units!) sugarcane.

Finally, let us plot these locations by area in Australia. Below, find labels specifying the location and the expected tonnage/hectare, fibre per rake, and sugar per rake for a farm whose predictor variables are all equal to 0, so they may be directly compared:



It seems the Northern area of Queensland provides a consistent harvest efficiency, while a few of the Southern plantations excelled in sugar content. Fibre content is evenly distributed by area.

Conclusion

If you want to grow sugarcane in Australia, the best areas seem to be in North Queensland. That's where areas like Edmonton offer tonnes per hectare almost triple the normal rate.

A farmer on a small, young farm should use Holloway, Inlet, Silkwood, or Coom soils to maximize harvest tonnage, but these soils appear to be rarer than standards like Liverpool, which accounts for more than 20% of the soil in the data. If a farmer wants fibrous sugarcane (perhaps to use the fibre for production), Ramleh soil is most efficient, and sugarcane varieties like 99 would be a good place to start. Lugger soil may be best for producing high-sugar content sugarcane.

Appendix

R-Code Used for Page 1

```
attach(cane) # Primes dataset for use
totalRain = `Jul-96` + `Aug-96` + `Sep-96` + `Nov-96` + `Dec-96` + `Jan-97` + `Feb-97` + `Mar-97`
    + `Apr-97` + `May-97` + `Jun-97` + `Jul-97` + `Aug-97` + `Sep-97` + `Nov-97` + `Dec-97`
    # For the model, combine rainfalls
caneTonnageLine <- lm(`Tonn/Hect` ~ Area + Age + HarvestMonth + HarvestDuration + totalRain)
    # First, naive model using all predictors
par(mfrow=c(2,2)) # Asks R for four plots in a 2x2 square
plot(caneTonnageLine) # Produces the diagnostic plot
```

R-Code Used for Page 2

```
par(mfrow=c(6,6)) # Asks R for 36 plots in a 6x6 square
pairs(`Tonn/Hect` ~ Area + Age + HarvestMonth + HarvestDuration + totalRain) # Matrix plots
step(lm(`Tonn/Hect` ~ Area + Age + HarvestMonth + HarvestDuration + `Jul-96` + `Aug-96` + `Sep-96`
    + `Nov-96` + `Dec-96` + `Jan-97` + `Feb-97` + `Mar-97` + `Apr-97` + `May-97` + `Jun-97` +
    `Jul-97` + `Aug-97` + `Sep-97` + `Nov-97` + `Dec-97`), direction="backward")
        # AIC backwards step regression
```

At this step, cane.txt was altered to remove two outliers. The points were replaced afterward.

```
caneTonnageLine <- lm(`Tonn/Hect` ~ Area + Age + HarvestDuration) # Reassign according to model
summary(caneTonnageLine) # Provides details for writing the model
```

R-Code Used for Page 3

```
caneFibreLine <- lm(Fibre ~ Area + Age + HarvestMonth + HarvestDuration + totalRain)
    # Naive model using all predictors
par(mfrow=c(2,2))
plot(caneFibreLine) # Diagnostic plot, not pictured
par(mfrow=c(6,6))
pairs(Fibre ~ Area + Age + HarvestMonth + HarvestDuration + totalRain) # Matrix plot, pictured
step(lm(Fibre ~ Area + Age + HarvestMonth + HarvestDuration + `Jul-96` + `Aug-96` + `Sep-96` +
    `Nov-96` + `Dec-96` + `Jan-97` + `Feb-97` + `Mar-97` + `Apr-97` + `May-97` + `Jun-97` +
    `Jul-97` + `Aug-97` + `Sep-97` + `Nov-97` + `Dec-97`), direction="backward")
        # Backwards AIC testing
caneFibreLine <- lm(Fibre ~ Area + HarvestMonth +  HarvestDuration) # Reassign according to model
summary(caneFibreLine) # Provides details for writing the model
```

R-Code Used for Page 4

```
caneSugarLine <- lm(Sugar ~ Area + Age + HarvestMonth + HarvestDuration + totalRain)
    # Naive model using all predictors
par(mfrow=c(2,2))
plot(caneSugarLine) # Diagnostic plot
par(mfrow=c(6,6))
pairs(Fibre ~ Area + Age + HarvestMonth + HarvestDuration + totalRain) # Matrix plot
step(lm(Sugar ~ Area + Age + HarvestMonth + HarvestDuration + `Jul-96` + `Aug-96` + `Sep-96`
    + `Nov-96` + `Dec-96` + `Jan-97` + `Feb-97` + `Mar-97` + `Apr-97` + `May-97` + `Jun-97`
    + `Jul-97` + `Aug-97` + `Sep-97` + `Nov-97` + `Dec-97`), direction="backward")
        # Backward AIC step regression
caneSugarLine <- lm(Sugar ~ Age + HarvestMonth) # Reassign according to model
summary(caneSugarLine) # Provides details for writing the model
```

R-Code Used for Page 5

```
par(mfrow=c(3,3))
pairs(`Tonn/Hect` ~ Fibre + Sugar) # Matrix plot of response variables
caneSoil <- subset(cane, SoilName==``Soil Type Here'')
    # Makes a subset of entries using a specific soil
summary(lm(caneSoil$`Tonn/Hect` ~ caneSoil$Area + caneSoil$Age + caneSoil$HarvestDuration))
summary(lm(caneSoil$Fibre ~ caneSoil$Area + caneSoil$HarvestMonth + caneSoil$HarvestDuration))
summary(lm(caneSoil$Sugar ~ caneSoil$Age + caneSoil$HarvestMonth))
    # Three summaries for observing changes in the model based on soil
```

R-Code Used for Page 6

```
caneSugarType <- subset(cane, Variety==``Sugarcane Variety Here'')
    # Makes a subset of entries using a specific sugarcane strain
summary(lm(caneSugarType$`Tonn/Hect` ~ caneSugarType$Area + caneSugarType$Age +
    caneSugarType$HarvestDuration))
summary(lm(caneSugarType$Fibre ~ caneSugarType$Area + caneSugarType$HarvestMonth +
    caneSugarType$HarvestDuration))
summary(lm(caneSugarType$Sugar ~ caneSugarType$Age + caneSugarType$HarvestMonth))
    # Three summaries for observing changes in the model based on sugarcane strain

caneLoci <- subset(cane, District==``Location Here'')
    # Makes a subset of entries from a specific Queensland district
summary(lm(caneLoci$`Tonn/Hect` ~ caneLoci$Area + caneLoci$Age + caneLoci$HarvestDuration))
summary(lm(caneLoci$Fibre ~ caneLoci$Area + caneLoci$HarvestMonth + caneLoci$HarvestDuration))
summary(lm(caneLoci$Sugar ~ caneLoci$Age + caneLoci$HarvestMonth))
    # Three summaries for observing changes in the model based on district
```

The images were produced using an image editor and publicly available images.