Logistic Regression 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:
    - PriceCategory - whether each average avocado sample is ‘Expensive’ or ‘Cheap’
    - 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

Goals

1) Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted

2) From the probabilities, predict whether a sample is expensive or cheap

3) Understand the effects of total volume, type and year on the probability that a sample is expensive or cheap

Linear Regression vs. Logistic Regression

Why Linear Models are not appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
  - That is, convert binary data to probability scores. Linear regression cannot do this.

Why Logistic Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
    - That is, convert binary data to probability scores. Linear regression cannot do this.

- By definition, logistic models convert binary data to probability scores.

Why Logistic Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
    - That is, convert binary data to probability scores. Linear regression cannot do this.

- By definition, logistic models convert binary data to probability scores

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times (\text{Explanatory Variable 1}) + \beta_2 \times (\text{Explanatory Variable 2}) + \ldots$$

where P is the probability being expensive.

Why Logistic Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
    - That is, convert binary data to probability scores. Linear regression cannot do this.

- By definition, logistic models convert binary data to probability scores

$$\log \boxed{\frac{P}{1-P}} = \beta_0 + \beta_1 \times (\text{Explanatory Variable 1}) + \beta_2 \times (\text{Explanatory Variable 2}) + \ldots$$

where P is the probability being expensive.

Odds is the ratio of the probability of being expensive to the probability of being cheap

- Odd > 1 implies the probability of being expensive is higher
- Odd < 1 implies the probability of being cheap is higher

Why Logistic Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
    - That is, convert binary data to probability scores. Linear regression cannot do this.

- By definition, logistic models convert binary data to probability scores

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times (\text{Explanatory Variable 1}) + \beta_2 \times (\text{Explanatory Variable 2}) + \ldots$$

where P is the probability being expensive.

- Goal 3: Understand the effects of total volume, type and year on the probability that a sample is expensive or cheap
    - After fitting, we can interpret the coefficients.
    - For example, holding all other variables fixed, the log odds changes by β_1 when Explanatory Variable 1 increases by 1.

Why Logistic Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
  - That is, convert binary data to probability scores. Linear regression cannot do this.

- By definition, logistic models convert binary data to probability scores.

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times (\text{Explanatory Variable 1}) + \beta_2 \times (\text{Explanatory Variable 2}) + \ldots$$

where P is the probability being expensive.

- With a few mathematical tricks, the model can be converted in terms of P.

Why Logistic Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
  - That is, convert binary data to probability scores. Linear regression cannot do this.

- By logistic models, convert binary data to probability scores

$$P = \frac{e^{\beta_0+\beta_1\times(\text{Explanatory Variable 1})+\beta_2\times(\text{Explanatory Variable 2})+\ldots}}{1+e^{\beta_0+\beta_1\times(\text{Explanatory Variable 1})+\beta_2\times(\text{Explanatory Variable 2})+\ldots}}$$

where P is the probability being expensive.

Why Linear Models are appropriate

- Goal 1: Predict the probability of a sample being “expensive” or “cheap” based on its type of avocado, total volume sold, year in which the sample was conducted
  - That is, convert binary data to probability scores. Linear regression cannot do this.

- By definition, logistic models convert binary data to probability scores

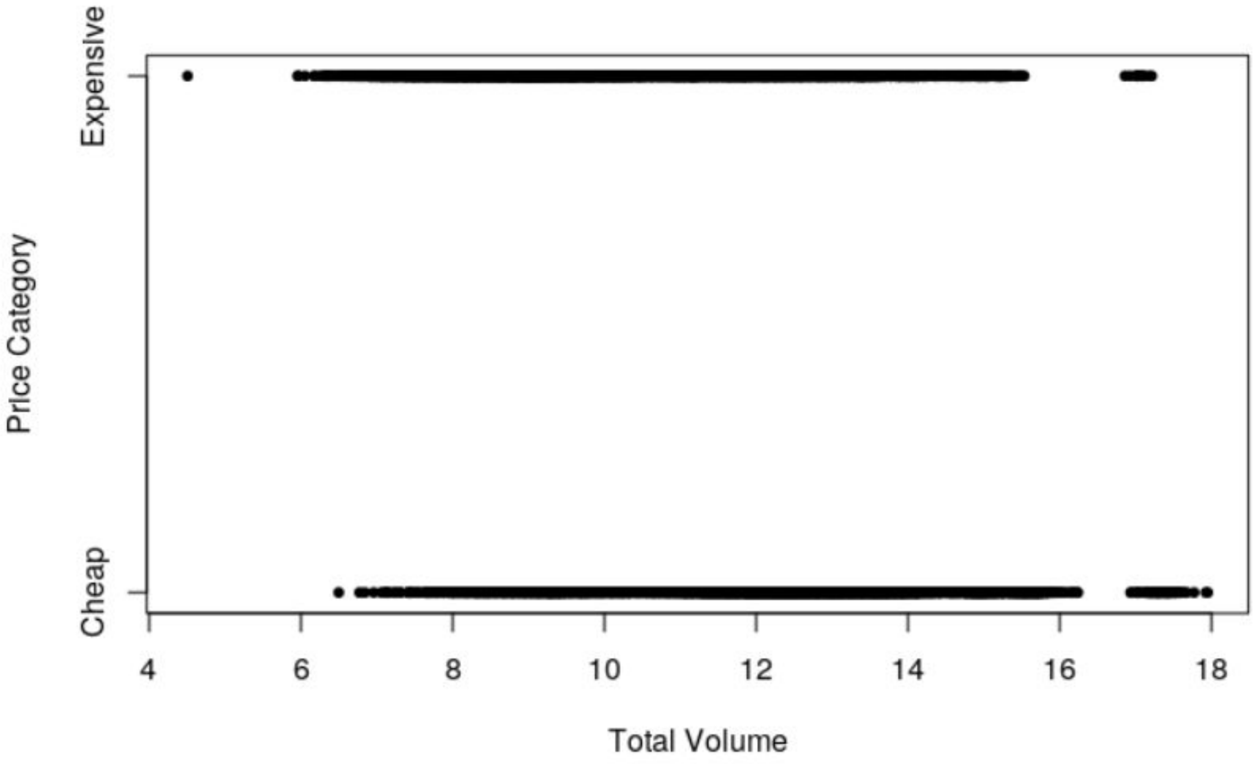
$$P = \frac{e^{\beta_0 + \beta_1 \times (\text{Explanatory Variable 1}) + \beta_2 \times (\text{Explanatory Variable 2}) + \ldots}}{1 + e^{\beta_0 + \beta_1 \times (\text{Explanatory Variable 1}) + \beta_2 \times (\text{Explanatory Variable 2}) + \ldots}}$$

where P is the probability being expensive.

