Linear Modelling in R

Andre Archer
Northwestern University
Research Computing Services

Format of the Online Workshop

- In this workshop, I will be using Google Slides and live coding in RStudio
  - I will be using the Rmd file, linear_model_code.Rmd, to teach the workshop.

- If you have an questions, please put them in the chat.
  - There are TAs monitoring the chat. They will respond to questions.
  - If necessary, I will be interrupted by a TA.

Contents

Data Description

Data we are working with

- Dataset contains 18729 samples of avocado prices and volume sold across U.S. cities
- The dataset set contains variables:
  - AveragePrice - average price of avocado
  - TotalVolume - total volume sold
  - Type - whether the avocado was organic or conventional
  - Year - year in which the recording was made
  - Region - region in the U.S. the recording was made
  - Month - month in the recording was made

Goals of this workshop

Goal

1) I would like to predict average price as function of total volume sold (proxy for demand) and other explanatory variables

2) I would also like to understand the effects of total volume, type and year on the average price

Why Linear Model is appropriate

1) I would like to predict average price as function of total volume sold (proxy for demand), Type and Year
   - Linear model expresses the response variable as a function of explanatory variable

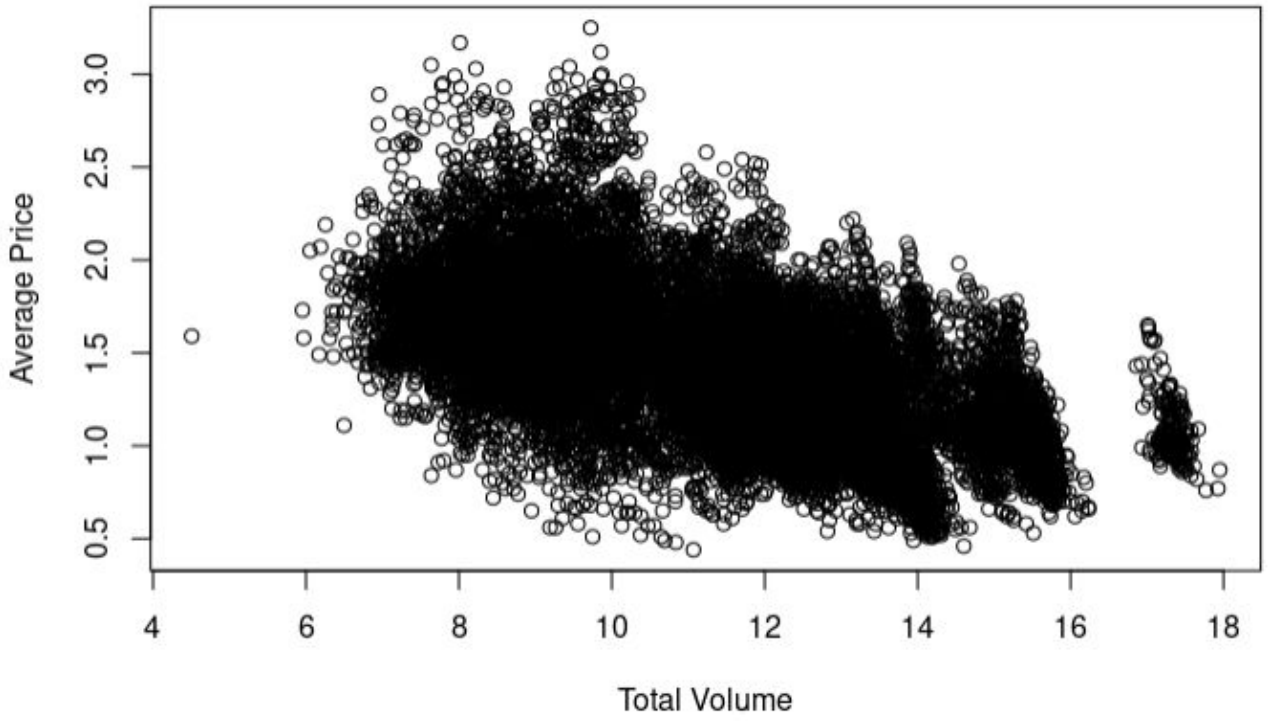
$$\text{Average Price} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Year} + \dots$$

   - Linear fitting learns the coefficients, $\beta_0, \beta_1, \beta_2, \beta_3 \dots$, from data

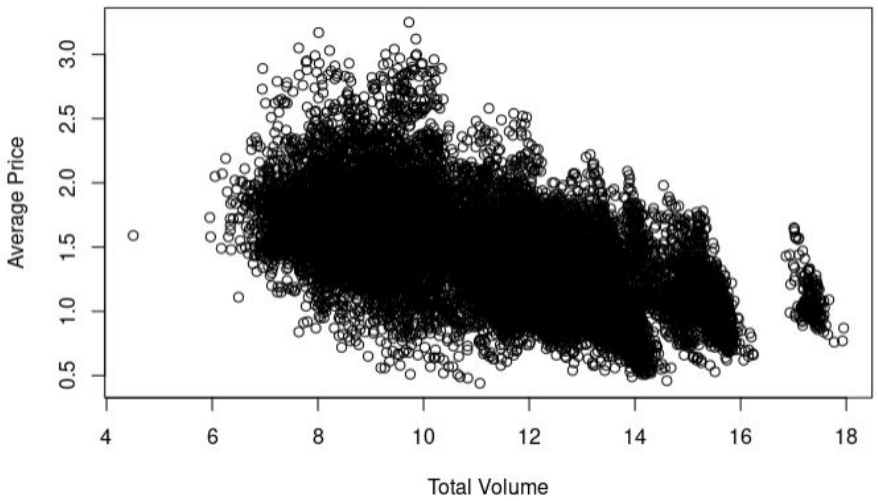
2) I would also like to understand the effects of total volume, type and year on the average price
   - After fitting, we can interpret the coefficients.
   - For example, holding all other variables fixed, average price changes by β_1 when total volume increase by 1.

Linear Model with a constant term

Scatter plot of Average Price vs. Total Volume

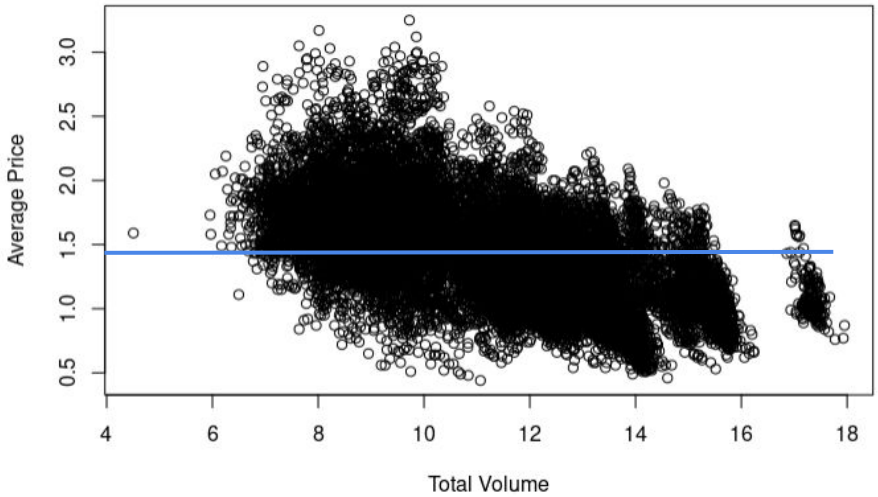


Scatter plot of Average Price vs. Total Volume



- The data looks like a linear of function of total volume
- Let’s fit it to a linear model!

