**Case 2: Organic Banana**

RETL 603

Professor Yu Ma

Ellen Jia

November 10th, 2023

表格

中度可信度描述已自动生成

1. **Visualize the relationship between banana sales and price overall. Does it appear to be a linear relationship?**

There is a linear relationship.

The plot depicts a trend where higher sales volumes cluster at lower prices, and there are two clusters in the low-price range—one centered around the 10-15 sales range and another within the 0-5 sales range. As prices increase, there is a visible decrease in sales volumes within both clusters. This observation of decreasing sales with increasing prices suggests linear relationship, but there might be other factors such as banana type (organic and regular) and retailer type (express and grocery) influencing the relationship between price and sales so it might not be single linear relationship.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成

1. **Perform a simple linear regression model analysis with the dependent variable (DV) as sales\_kkg, and the only predictor as price\_kg. Explain your findings, including R-squared, significance and the interpretation of the intercept and slope.**

**图形用户界面, 文本, 应用程序

描述已自动生成表格

描述已自动生成**

**R-squared:** The R-squared value is 0.127, which means that approximately 12.7% of the variability in banana sales is explained by the model that includes price per kilogram as the predictor. This is consistent with the analysis in question one that although this relationship appear to be linear, the single linear regression cannot explain the data well.

**Significance:** The p-values for both the intercept and the slope are 0.000, indicating that they are statistically significant. This means there is a very low probability that these results could be due to random chance, which implies that there is a relationship between the DV and the IV.

**Intercept (const):** The intercept value of 15.9856 theoretically represents the expected sales of bananas in thousands of kilograms when the price is $0/kg. This scenario is not practical for prediction purposes as it implies free bananas. So the intercept here does not help with the sales forecasting, it is just the point where the extended regression line intersects with the y-axis.

**Slope (price\_kg):** The slope is -5.1088, meaning that for every one dollar increase in price per kilogram, the sales of bananas are expected to decrease by 5.1088 thousand kilograms. This shows an inverse relationship between price and sales.

1. **Perform a multiple linear regression model analysis by incorporating the ‘retailer’ variable into the model mentioned in Q2. Since “retailer” is a categorical variable, set the base level (reference level) to ‘express’. Explain your findings, including R-squared, significance and the interpretation of all model parameters. What additional insights do you gain by including this variable?**

图形用户界面, 文本

描述已自动生成

手机屏幕截图

描述已自动生成

文本

描述已自动生成

文本

描述已自动生成

图表, 折线图

描述已自动生成

**R-squared:** The R-squared value is 0.946, indicating that 94.6% of the variance in banana sales is explained by the model, which is a substantial improvement from the simple regression model. This shows that the relationship between sales and price is influenced by the retailer type.

**Intercept (const):** The coefficient of 4.7241 indicates the theoretical expected sales at express stores when the price is $0/kg. It is the point where the extended regression line intersects with the y-axis and has no reality meanings. The p-value of 0.000 suggests this value is statistically significant.

**Price\_kg:** The coefficient of -1.3754 suggests that for each $1 increase in price, sales decrease by about 1.3754 thousand kg. The p-value of 0.000 indicates that price is a significant predictor of sales

**Grocery:** The coefficient of 9.9333 means that sales at grocery stores are 9.9333 thousand kg higher than at express stores which means that in general the sales in grocery stores is higher than that of express stores. The p-value of 0.000 confirms that the type of store is a significant predictor of sales.

Including the retailer variable shows that the type of store (grocery and express) has a significant impact on sales, which wasn't accounted for in the simple model. The model's predictive power is greatly enhanced by including the store type, as reflected in the increased R-squared value.

1. **One might expect consumers to be less price-sensitive at express stores due to their typically higher prices and the convenience factor. How can this be statistically tested? You can introduce an interaction term between 'retailer' and 'price' into the model of Q3. Perform this analysis and explain your findings, including what additional insights you gain by including this interaction term.**

****

**手机屏幕截图

描述已自动生成**

**文本

描述已自动生成**

**图表, 折线图

描述已自动生成**

The formula is sales=intercept+(retailer coefficient)\*retailer + (price coefficient)\*price+(interaction term)\*retailer\*price

For express stores, sales=intercept+(price coefficient)\*price

For grocery stores, sales=[intercept+(retailer coefficient)]+[(price coefficient)+(interaction term)]\*price

**Intercept:** The coefficient of 3.8080 indicates the theoretical expected sales at express stores when the price is $0/kg. It is the point where the extended regression line intersects with the y-axis and has no reality meanings. The p-value of 0.000 suggests this value is statistically significant.

**Retailer Coefficient (Grocery):** The coefficient for grocery stores is 14.0440. It suggests that, all else being equal, theoretically, sales of bananas in grocery stores are 14,044 kilograms higher than in express stores when the price is $0/kg although this has no meanings in reality.

**Price Coefficient:** The price coefficient of -0.8629 indicates that for each $1/kg increase in the price of bananas, sales decrease by 862.9 kilograms in express stores

**Interaction Term:** The interaction term coefficient of -2.5285 shows that the price sensitivity is different for grocery stores compared to express stores. Specifically, for every $1/kg increase in price, the decrease in sales is an additional 2,528.5 kilograms in grocery stores compared to express stores. This supports the hypothesis that consumers at grocery stores are more price-sensitive than at express stores. The negative sign indicates that as price increases, sales decrease more steeply in grocery stores.