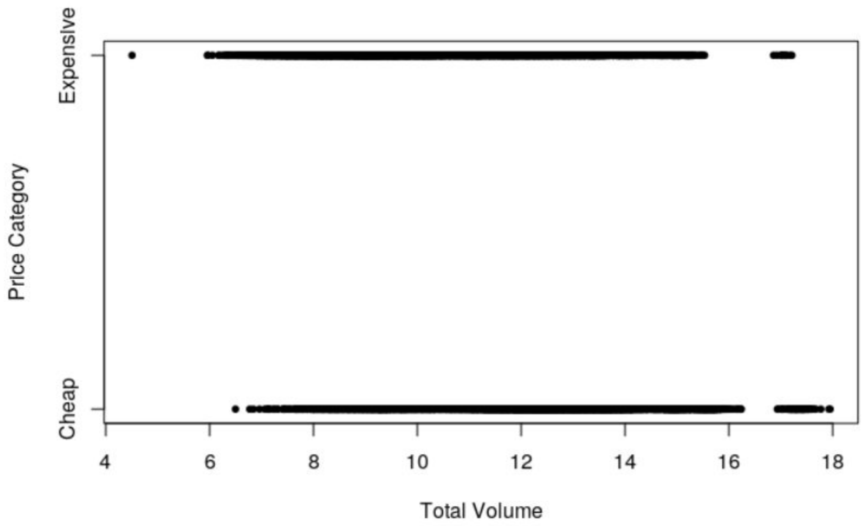
- Goal 2: From the probabilities, predict whether a sample is expensive or cheap

Logistic Model with Total Volume

Scatter plot of Price Category vs. Total Volume



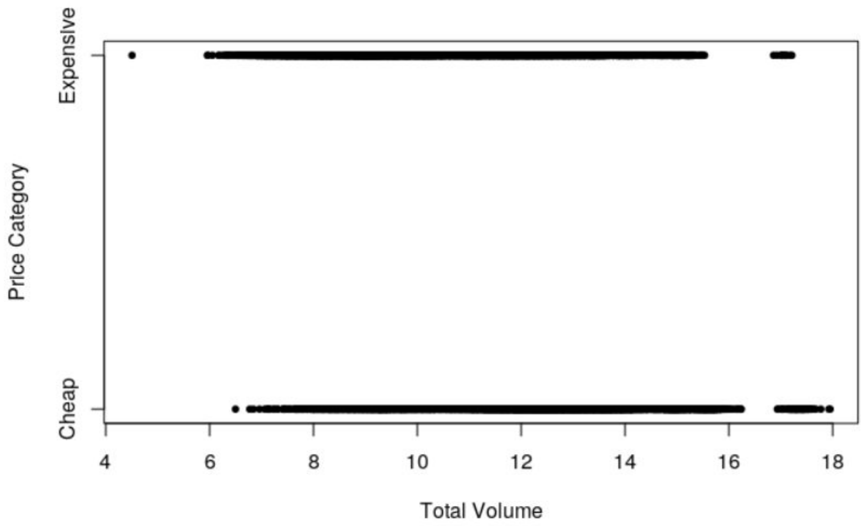
Scatter plot of Average Price vs. Total Volume



- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

Scatter plot of Average Price vs. Total Volume

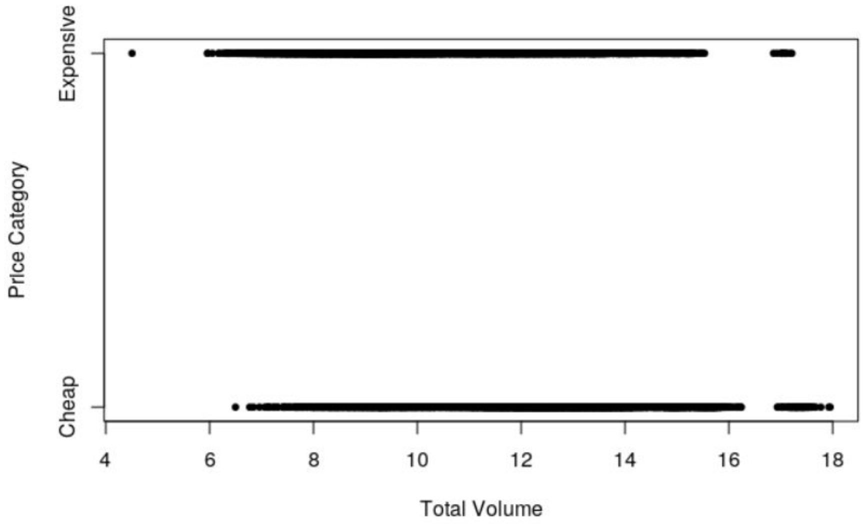


- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let’s use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume}$$

Scatter plot of Average Price vs. Total Volume



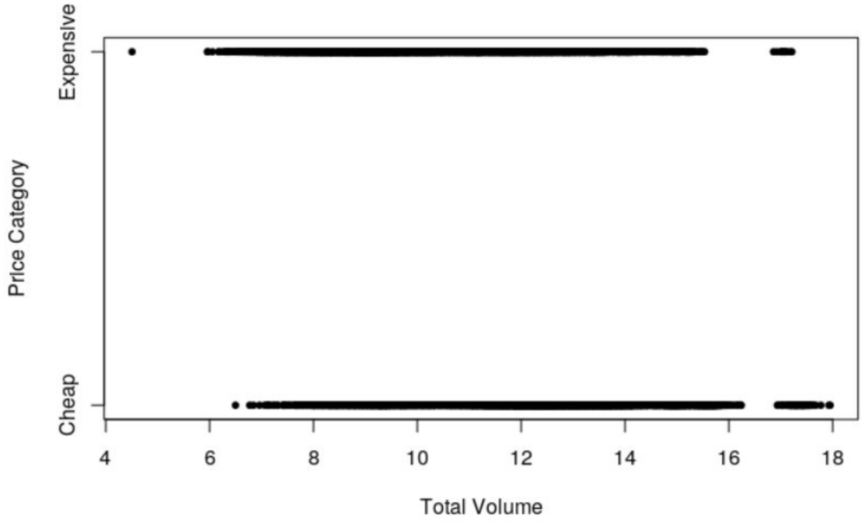
- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume}$$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.961950   0.109664   63.48   <2e-16 ***
TotalVolume -0.617017   0.009562  -64.53   <2e-16 ***
```

Scatter plot of Average Price vs. Total Volume



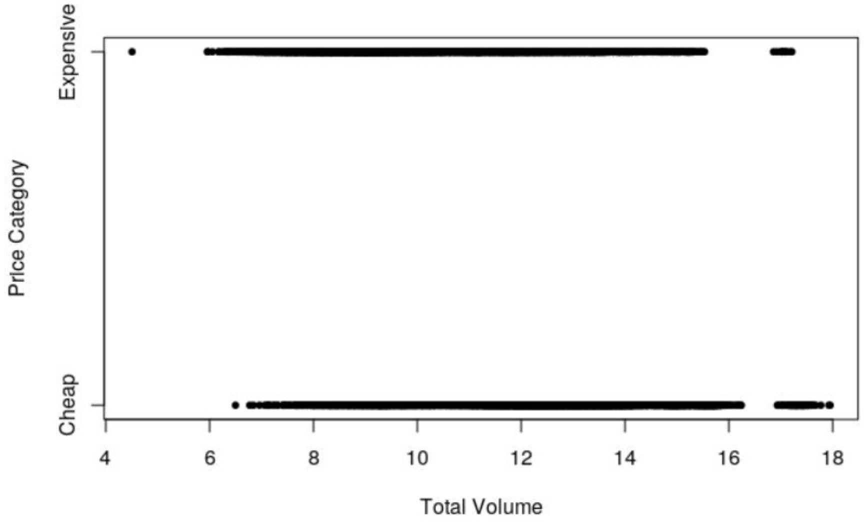
- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let’s use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume}$$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.961950   0.109664   63.48   <2e-16 ***
TotalVolume -0.617017   0.009562  -64.53   <2e-16 ***
```