Linear Modeling with a constant term only

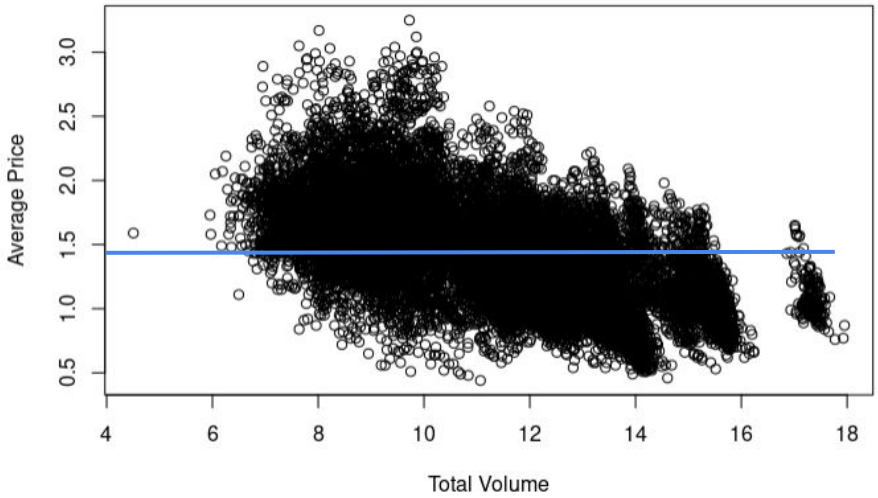


- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

- *As a first pass, let's fit the data to a flat line.*

$$\text{AveragePrice} = \beta_0$$

Linear Modeling with a constant term only



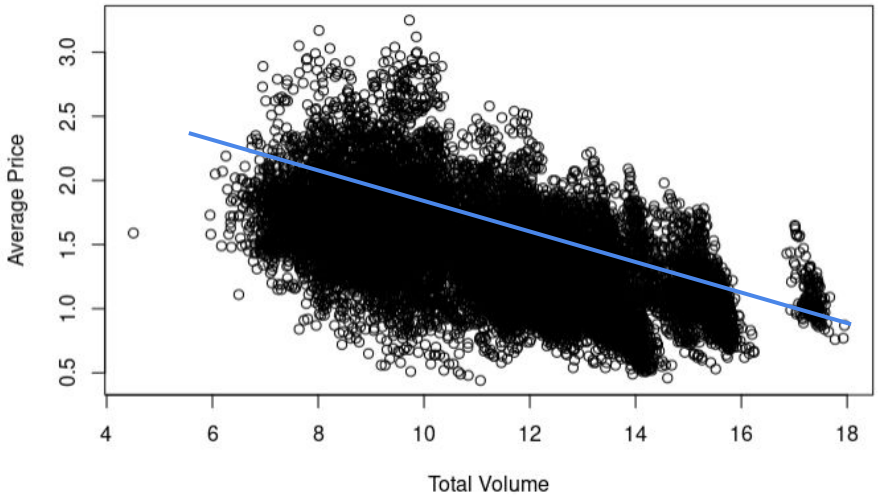
- The data looks like a linear of function of total volume
- Let's fit it to a linear model!
- *As a first pass, let's fit the data to a flat line.*

$$\text{AveragePrice} = \beta_0$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.405978   0.002981   471.7   <2e-16 ***
```

Linear Model with Total Volume

Linear Modeling with Total Volume



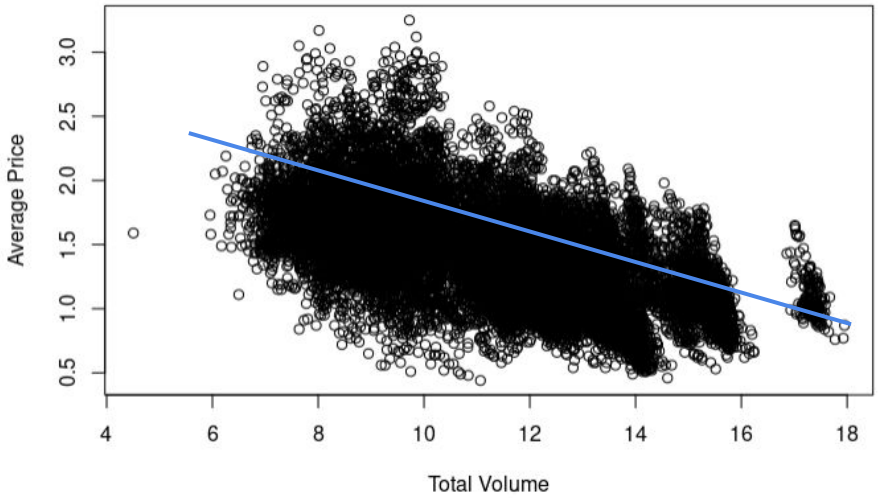
- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

- *Let's have a second go at it. Let's fit AveragePrice to a straight line function of Total Volume.*

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume}$$

Let's do this in R!

Linear Modeling with Total Volume



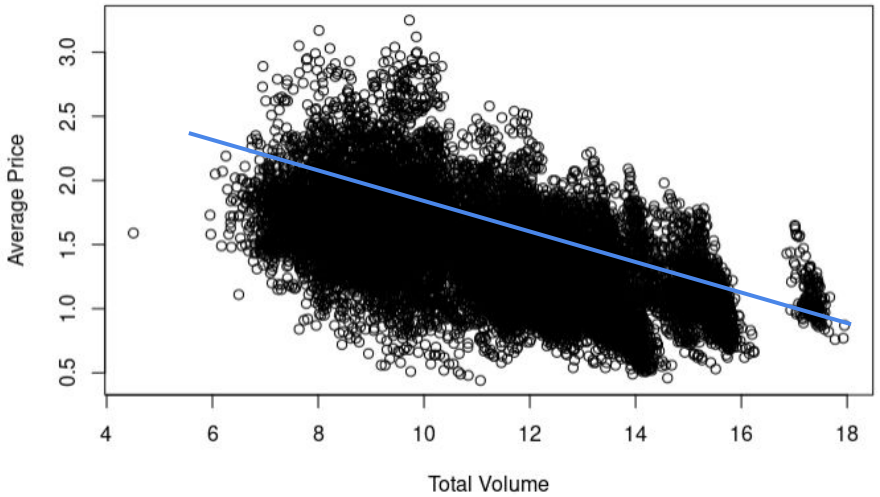
- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

- *Let's have a second go at it. Let's fit AveragePrice to a straight line function of Total Volume.*

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume}$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.565532   0.012193  210.42   <2e-16 ***
TotalVolume -0.102463   0.001056  -97.03   <2e-16 ***
```

Linear Modeling with Total Volume



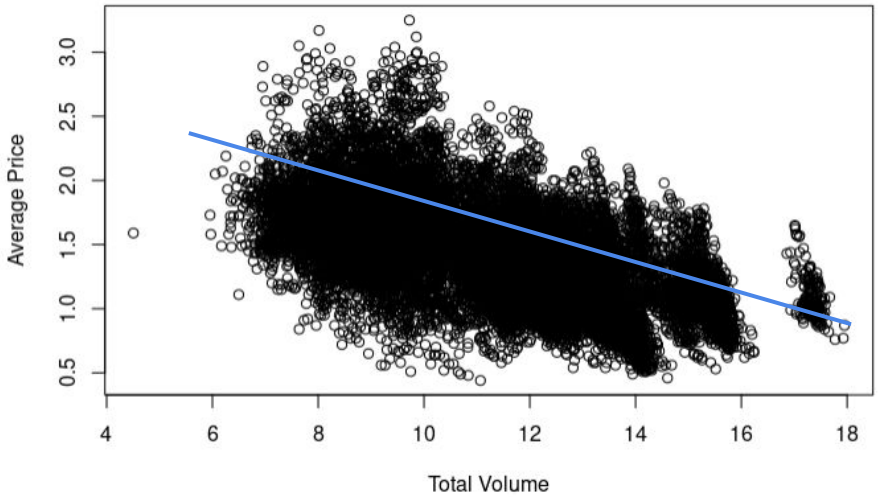
- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

- *Let's have a second go at it. Let's fit AveragePrice to a straight line function of Total Volume.*