**Statistical Significance:** All p-values are less than 0.05, which means that the coefficients for the intercept, retailer type, price, and the interaction term are statistically significant.

1. **It's also possible to suspect that consumers who purchase organic bananas differ from those who buy regular bananas. The former may prioritize health and well-being, and they may be willing to pay a higher price, suggesting potentially lower price sensitivity. Can you test this hypothesis using the data and multiple linear regression? Perform the corresponding analysis and explain your findings.**

**表格

描述已自动生成**

**图表, 折线图

描述已自动生成**

The formula for bananas in express stores is:

sales=intercept+(organic coefficient)\*organic + (price coefficient)\*price+(interaction term)\*organic\*price

For regular bananas, sales=intercept+(price coefficient)\*price

For organic bananas, sales=[intercept+(organic coefficient)]+[(price coefficient)+(interaction term)]\*price

**Intercept:** The coefficient of 3.5288 indicates the theoretical expected sales when the price is $0/kg. It is the point where the extended regression line intersects with the y-axis and has no reality meanings. The p-value of 0.000 suggests this value is statistically significant.

**Organic Coefficient:** This significant negative coefficient of -1.9522 suggests that, holding price constant, sales of organic bananas are lower than regular bananas by 1952.2 kilograms, reflecting less demand or smaller market size for organic bananas.

**Price Coefficient:** The price coefficient of -0.5448 indicates that for each $1/kg increase in price, sales of regular bananas decrease by 544.8 kilograms, suggesting more price sensitivity among consumers of regular bananas.

**Price Organic Interaction Term:** The interaction term has a positive and significant coefficient of 0.6434, indicating that the price sensitivity of organic bananas differs from that of regular bananas. It shows that every $1/kg increase in price has a negative impact on sales of 643.4 kilograms less for organic bananas than for normal bananas. This supports the hypothesis that consumers of organic bananas are potentially less price-sensitive, possibly due to factors such as health and well-being priorities.

**Statistical Significance:** All variables, including the interaction term, have p-values less than 0.05, indicating that the findings are statistically significant.

1. **Building on your findings in Q5, demonstrate how to manually forecast the sales of organic bananas at the express stores when the price of organic bananas is $2. Similarly, demonstrate how to arrive at the sales of regular bananas at the express stores when the price of regular bananas is $2.**

For regular bananas, sales=intercept+(price coefficient)\*price

Sales=3.5288+(-0.5448)\*2=2.4392

For organic bananas, sales=[intercept+(organic coefficient)]+[(price coefficient)+(interaction term)]\*price

Sales=(3.5488-1.9522)+(-0.5448+0.6434)\*2=1.7738

According to the calculations above, the predicted sales of bananas at express stores with a price set at $2/kg are 2439.2 kilograms for regular bananas and 1773.8 kilograms for organic bananas.  
文本

描述已自动生成

1. **Visualize the relationship between sales and week for non-organic bananas in express stores. Does it suggest a linear or nonlinear relationship?**

Nonlinear relationship.

The sales appear to fluctuate, with periods of increase and decrease rather than following a consistent trend throughout the year. Specifically, sales increase from weeks 0-5, decrease from 5-10, rise again from 10-20, fall from 20-30, and then rise from 30-45 before declining again towards the end of the year. This fluctuation pattern cannot be captured by a straight line, which is what we would expect in a linear relationship.

Moreover, the R-squared value of 0.300 indicates a relatively weak fit for the linear model, meaning that only 30% of the variance in sales is explained by the week of the year, leaving a substantial 70% potentially influenced by other factors or nonlinear patterns. This further supports the assessment that the relationship between week and sales is nonlinear.

图形用户界面, 文本, 应用程序

描述已自动生成

图表, 散点图

描述已自动生成

1. **Take sales as the dependent variable (DV) and price as the predictor variable. Add up to 5 degrees of polynomial functions to see if the R-squared (R^2) improves. For example, for 2 degrees, the model would be: sales\_kkg ~ price\_kg + week + I(week\*\*2). Which model will you choose and why?**

手机屏幕截图

描述已自动生成

The codes and the tables follow the text and analysis.

The Polynomial Degree 5 Model has the highest R-squared value (0.775), indicating that it explains the most variation in the data, and the lowest AIC and BIC values. This model is the best among the polynomial models. It suggests a high-degree polynomial relationship between sales, price, and week.

文本

描述已自动生成

手机屏幕截图

描述已自动生成

表格

描述已自动生成手机屏幕截图

描述已自动生成

表格

描述已自动生成表格

描述已自动生成

手机屏幕截图

描述已自动生成表格

描述已自动生成

表格

描述已自动生成表格

描述已自动生成

**Question 9: Perform a 5-fold cross-validation and select the optimal degree for the polynomial function based on validation Mean Squared Error (MSE).**

According to the five 5-fold MSEs, the 5 degree polynomial model is the best since it has the smallest MSE, so it is the best one.

图形用户界面, 文本, 应用程序

描述已自动生成

手机屏幕截图

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成

表格

描述已自动生成

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

一些文字和图片的手机截图

描述已自动生成

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

一些文字和图片的手机截图

描述已自动生成

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成一些文字和图片的手机截图

中度可信度描述已自动生成