Scatter plot of Average Price vs. Total Volume



- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = 6.96 - 0.617 \times \text{Total Volume}$$

Deviance

- For a linear model, deviance is sum of squares of the residuals
- Deviance is a more generalized “sum of squares of the residuals” for GLMs, like logistic models and Poisson models
    - Significant reduction of deviance is important
    - Deviance allows us to compare nested models

Predict Function

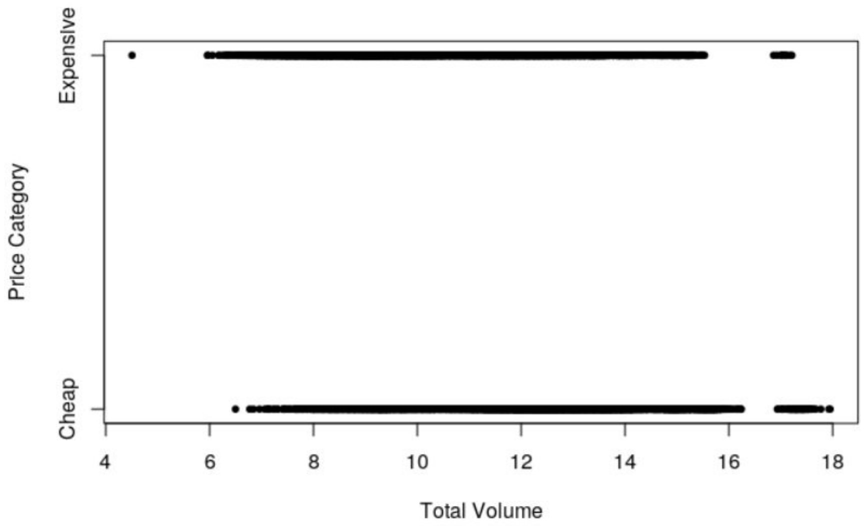
- We can use the “predict” to determine the probability of any dataset with the same terms as our model
- “predict” returns the predicted log-odds

$$\text{log odds} = \log \frac{P}{1-P}$$

- To get the probability, we need to simple transformation

$$P = \frac{e^{\text{log odds}}}{1 + e^{\text{log odds}}}$$

Predicting Classes



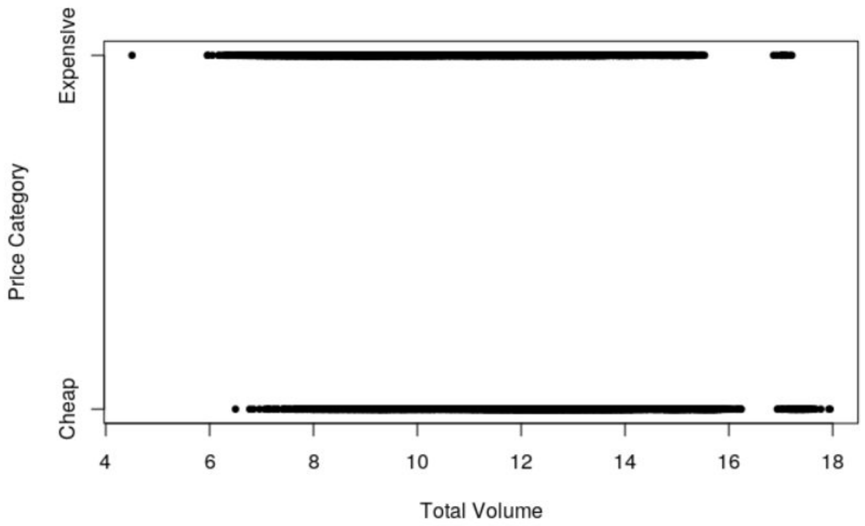
- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = 6.96 - 0.617 \times \text{Total Volume}$$

- How do we get the probabilities of the fitted data?
- How do we use the model to predict classes?

Predicting Class Probabilities



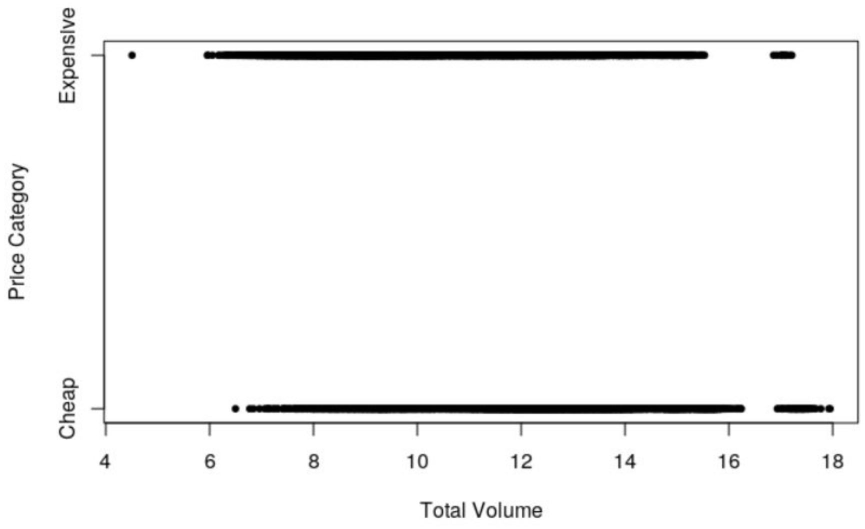
- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = 6.96 - 0.617 \times \text{Total Volume}$$

- How do we get the probabilities of the fitted data?
- How do we use the model to predict classes?
  - We have decide on a threshold probability.

Predicting Class Probabilities



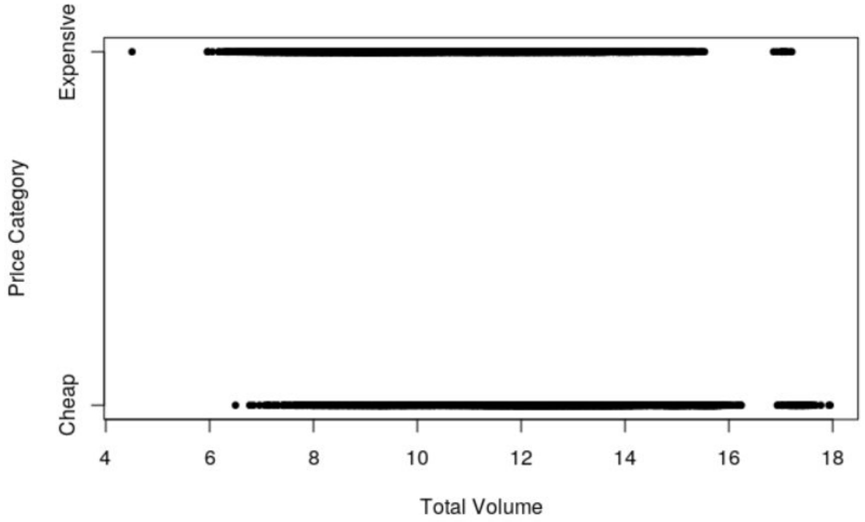
- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = 6.96 - 0.617 \times \text{Total Volume}$$

- How do we get the probabilities of the fitted data?
- How do we use the model to predict classes?
  - We have decide on a threshold probability.
  - If P >= 0.5, then the sample is expensive.

