Problem Set #5

Econ 815
Aditya Balaram

The problem that I am considering for this assignment is that of predicting the number of returned products for an online retailer. More specifically, I try to establish characteristics of baskets that drive the fraction of products that are returned. This is a problem of importance for online retailers since returns pose a number of problems that can be hard to deal with. While the goal here isn't to limit returns, being able to forecast how many products are going to be returned at a given time can help a retailer with staffing decisions, the amount of financial resources to devote to returns processing, and building of contracts with shipping companies.

The dataset that I use is from an online jewelry retailer based in the US. This dataset includes data on product level characteristics, product images, and so on but I focus on the subset that deals with order characteristics. In the data that I use for this assignment, I have access to information on the consumer ID, order ID, products that are ordered, physical location to which the order is shipped, amount paid for shipping, the shipping option that is chosen by the consumer, the price paid for the product, the amount of store credit used, the amount of rewards discount that is applied to the order, the amount paid using a gift card, the quantity returned, and the amount refunded to the consumer.

For this assignment, I focus on predicting the number of returns from a given basket based on a few characteristics of the basket. The characteristics that I consider are as follows:

1. Basket size (# of products in the basket) – I expect larger baskets to have a higher number of returns as consumers might be "trying-on" products to ascertain a match between product characteristics and their expectation.

2. Total $ value of the basket – I expect consumers who spend large amounts to have higher expectations, which will result in more products being returned due to a mismatch between product characteristics and expectation. Further, the risk associated with these products is more and consumers may not be willing to incur the loss associated with just keeping the product.

3. Store credit used for purchase – Store credit is usually provided to consumers who have had transactions with the retailer in the past. This indicates that these consumers are more familiar with the retailer's products, which possibly reduces their likelihood to return.

4. Shipping price paid – Consumers who are willing to pay for higher shipping are probably those that aren't very cash strapped and have high expectations of the products being ordered. This could in turn lead to them returning products more frequently.

I begin by providing some visual evidence of three of these relationships. Firstly, in Figure 1, I illustrate the distribution of the fraction of returned products from an order. I exclude all those baskets with no returns as there is a very large number of such baskets and it becomes hard to visualize the distribution of returns.

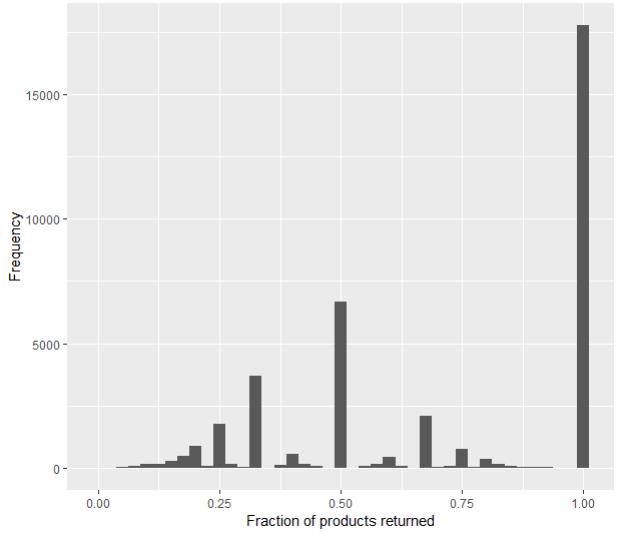

Figure 1: Histogram of fraction of returns

Figure 1 indicates that among baskets that have at least one product returned, a large number have the entire basket being returned. Further, there is a reasonable number of baskets with 20-60% of products being returned. Next, I present a scatter plot to look at the relationship between the size of the basket and the fraction of returned products.

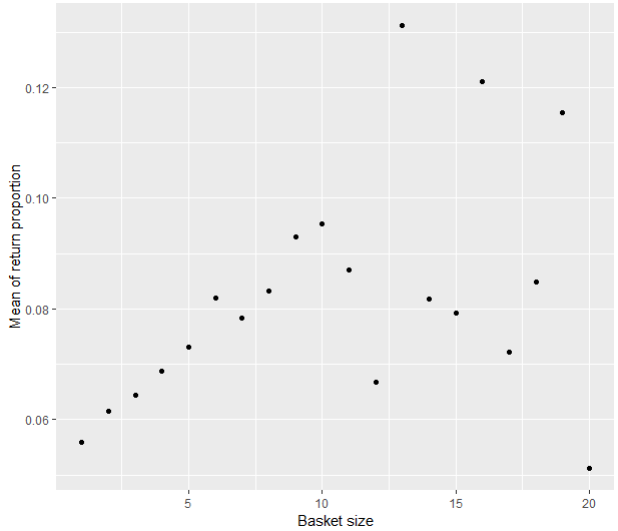

Figure 2: Scatter plot of basket size and mean proportion of returns

The horizontal axis in Figure 2 represents the basket size, and the vertical axis

represents the mean proportion of returned products for a given basket size. Figure 2 seems to suggest that as the basket size increases, so does the fraction of products that are returned.