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume}$$

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.565532   0.012193  210.42   <2e-16 ***
TotalVolume -0.102463   0.001056  -97.03   <2e-16 ***
```

Linear Modeling with Total Volume



- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

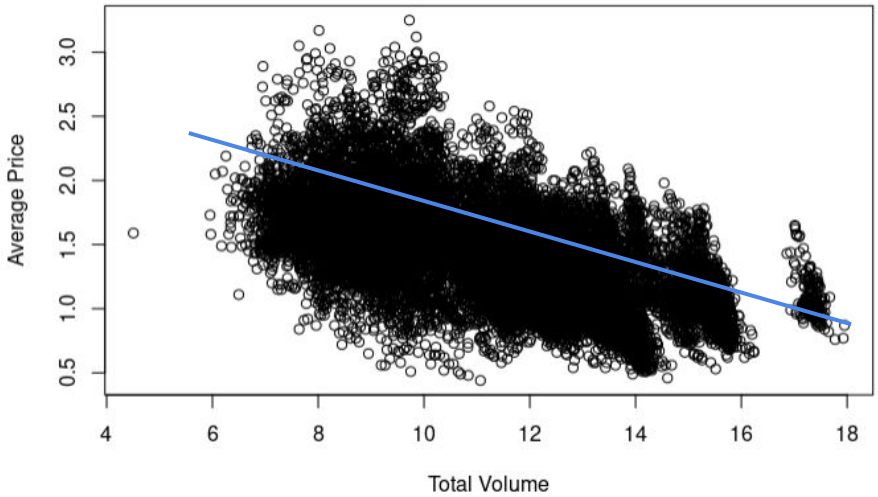
- *Let's have a second go at it. Let's fit AveragePrice to a straight line function of Total Volume.*

$$\text{Average Price} = 2.566 - 0.102 \times \text{Total Volume}$$

We can use the model to:

1) plot the straight line fit

Linear Modeling with Total Volume



- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

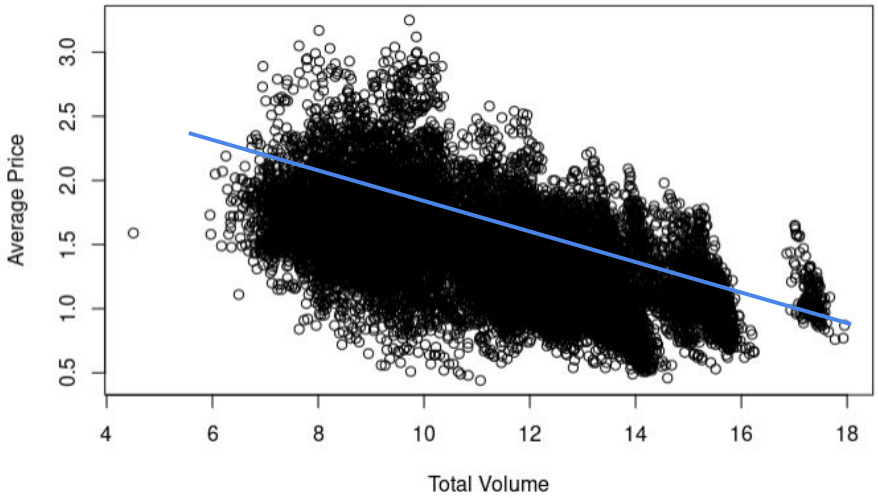
- *Let's have a second go at it. Let's fit AveragePrice to a straight line function of Total Volume.*

$$\text{Average Price} = 2.566 - 0.102 \times \text{Total Volume}$$

We can use the model to:

1) plot the straight line fit
2) predict Average Price values

Linear Modeling with Total Volume



- The data looks like a linear of function of total volume
- Let's fit it to a linear model!

- *Let's have a second go at it. Let's fit AveragePrice to a straight line function of Total Volume.*

$$\text{Average Price} = 2.566 - 0.102 \times \text{Total Volume}$$

We can use the model to:

1) plot the straight line fit
2) predict Average Price values
3) do model comparison using ANOVA

Model Comparison using ANOVA

- Adding any explanatory variable will reduce the sum of squares of the residuals
  - Often, it is important to know if the reduction in sum of squares is meaningful
  - We use ANOVA to tell if the reduction in sum of square is statistically significant

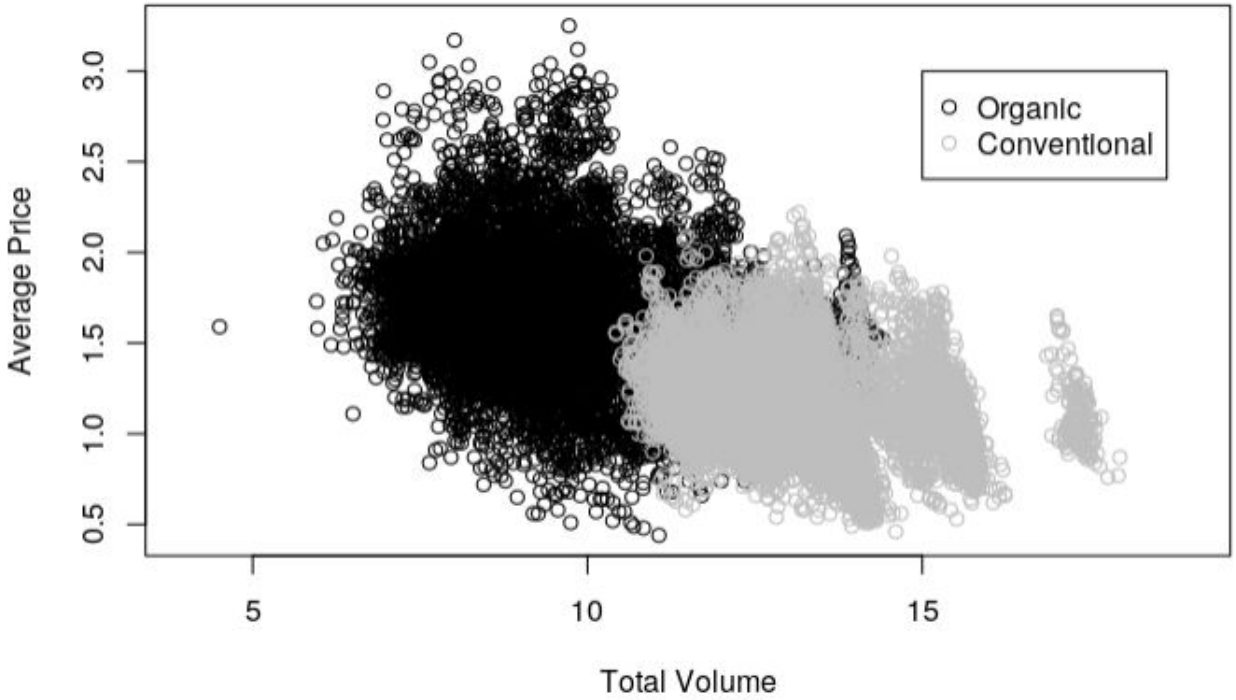
Statistical Output of Linear Models in R

The statistical output of the summary function are close to true if

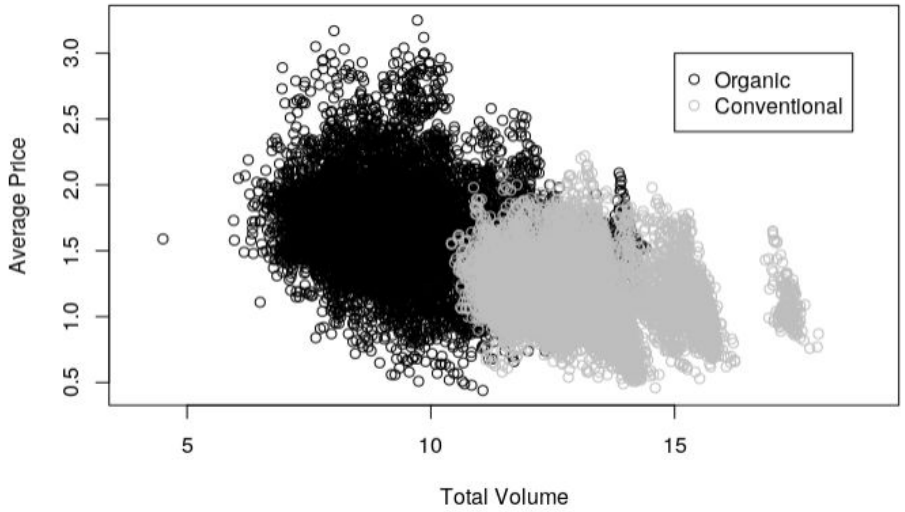
- we have sufficient data
- the response variable is normally distributed
- a linear model describes the average behaviour of the response variable
- each response variable has the same standard deviation, etc.

Linear Model with Total Volume and Type

Scatter plot of Price vs. Total Volume for each type of avocado



Linear Modeling with the categorical variable, Type.

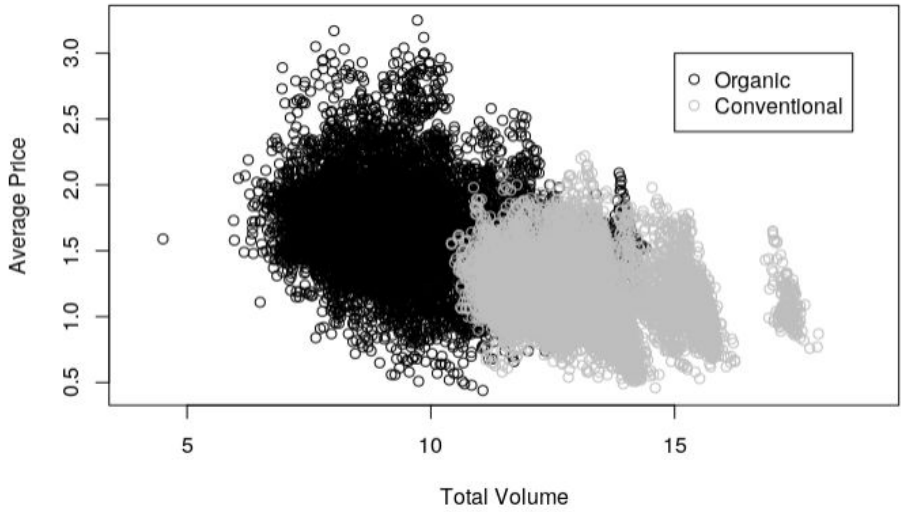


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type}$$

Type = 0 if conventional
Type = 1 if organic

Linear Modeling with the categorical variable, Type.



- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type}$$

- If conventional,

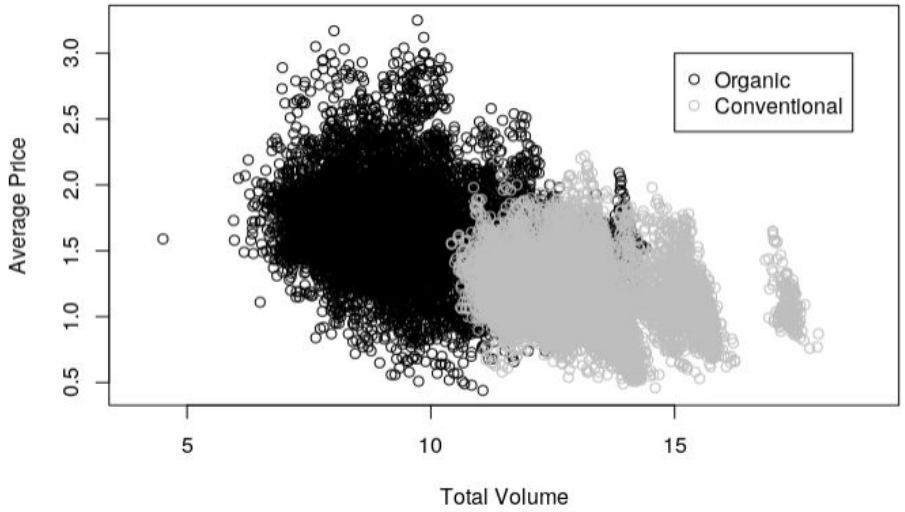
$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume}$$

- If organic,

$$\text{AveragePrice} = \beta_0 + \beta_2 + \beta_1 \text{Total Volume}$$

Let's do this in R!

Linear Modeling with the categorical variable, Type.

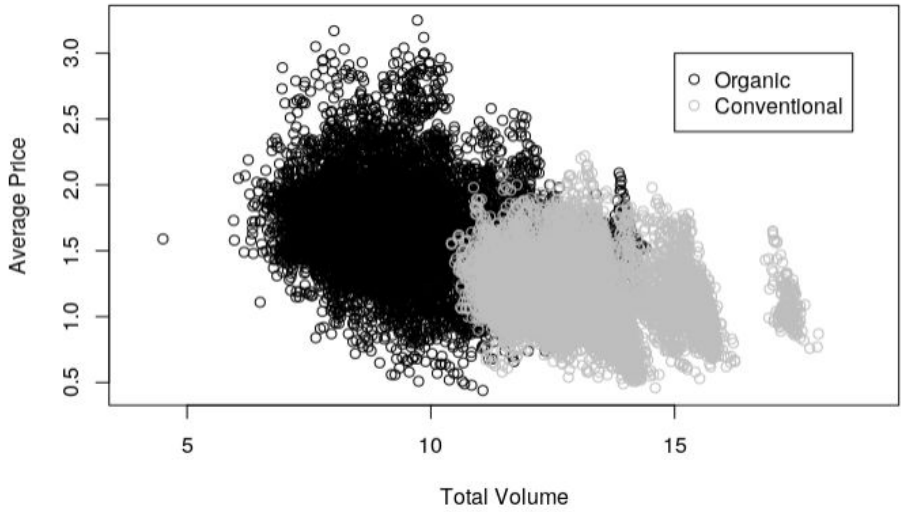


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.744565	0.022081	79.01	<2e-16	***
TotalVolume	-0.044627	0.001662	-26.86	<2e-16	***
Typeorganic	0.332961	0.007620	43.70	<2e-16	***

Linear Modeling with the categorical variable, Type.

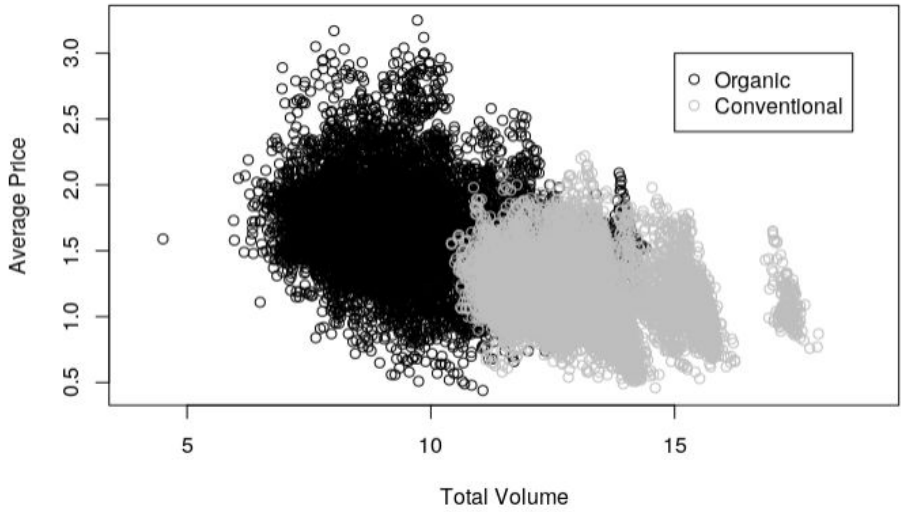


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.744565	0.022081	79.01	<2e-16	***
TotalVolume	-0.044627	0.001662	-26.86	<2e-16	***
Typeorganic	0.332961	0.007620	43.70	<2e-16	***

Linear Modeling with the categorical variable, Type.

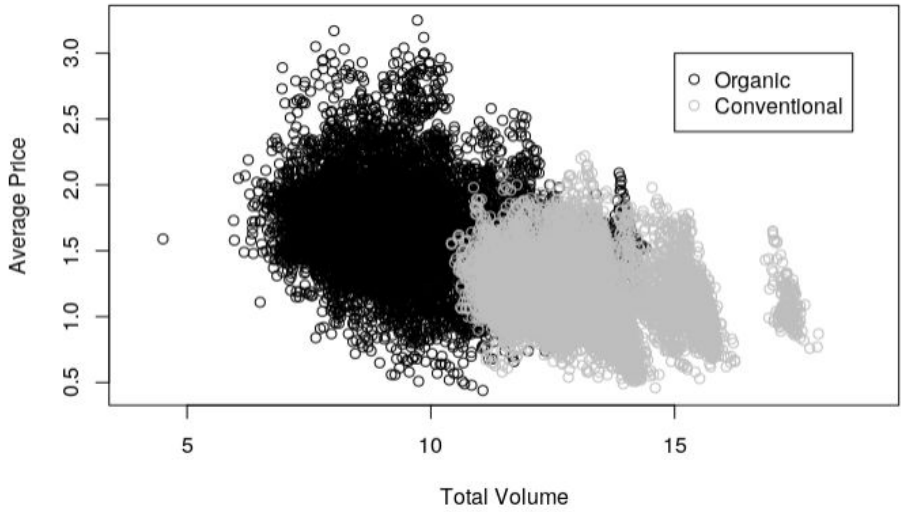


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type}$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.744565   0.022081   79.01   <2e-16 ***
TotalVolume -0.044627   0.001662  -26.86   <2e-16 ***
Typeorganic  0.332961   0.007620   43.70   <2e-16 ***
```