Predicting Classes



- Samples with lower volume are likely to be expensive
- Samples with higher volume are likely to be cheap

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = 6.96 - 0.617 \times \text{Total Volume}$$

- How do we get the probabilities of the fitted data?
- How do we use the model to predict classes?
  - We have decide on a threshold probability.
  - If P >= 0.5, then sample is expensive.
  - If P < 0.5, then the sample is cheap.

Confusion Matrix

	Predicted Class 0	Predicted Class 1
Actual Class 0		
Actual Class 1		

Confusion Matrix

	Predicted Class 0	Predicted Class 1
Actual Class 0	*Total number of correctly predicted cheap avocados*	
Actual Class 1		

Confusion Matrix

	Predicted Class 0	Predicted Class 1
Actual Class 0	Total number of correctly predicted cheap avocados	
Actual Class 1		*Total number of correctly predicted expensive avocados*

Confusion Matrix

	<u>Predicted Class 0</u>	<u>Predicted Class 1</u>
<u>Actual Class 0</u>	Total number of correctly predicted cheap avocados	*Total number of cheap avocados incorrectly predicted as expensive*
<u>Actual Class 1</u>		Total number of correctly predicted expensive avocados

Confusion Matrix

	<u>Predicted Class 0</u>	<u>Predicted Class 1</u>
<u>Actual Class 0</u>	Total number of correctly predicted cheap avocados	Total number of cheap avocados incorrectly predicted as expensive
<u>Actual Class 1</u>	*Total number of expensive avocados incorrectly predicted as cheap*	Total number of correctly predicted expensive avocados

Logistic Model with only constant term

Logistic Model with constant term

- Let’s use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0$$

Logistic Model with constant term

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0$$

- Model assigns constant probability to each class. Why would we do this?

Logistic Model with constant term

- Let’s use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0$$

- Model assigns constant probability to each class. Why would we do this?
  - This is known as the null model. It is worst possible model since we are always going to make a class prediction error
    - Let’s say we have two samples: one is expensive and the other is cheap.
    - The null model will assign both the same probability score so any threshold decision will make at least one error.

Logistic Model with constant term

- Let's use a logistic model to predict the probability of being expensive using total volume sold

$$\log \frac{P}{1-P} = \beta_0$$

- Model assigns constant probability to each class. Why would we do this?
    - This is known as the null model. It is worst possible model since we are always going to make a class prediction error
    - We can run an ANOVA to see if other models significantly reduce the deviance relative to the null model

Logistic Model with constant term

- Let's use a logistic model to predict the probability of being expensive using total volume sold

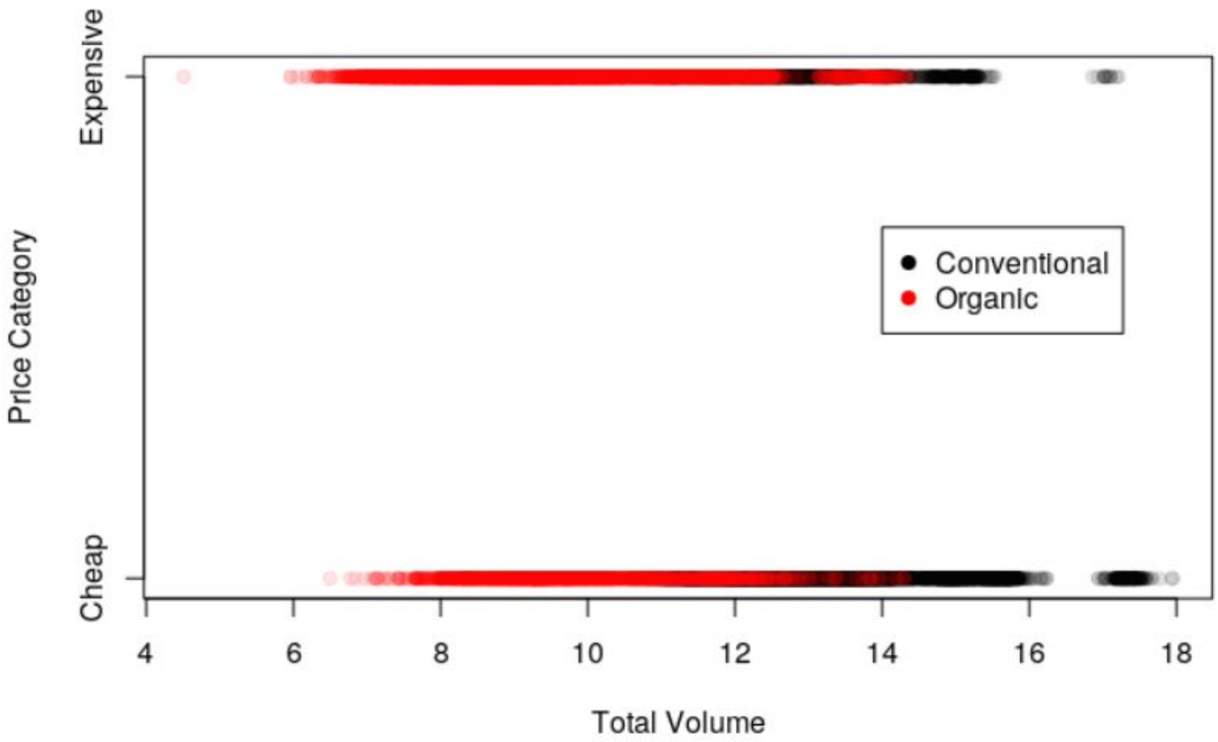
$$\log \frac{P}{1-P} = \beta_0$$

	Estimate	Std. Error	z value
(Intercept)	-0.01304	0.01481	-0.881

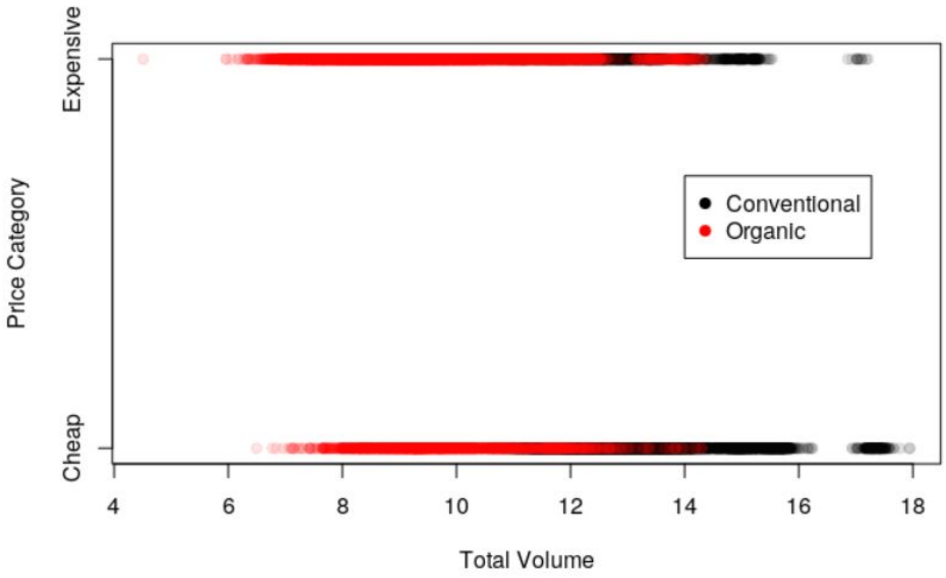
- Model assigns constant probability to each class. Why would we do this?
  - This is known as the null model. It worst possible model since we are always going to make a class prediction error
  - We can run an ANOVA to see if other models significantly reduce the deviance relative to the null model

