BZAN 535 - Homework 10

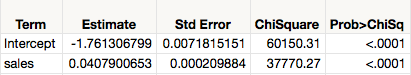
Kevin Gardner

Due on 11/23/2016

## Question 1

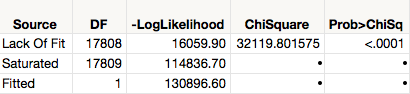
Using the attached file from the journey data, fit a logistic regression model for Y = Produce (based on the basket containing one item or more from the Produce Department) and X = basket $ total.

1. Report the slope coefficient and interpret this corresponding odds.



The coefficient of 0.0408 (p<.0001) for basket $ total indicates that the odds of a purchase increase by about 4.16% (or a factor of 1.0416) for each additional $ in sales.

1. Does this model have significant lack of fit? Are there any outliers evident in the deviance residual plot? What do you conclude? Is this model generally satisfactory?



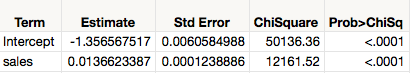
The chi-square for lack of fit is about 32120 (p<.0001), so the model has significant lack of fit. From the residuals plot we observe that the most extreme residuals are large baskets with no produce.

Based on the lack of fit and studentized deviance residuals greater than 3 in magnitude, we conclude that the model is not generally satisfactory.



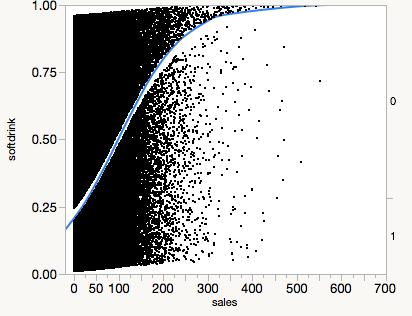
## Question 2

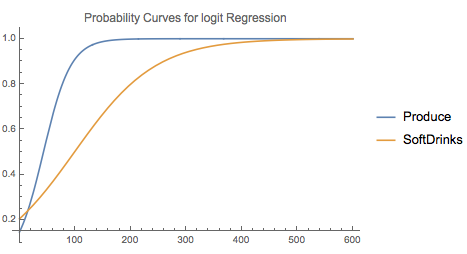
Fit a logistic regression model for Y = Soft drinks and X= basket $ total. Report the equation of the model for logit(Soft drinks). Then re-write this equation for the probability of a basket containing soft drinks.



1. Contrast the model for predicting purchase of soft drinks with the model for predicting purchase of produce. In particular, which are more common in baskets and how is this reflected in the logistic regression model? What does the difference in the slopes tell you about shopping behaviors related to produce and soft drinks?

The slope for softdrink () is smaller than the slope for produce (0). This indicates that as the size of basket increases, log odds for produce increase more quickly than log odds for soft drinks. This is illustrated below by the steepness the regression curves.



Based on overall probability produce is more common: The probability curve for produce is higher than the curve for soft drinks at almost all basket sizes.

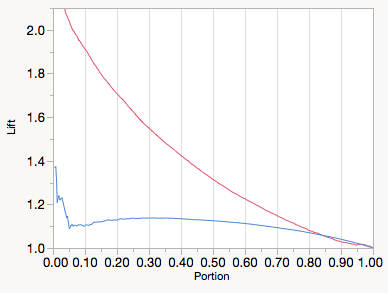
1. Show and interpret the lift plots for the model in 1a and 2. Explain what about the fitted logistic regression models cause the differences you see in lift.

If we take the largest 10% of baskets with the highest estimated P(Produce|$size), we will have about 2.4 times the chance of Produce versus if we ignore basket size.

If we take the largest 10% of baskets with the highest estimated P(Softdrink|$size), we will have about 1.9 times the chance of Softdrink versus if we ignore basket size.

The odds of including produce is much more dependent on the size of the basket than is the odds of including soft drinks.

**Lift Curve (Nominal Logistic Fit for produce) Lift Curve (Nominal Logistic Fit for softdrink)**

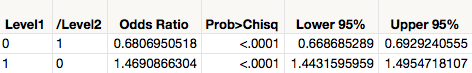
**** 

## Question 3

When we discussed the odds ratios in Lectures 20-22, we noted that ignoring basket size made commodities all appear to have affinity for one another.

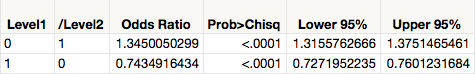
1. What is the odds ratio for produce and soft drinks? Interpret this quantity.

**Odds Ratios for softdrink (not including $size)**



According to this model, the odds of having soft drinks with produce are about 47% greater than the odds of having soft drinks without produce.

1. Use a logistic regression to show that the answer in a) is misleading, and that when we control for basket size, the correct value is less than 1. Show the results of such an analysis, and interpret the resulting odds ratio. [Hint: Think about how a model containing basket size (in $) provides estimates for other effects in the model, controlling for basket size.]

**Odds Ratios for softdrink (including $size)**

The model used in part (a) is misleading because it conflates basket size effect with a true affinity between soft drinks and produce.

Controlling for basket size, the odds of having soft drinks with produce are about 26% lower than the odds having soft drinks without produce.

## Question 4

Write the query for producing the data I provided for questions 1-3. Be sure to exclude baskets that involve gas, as we did in a query for HW9.

SELECT jt.basket\_id, sum(jt.sales\_value) as sales

, max(case when jp.department='PRODUCE' then 1 else 0 end) Produce

, max(case when jp.commodity\_desc='SOFT DRINKS' then 1 else 0 end) softdrink

FROM journey\_transaction\_data jt JOIN journey\_product jp

ON jt.product\_id=jp.product\_id

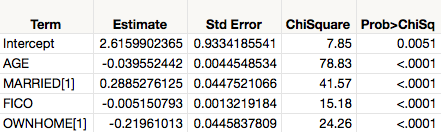
WHERE jp.department NOT IN ('KIOSK-GAS', 'MISC SALES TRAN')

GROUP BY basket\_id;

## Question 5

Using logistic regression for the buytest data, which factors affect the odds of buying? Answer this question by:

1. Reporting a logistic regression model that includes all the explanatory variables that have a statistically significant effect on P(Purchase).

**Nominal Logistic Fit for Purchase**

1. Describing the most likely buyer according to your model; include the estimated probability of a purchase for such individuals.

Based on our model, the most likely buyer is married, young, does not own a home and has a low FICO score. Evaluating the equation, we find the probability of purchase for such an individual is about 36%.