Linear Modeling with the categorical variable, Type.



- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{Average Price} = 1.744 - 0.044 \times \text{Total Volume} + 0.33 \times \text{Type}$$

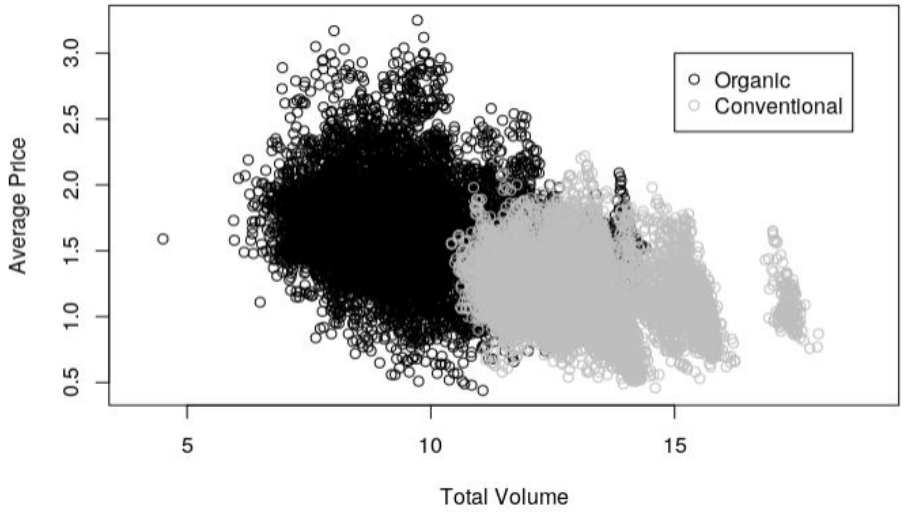
- If conventional,

$$\text{Average Price} = 1.744 - 0.044 \times \text{Total Volume}$$

- If organic,

$$\text{Average Price} = 2.074 - 0.044 \times \text{Total Volume}$$

Linear Modeling with the categorical variable, Type.



- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{Average Price} = 1.744 - 0.044 \times \text{Total Volume} + 0.33 \times \text{Type}$$

- If conventional,

$$\text{Average Price} = 1.744 - 0.044 \times \text{Total Volume}$$

- If organic,

$$\text{Average Price} = 2.074 - 0.044 \times \text{Total Volume}$$

We can use the model to:

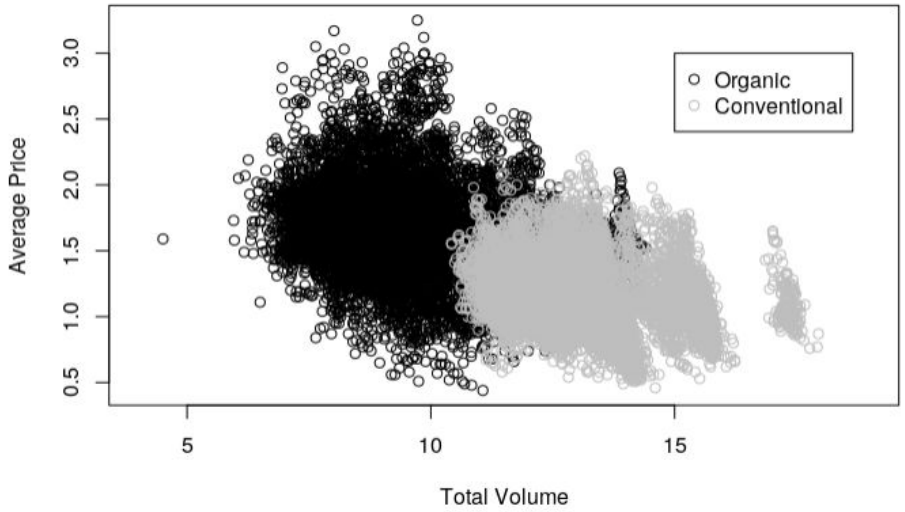
1) plot the straight line fit
2) predict Average Price values
3) do model comparison using ANOVA

Linear Modeling with the categorical variable, Type.

- The previous model assumes that curves have the same slopes but are shifted up/down from each other
- What if it’s possible for the curves to have different slopes? How do we include this in our model?

Linear Model with Total Volume and Type interaction

Linear Modeling with an interaction term

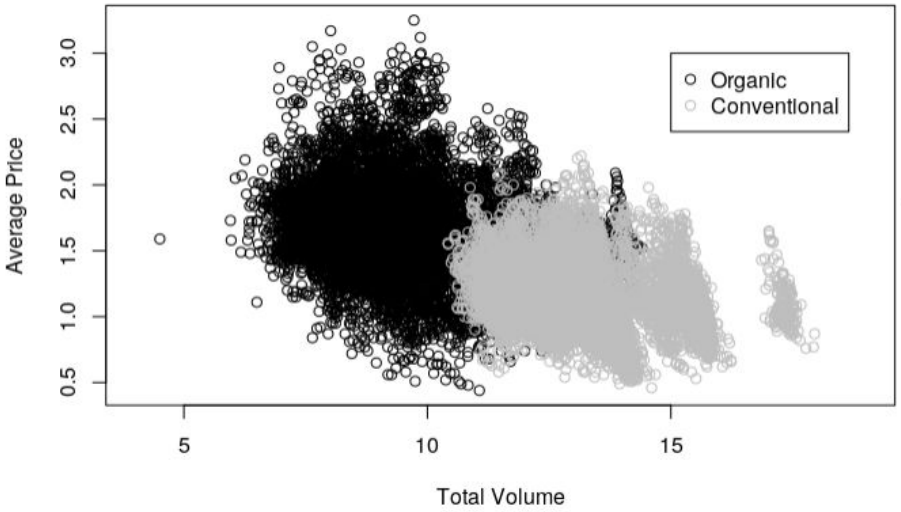


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Total Volume} \times \text{Type}$$

Type = 0 if conventional
Type = 1 if organic

Linear Modeling with an interaction term



- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Total Volume} \times \text{Type}$$

- If conventional,

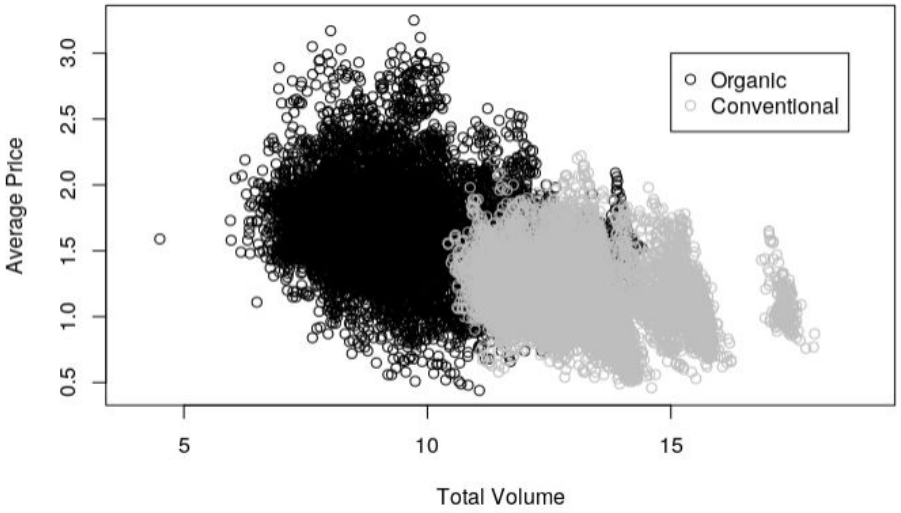
$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume}$$

- If organic,

$$\text{AveragePrice} = \beta_0 + \beta_2 + (\beta_1 + \beta_3) \times \text{Total Volume}$$

Let's do this in R!

Linear Modeling with an interaction term

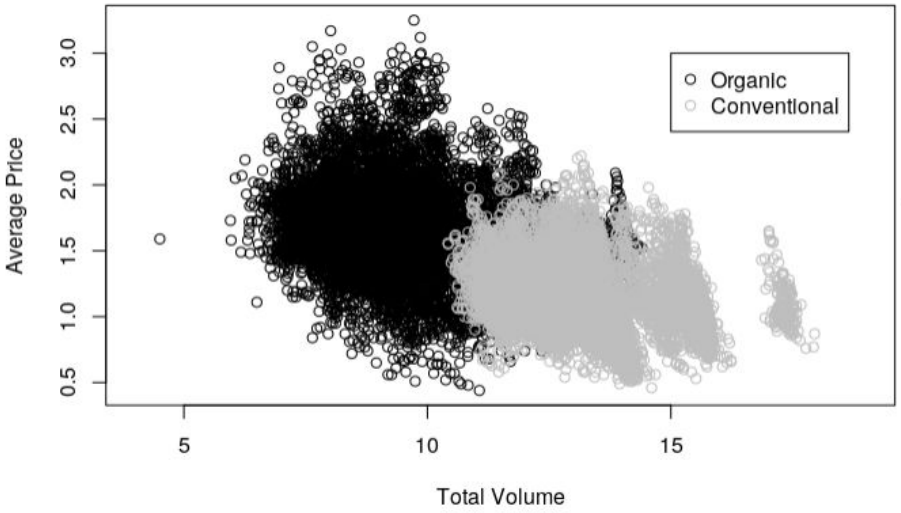


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Total Volume} \times \text{Type}$$