Logistic Model with Total Volume and Type

Scatter plot of Price Category vs. Total Volume colored by Type

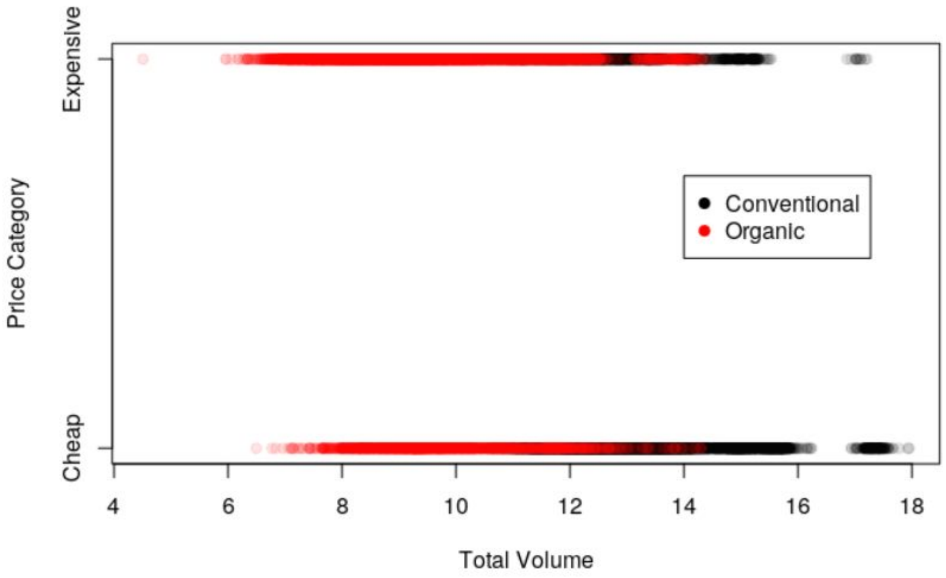


Logistic Model with Total Volume and Type



- The type variable adds more information about which samples are “cheap” and “expensive”
- If organic and lower volume sold, then the avocado is likely to be expensive
- If conventional and high volume sold, the the avocado is likely to be cheap

Logistic Model with Total Volume and Type



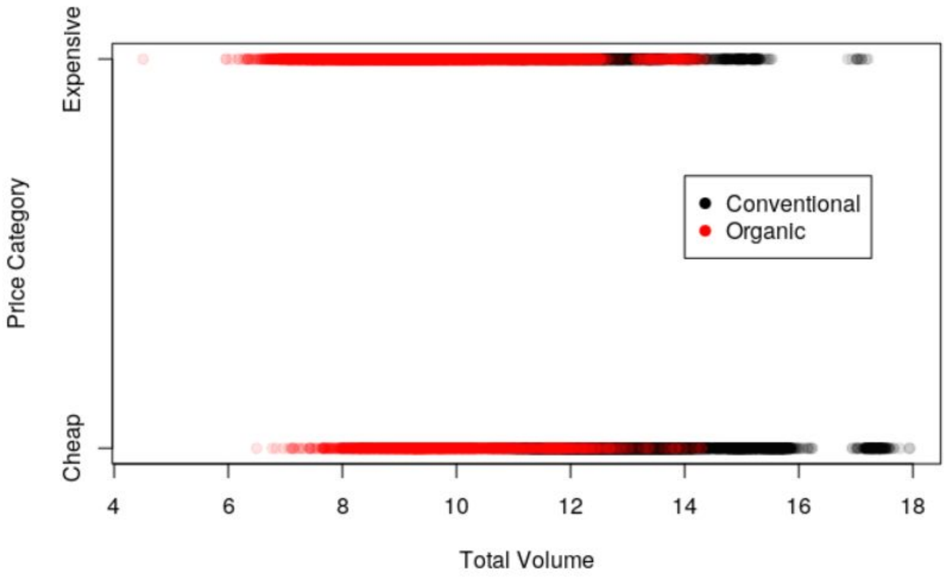
- The type variable adds more information about which samples are “cheap” and “expensive”
- If organic and lower volume sold, then the avocado is likely to be expensive
- If conventional and high volume sold, the the avocado is likely to be cheap

- Let’s use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume} + \beta_2 \times \text{Type}$$

Type = 0 if conventional
Type = 1 if organic

Logistic Model with Total Volume and Type



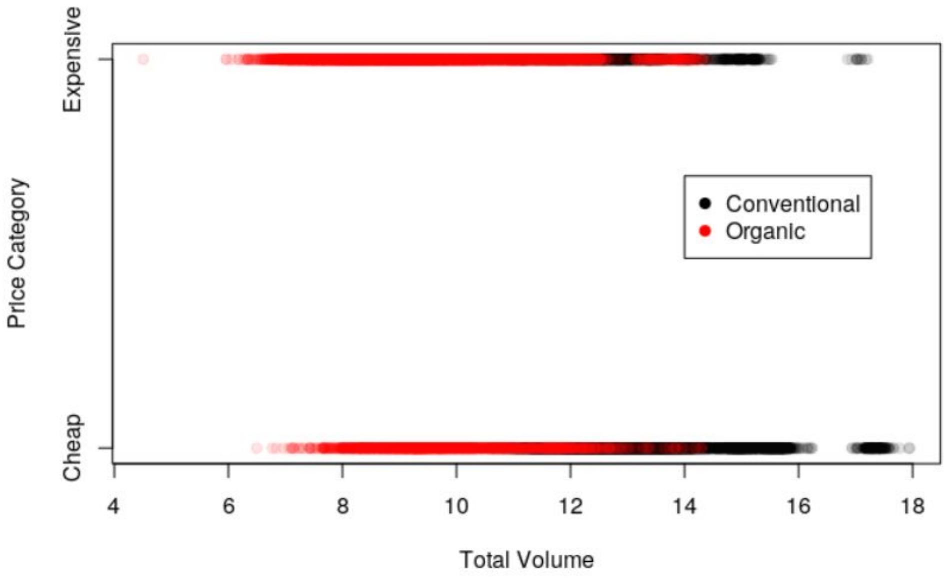
- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume} + \beta_2 \times \text{Type}$$

- If conventional avocado,

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume}$$

Logistic Model with Total Volume and Type



- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume} + \beta_2 \times \text{Type}$$