Finally, Figure 3 seems to provide minimal evidence that the higher the basket value, the larger the proportion of returned products. I say this because the point cloud doesn't remain as concentrated for more valuable baskets, and instead moves rightward.

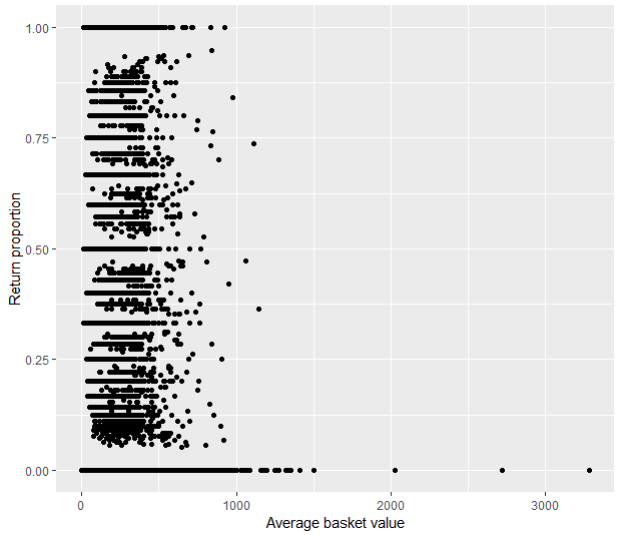

Figure 3: Scatter plot of basket value and proportion of returns

I now present three regression models that help with understanding how these factors affect the fraction of products that are returned. In the first model, I use OLS estimation to obtain the parameters of the model. Since the dependent variable (fraction of products that are returned) is a proportion i.e., it lies in $[0, 1]$. The downside with using OLS for such a dependent variable is that it fails to capture the fact that the variable is constrained to $[0, 1]$. One possible way to deal with this problem is to use a logistic regression model which is the second model that I use. While this does restrict the dependent variable in the required manner, an assumption that we are making here is that each product in a basket is governed by the same "success" or return probability. The final model, which I believe is the most appropriate, that I use is a Tobit model since the dependent variable is censored to the left at 0 and to the right at 1. I make a few modifications to some of the dependent variables in order to have them in similar scales to the dependent variable. I use the natural logarithm of all price related variables but since a lot of these variables can take the value 0, I consider $\ln(\text{Price Variable} + 1)$.

Rather than making an assumption like that in the logit model, we really treat the basket as a whole in the Tobit model. The results from the three models are presented in Table 1.

	OLS	Logit	Tobit
Basket Value	0.02***	0.55***	0.88***
	(0.00)	(0.01)	(0.01)
Basket Size	-0.00***	-0.07***	-0.04***
	(0.00)	(0.00)	(0.00)
Store credit used	-0.00***	-0.05***	-0.04**
	(0.00)	(0.01)	(0.01)
Price of shipping option	-0.00**	-0.03**	-0.04***
	(0.00)	(0.01)	(0.01)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Three regression models. Standard errors are shown in parentheses

The parameter estimates are surprisingly consistent across the three models. We see that only basket value has a positive impact on the fraction of products that are returned. This indicates that the retailer should expect more returns from baskets that have higher value. On the other hand, basket size, the amount of store credit used, and the price of the shipping option used, have a negative impact on the fraction of products returned. This suggests that retailers should expect fewer returns from consumers who order large quantities, pay larger amounts through store credit, and pay larger amounts for shipping. Two of these negative effects are the opposite of what I had expected.

It is also worth looking a little deeper at the impacts of basket value and size. Obviously, as the size increases, so does value. Yet, these results suggest that a consumer who orders a lot of cheap products will return fewer products than one who orders a similar quantity of expensive products. Also, the magnitude of the coefficient of value is much larger than size, thereby indicating that the impact of basket value is far more pronounced than basket size.