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.753568   0.032210  54.441  < 2e-16
TotalVolume              -0.045312   0.002438 -18.584  < 2e-16
Typeorganic               0.318321   0.038891   8.185  2.9e-16
TotalVolume:Typeorganic   0.001279   0.003332   0.384    0.701
```

Linear Modeling with an interaction term

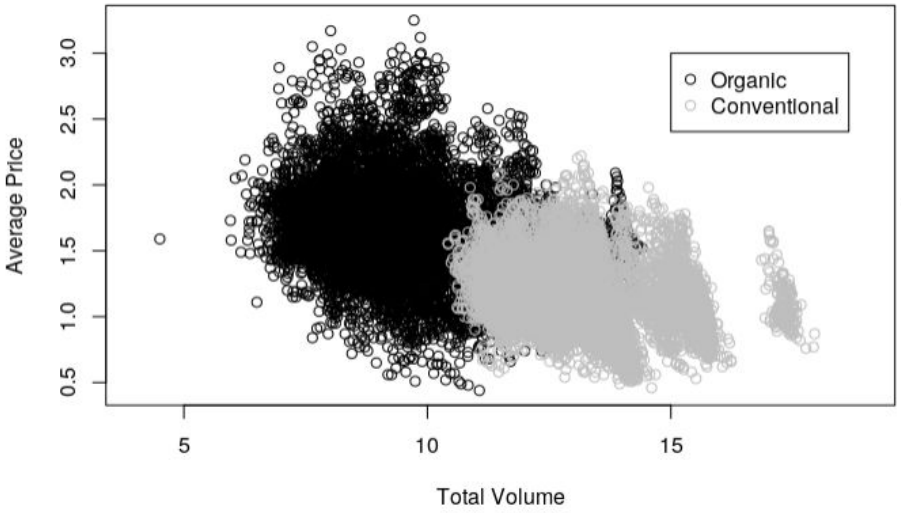


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Total Volume} \times \text{Type}$$

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.753568   0.032210  54.441  < 2e-16
TotalVolume          -0.045312   0.002438 -18.584  < 2e-16
Typeorganic           0.318321   0.038891   8.185  2.9e-16
TotalVolume:Typeorganic 0.001279   0.003332   0.384    0.701
```

Linear Modeling with an interaction term



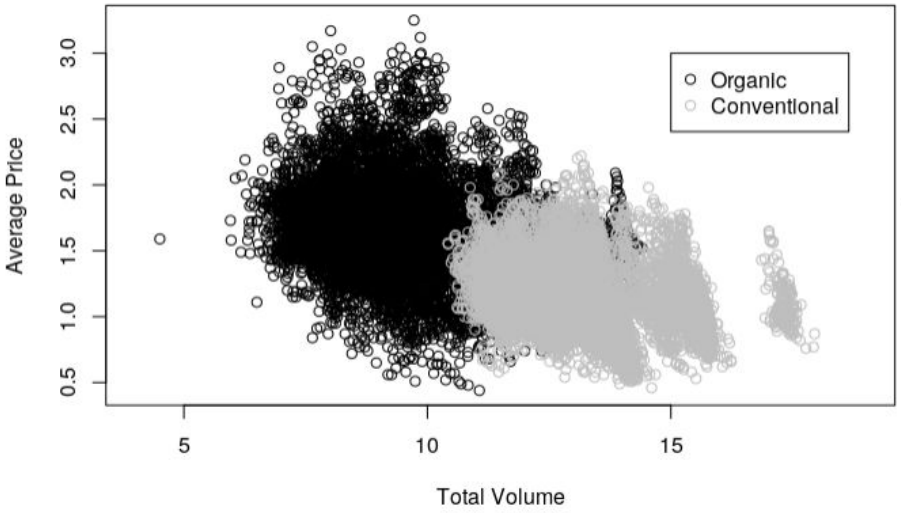
- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Total Volume} \times \text{Type}$$

```
Coefficients:

                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.753568   0.032210  54.441  < 2e-16
TotalVolume              -0.045312   0.002438 -18.584  < 2e-16
Typeorganic               0.318321   0.038891   8.185  2.9e-16
TotalVolume:Typeorganic   0.001279   0.003332   0.384    0.701
```

Linear Modeling with an interaction term

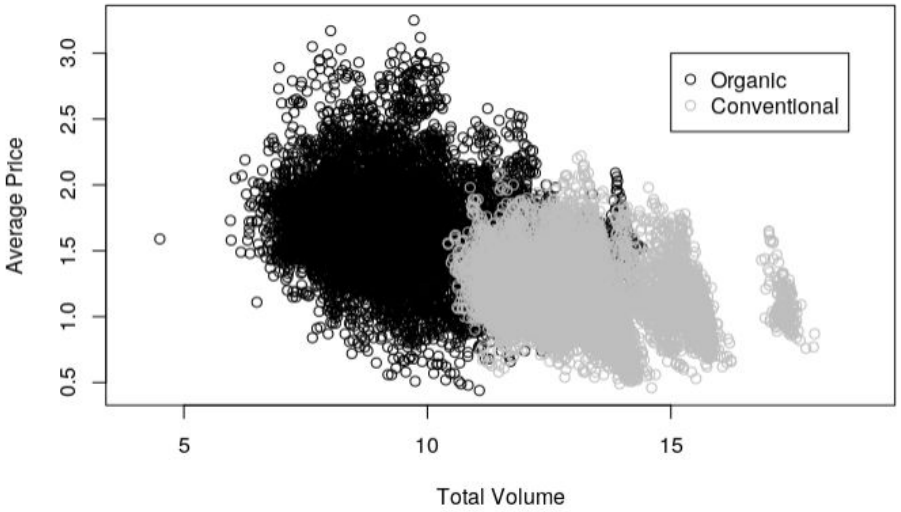


- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Type} + \beta_3 \text{Total Volume} \times \text{Type}$$