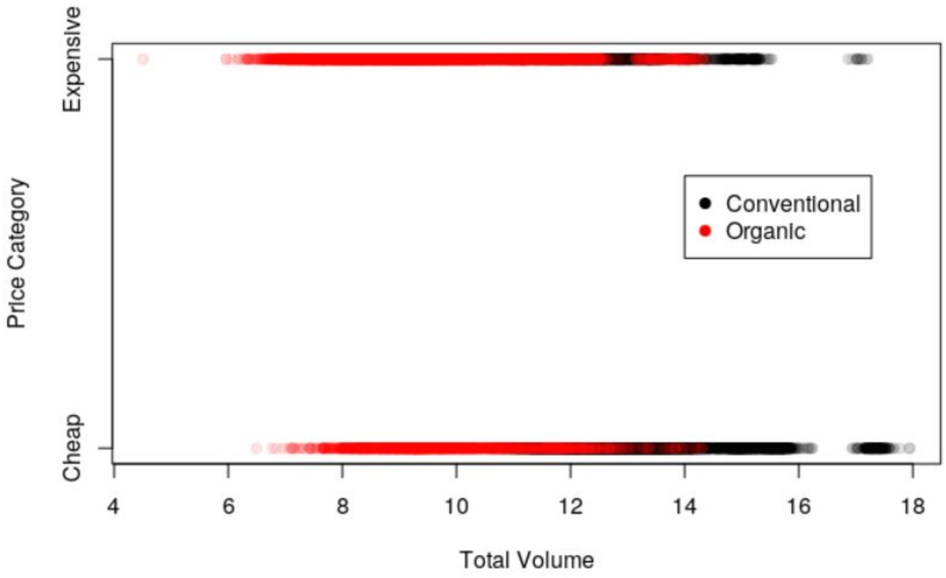
- If conventional avocado,

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume}$$

- If organic avocado,

$$\log \frac{P}{1-P} = \beta_0 + \beta_2 + \beta_1 \times \text{Total Volume}$$

Logistic Model with Total Volume and Type

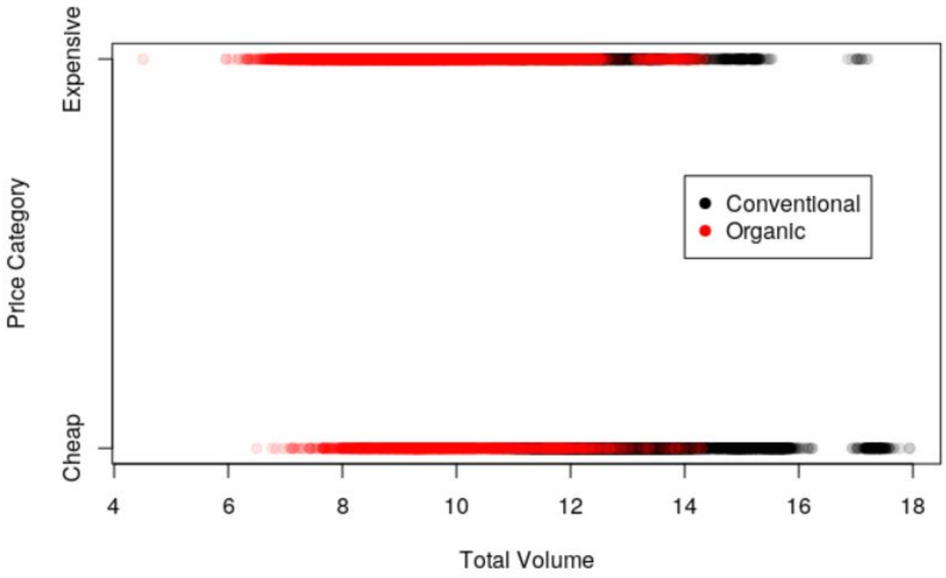


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume} + \beta_2 \times \text{Type}$$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.83471    0.17872   10.27   <2e-16 ***
TotalVolume  -0.24828    0.01371  -18.10   <2e-16 ***
Typeorganic   1.91438    0.05662   33.81   <2e-16 ***
```

Logistic Model with Total Volume and Type

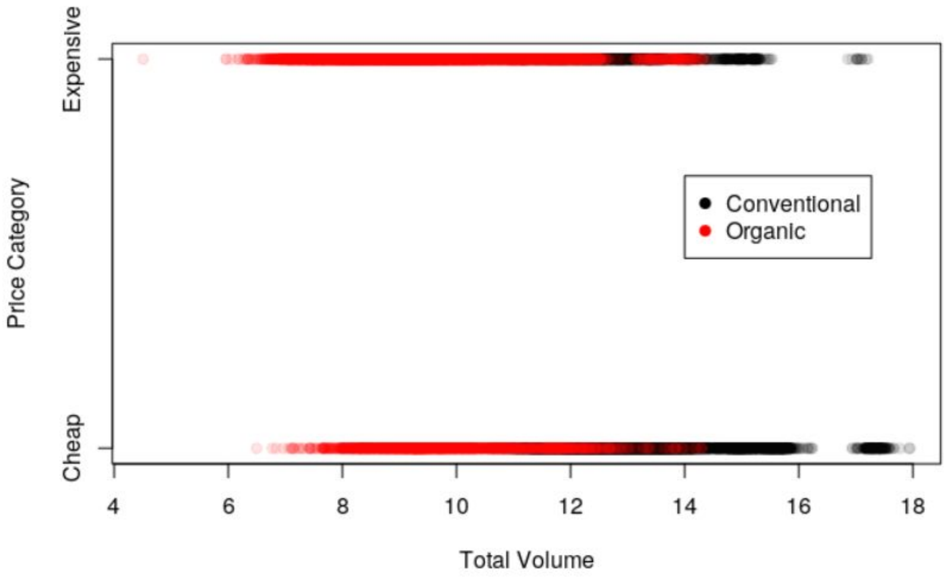


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume} + \beta_2 \times \text{Type}$$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.83471    0.17872   10.27   <2e-16 ***
TotalVolume -0.24828    0.01371  -18.10   <2e-16 ***
Typeorganic  1.91438    0.05662   33.81   <2e-16 ***
```