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.753568   0.032210  54.441  < 2e-16
TotalVolume                -0.045312   0.002438 -18.584  < 2e-16
Typeorganic                 0.318321   0.038891   8.185  2.9e-16
TotalVolume:Typeorganic     0.001279   0.003332   0.384    0.701
```

Linear Modeling with an interaction term



- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{Average Price} = 1.753 - 0.045 \times \text{Total Volume} + 0.32 \times \text{Type} + 0.0012 \times \text{Total Volume} \times \text{Type}$$

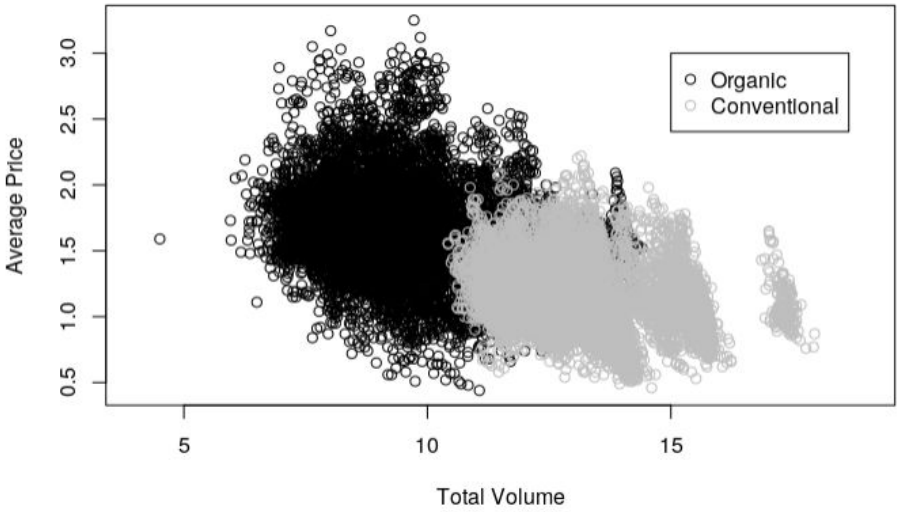
- If conventional,

$$\text{Average Price} = 1.75 - 0.045 \times \text{Total Volume}$$

- If organic,

$$\text{Average Price} = 2.07 - 0.046 \times \text{Total Volume}$$

Linear Modeling with an interaction term



- Each type of avocado appear to follow its own linear model
- How do we write this mathematically?

$$\text{Average Price} = 1.753 - 0.045 \times \text{Total Volume} + 0.32 \times \text{Type} + 0.0012 \times \text{Total Volume} \times \text{Type}$$

- If conventional,

$$\text{Average Price} = 1.75 - 0.045 \times \text{Total Volume}$$

- If organic,

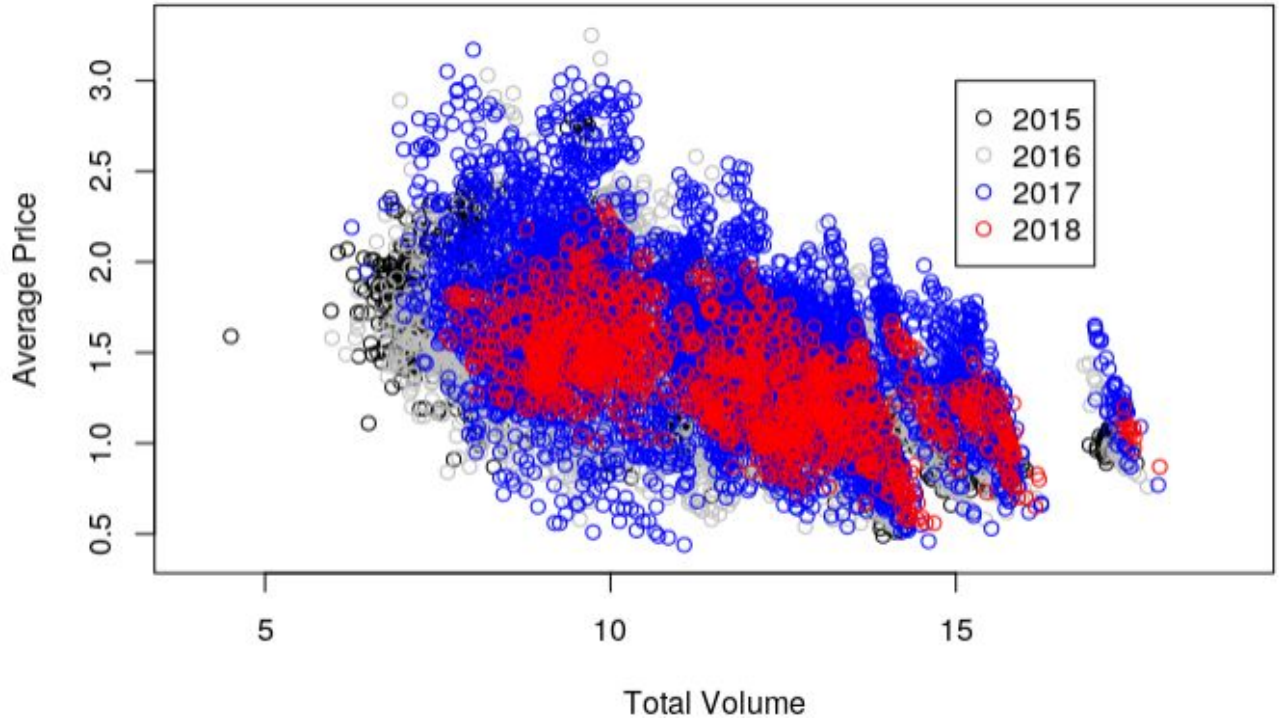
$$\text{Average Price} = 2.07 - 0.046 \times \text{Total Volume}$$

We can use the model to:

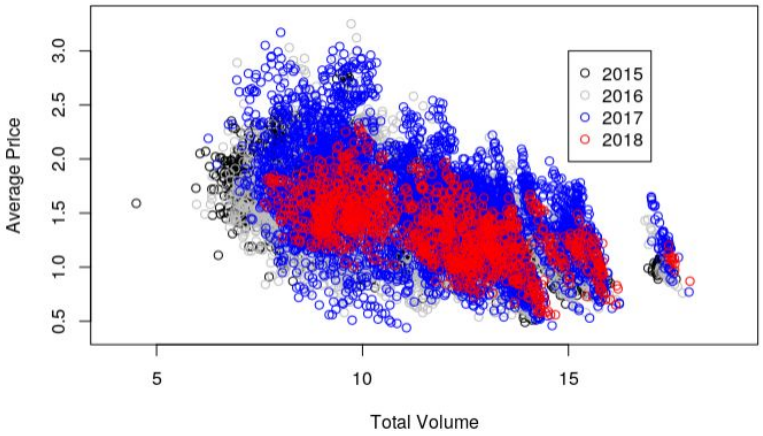
1) plot the straight line fit
2) predict Average Price values
3) do model comparison using ANOVA

Linear Model with Total Volume and Year

Linear Modeling with the categorical variable, Year.



Linear Modeling with the categorical variable, Year.

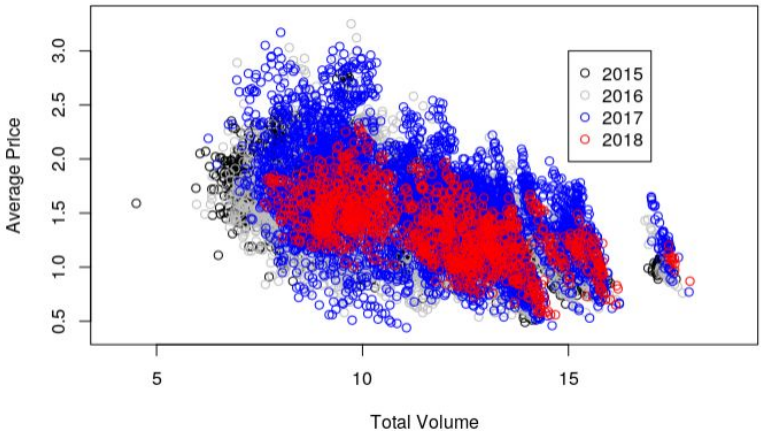


- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

Year2016 = 1 if Year = 2016
Year2016 = 0 if Year is not 2016

Linear Modeling with the categorical variable, Year.



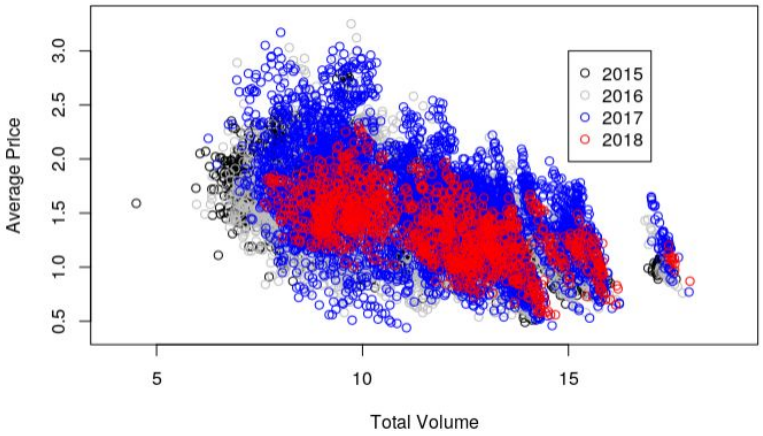
- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

Year2017 = 1 if Year = 2017

Year2017 = 0 if Year is not 2017

Linear Modeling with the categorical variable, Year.



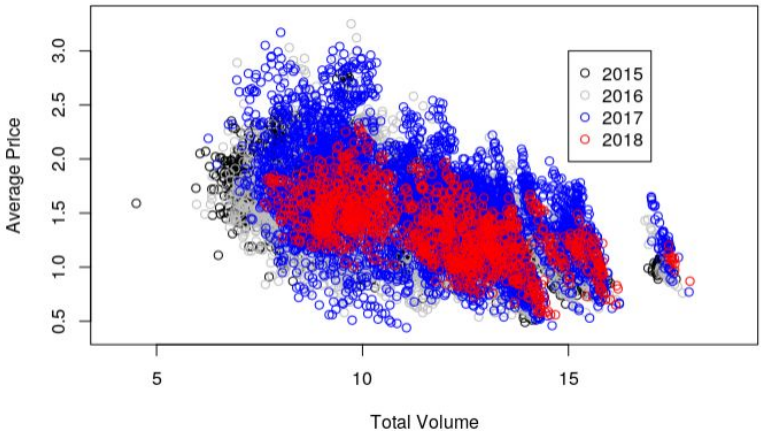
- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

Year2018 = 1 if Year = 2018

Year2018 = 0 if Year is not 2018

Linear Modeling with the categorical variable, Year.



- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

- If year is 2015,

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume}$$

- If year is 2016,

$$\text{AveragePrice} = \beta_0 + \beta_2 + \beta_1 \text{Total Volume}$$

- If year is 2017,

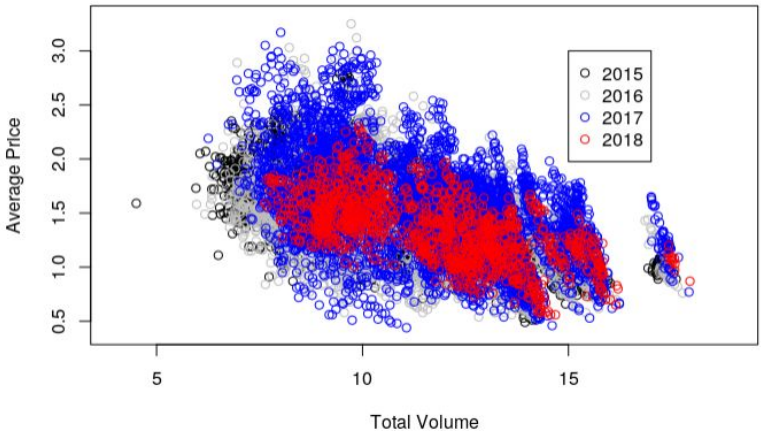
$$\text{AveragePrice} = \beta_0 + \beta_3 + \beta_1 \text{Total Volume}$$

- If year is 2018,

$$\text{AveragePrice} = \beta_0 + \beta_4 + \beta_1 \text{Total Volume}$$

Let's do this in R!

Linear Modeling with the categorical variable, Year.