Logistic Model with Total Volume and Type

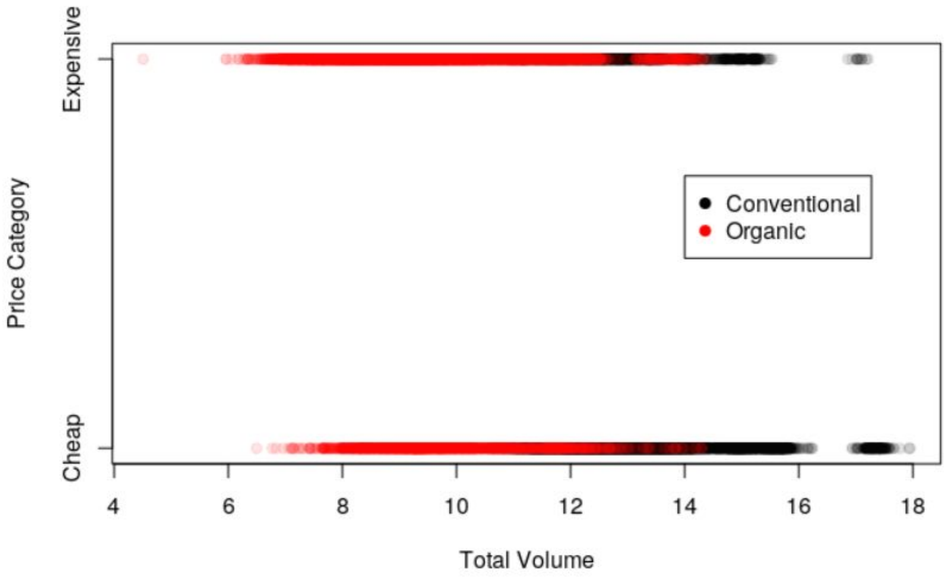


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 \times \text{Total Volume} + \beta_2 \times \text{Type}$$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.83471    0.17872   10.27   <2e-16 ***
TotalVolume -0.24828    0.01371  -18.10   <2e-16 ***
Typeorganic  1.91438    0.05662   33.81   <2e-16 ***
```

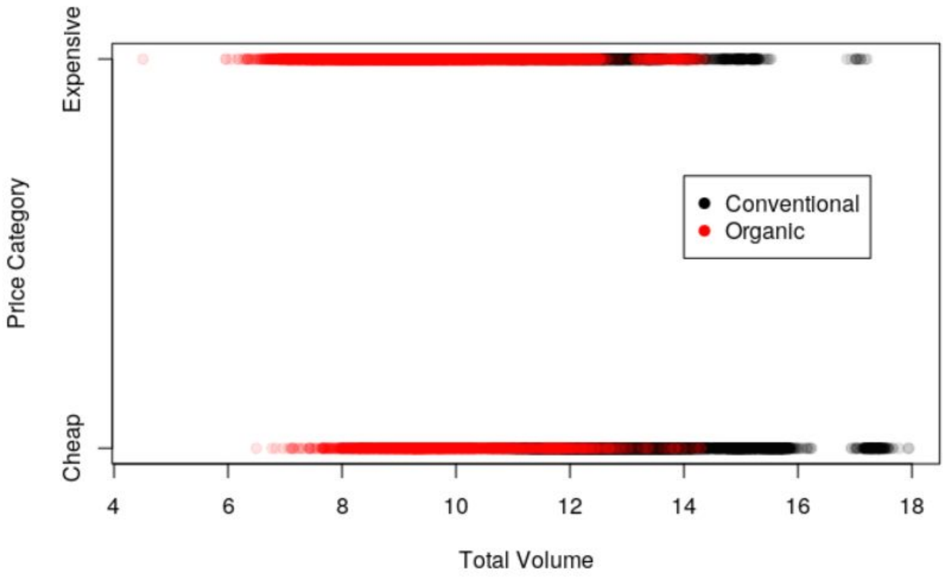
Logistic Model with Total Volume and Type



- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

Logistic Model with Total Volume and Type



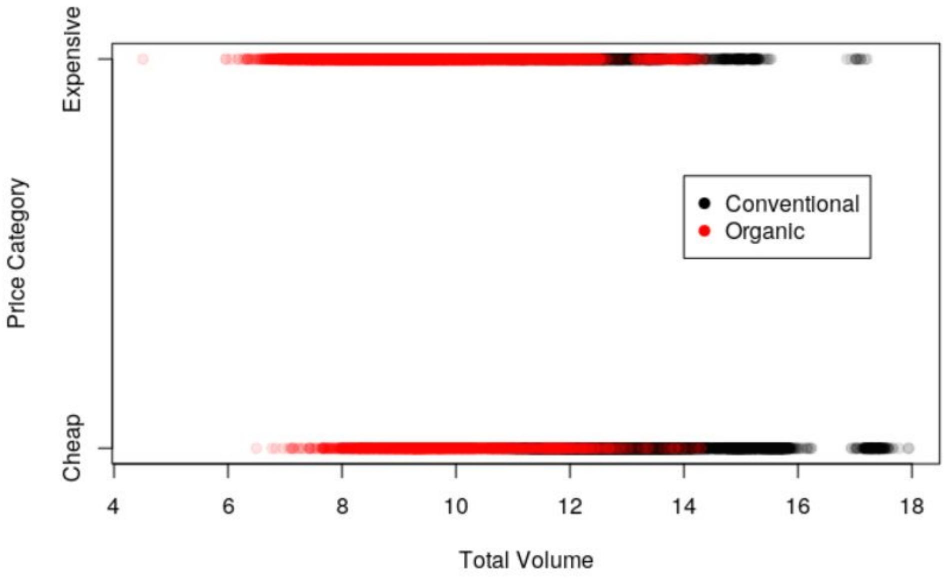
- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

- If conventional avocado,

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume}$$

Logistic Model with Total Volume and Type



- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

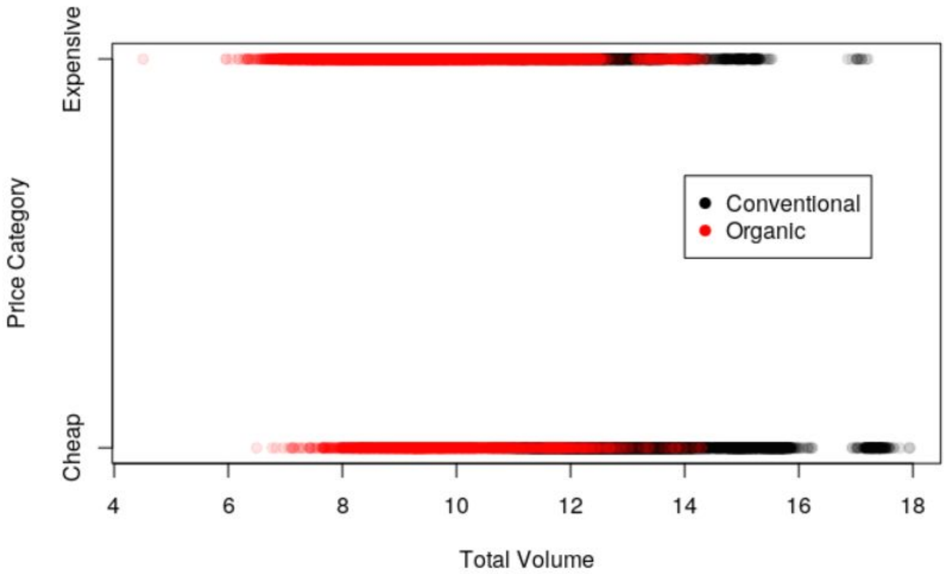
- If conventional avocado,

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume}$$

- If organic avocado,

$$\log \frac{P}{1-P} = 3.74 - 0.24 \times \text{Total Volume}$$

Logistic Model with Total Volume and Type

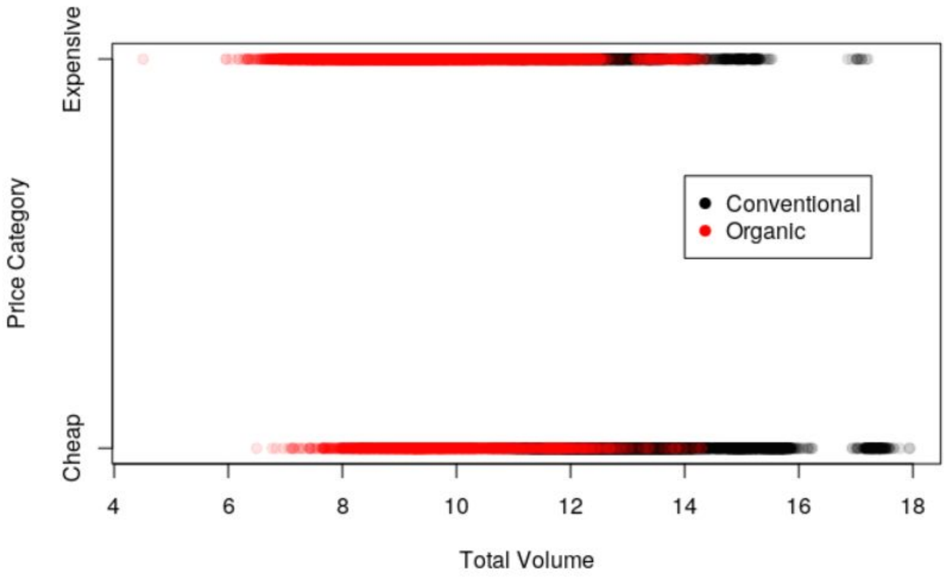


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

1) Print the model summary

Logistic Model with Total Volume and Type

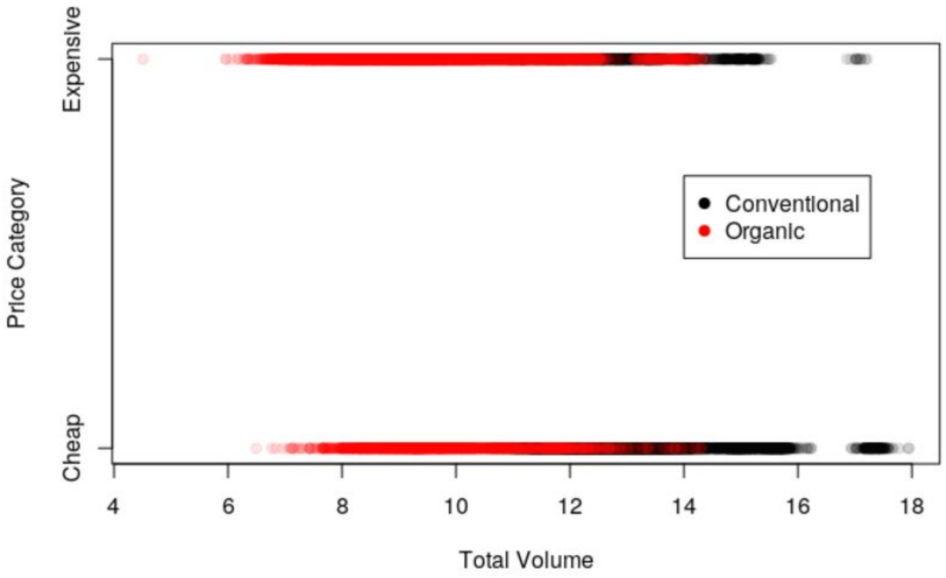


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

1) Print the model summary
2) Compute the predicted log-odds and probabilities on unseen data

Logistic Model with Total Volume and Type

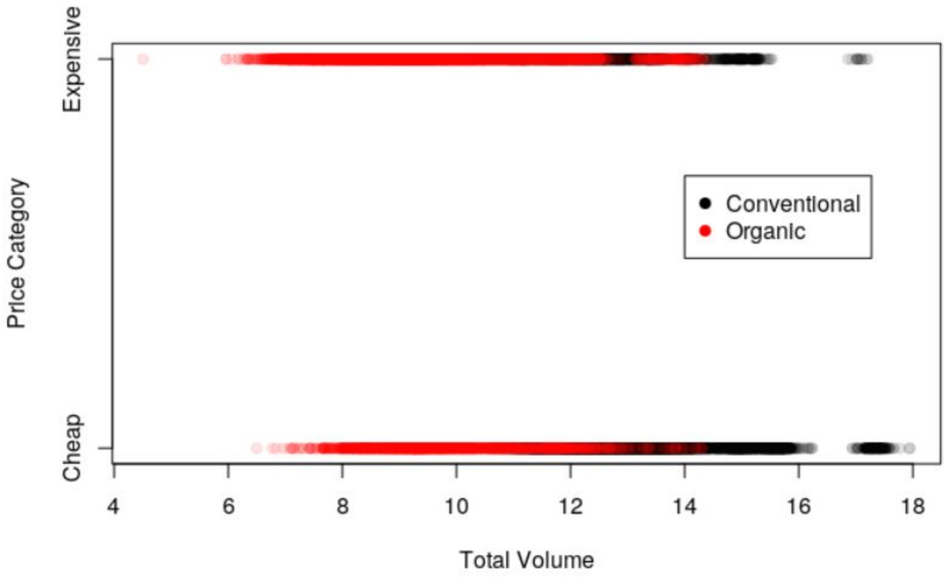


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

1) Print the model summary
2) Compute the predicted log-odds and probabilities on unseen data
3) Predict classes from the probabilities scores assigned to trained data

Logistic Model with Total Volume and Type

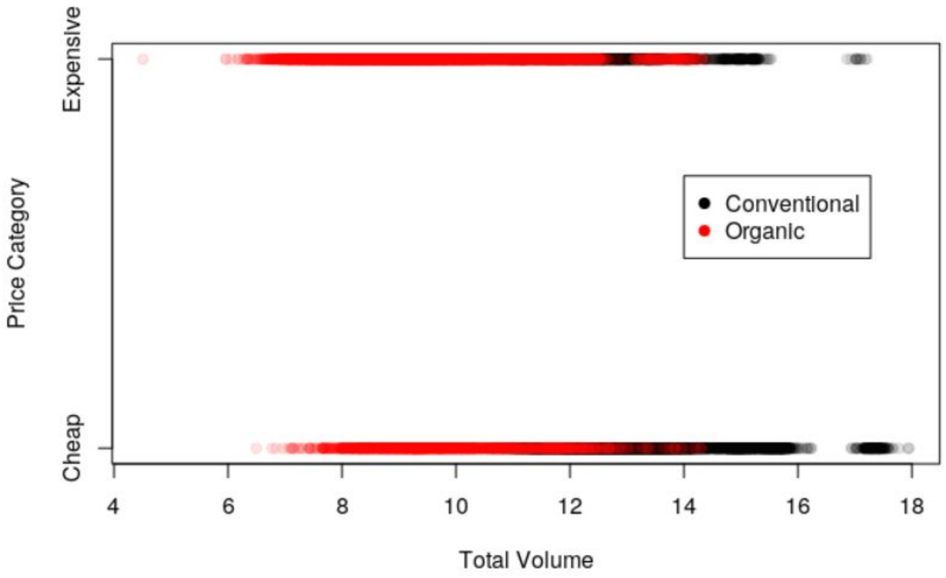


- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

1) Print the model summary
2) Compute the predicted log-odds and probabilities on unseen data
3) Predict classes from the probabilities scores assigned to trained data
4) Compute the confusion matrix of predicted classes

Logistic Model with Total Volume and Type



- Let's use a logistic model to predict the probability of being expensive using total volume sold and type

$$\log \frac{P}{1-P} = 1.83 - 0.24 \times \text{Total Volume} + 1.91 \times \text{Type}$$

1) Print the model summary
2) Compute the predicted log-odds and probabilities on unseen data
3) Predict classes from the probabilities scores assigned to trained data
4) Compute the confusion matrix of predicted classes
5) Use ANOVA to compare this model to the null model and the model with total volume.

Conclusion and Next Steps

- We found reasonable logistic models of whether an avocado was cheap or expensive using total volume and type.
- Exercises will allow you to experiment further with year, volume and type.
- Going further, you might want to consider
  - Model selection for logistic regression
  - Goodness of fit measures: AIC, BIC
  - Statistical tests for goodness of fit
  - Comparing logistic models using ROC curves and AUC
  - Picking the best threshold value

Exercises

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