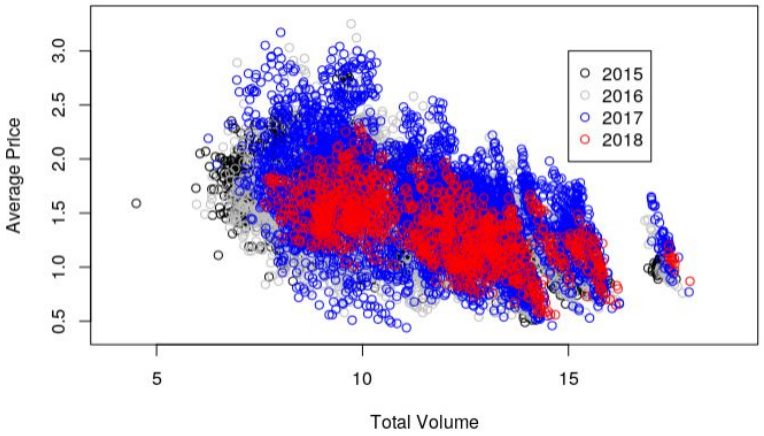


- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.529083	0.012072	209.497	< 2e-16	***
TotalVolume	-0.104349	0.001023	-101.963	< 2e-16	***
Year2016	-0.008084	0.005966	-1.355	0.175	
Year2017	0.182901	0.005947	30.757	< 2e-16	***
Year2018	0.041271	0.009755	4.231	2.34e-05	***

Linear Modeling with the categorical variable, Year.

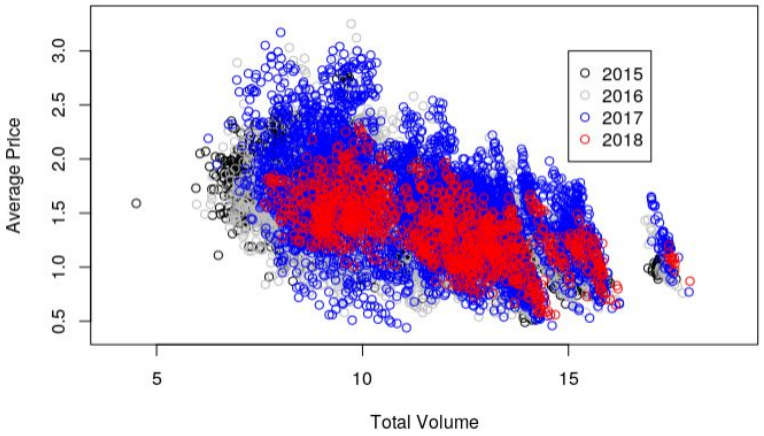


- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.529083	0.012072	209.497	< 2e-16	***
TotalVolume	-0.104349	0.001023	-101.963	< 2e-16	***
Year2016	-0.008084	0.005966	-1.355	0.175	
Year2017	0.182901	0.005947	30.757	< 2e-16	***
Year2018	0.041271	0.009755	4.231	2.34e-05	***

Linear Modeling with the categorical variable, Year.

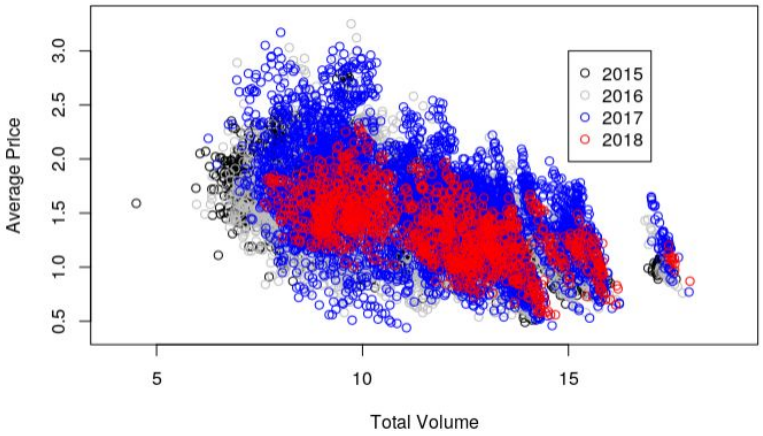


- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.529083	0.012072	209.497	< 2e-16	***
TotalVolume	-0.104349	0.001023	-101.963	< 2e-16	***
Year2016	-0.008084	0.005966	-1.355	0.175	
Year2017	0.182901	0.005947	30.757	< 2e-16	***
Year2018	0.041271	0.009755	4.231	2.34e-05	***

Linear Modeling with the categorical variable, Year.

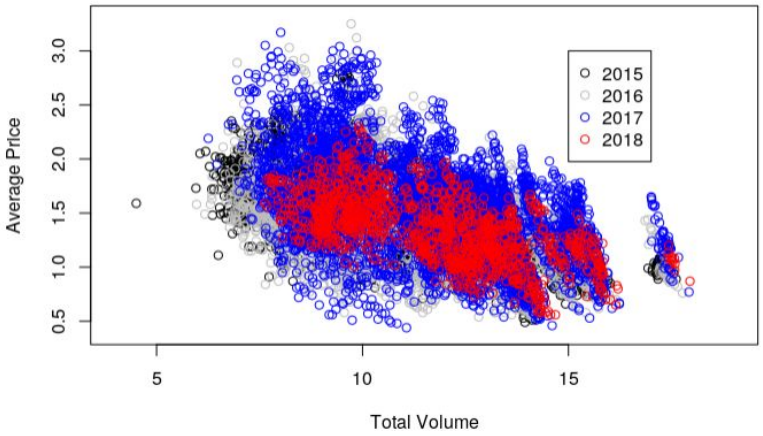


- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.529083	0.012072	209.497	< 2e-16	***
TotalVolume	-0.104349	0.001023	-101.963	< 2e-16	***
Year2016	-0.008084	0.005966	-1.355	0.175	
Year2017	0.182901	0.005947	30.757	< 2e-16	***
Year2018	0.041271	0.009755	4.231	2.34e-05	***

Linear Modeling with the categorical variable, Year.

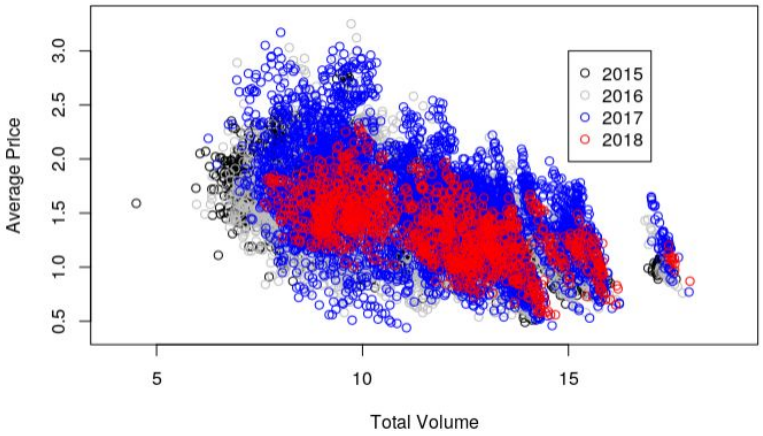


- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{AveragePrice} = \beta_0 + \beta_1 \text{Total Volume} + \beta_2 \text{Year2016} + \beta_3 \text{Year2017} + \beta_4 \text{Year2018}$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.529083	0.012072	209.497	< 2e-16	***
TotalVolume	-0.104349	0.001023	-101.963	< 2e-16	***
Year2016	-0.008084	0.005966	-1.355	0.175	
Year2017	0.182901	0.005947	30.757	< 2e-16	***
Year2018	0.041271	0.009755	4.231	2.34e-05	***

Linear Modeling with the categorical variable, Year.



- Each recording year of avocado appear to follow their own linear model
- How do we write this mathematically?

$$\text{Average Price} = 2.52 - 0.104 \times \text{Total Volume} - 0.008 \times \text{Year2016} + 0.18 \times \text{Year2017} + 0.041 \times \text{Year2018}$$

Conclusion and Next Steps

- We found reasonable linear models of average price using total volume, year and type.
- Going further, you might want to consider
  - Model selection, eg. LASSO, ridge regression, -- workshop on this in the summer
  - Goodness of fit measures: AIC, BIC, R^2, adjusted R^2
  - Statistical tests for goodness of fit
  - Checking MLE conditions

Exercises

1. Open the file linear_model_exercises.Rmd
2. Get cracking!