Goodbelly Case: Technical Summary

Excerpt from a report submitted April 30, 2019 for a graduate Marketing Analytics class at Wayne State University

Contact: Devon Ankar (devon@ankar.io)

Approach

This course project required that we use multiple linear regression (MLR) to develop our models. It also required that we divide our analysis into two stages:

Stage 1: explanatory modeling based on strong hypotheses (called "theories" in this course)

Stage 2: data mining to look for insights that may be unexpected based on theory

Our project also required that we use explanatory modeling techniques (as opposed to predictive modeling) to look for explanations about phenomena: to explain what drives existing sales, not to predict future sales.

Strongest Theories (Hypotheses): Stage 1, Substage 1

In this stage, we begin with a theoretical analysis and then move to modeling. First, we try to understand which independent variables have the strongest impact – in theory – on the dependent variable (sales) and include only those IVs in the model.

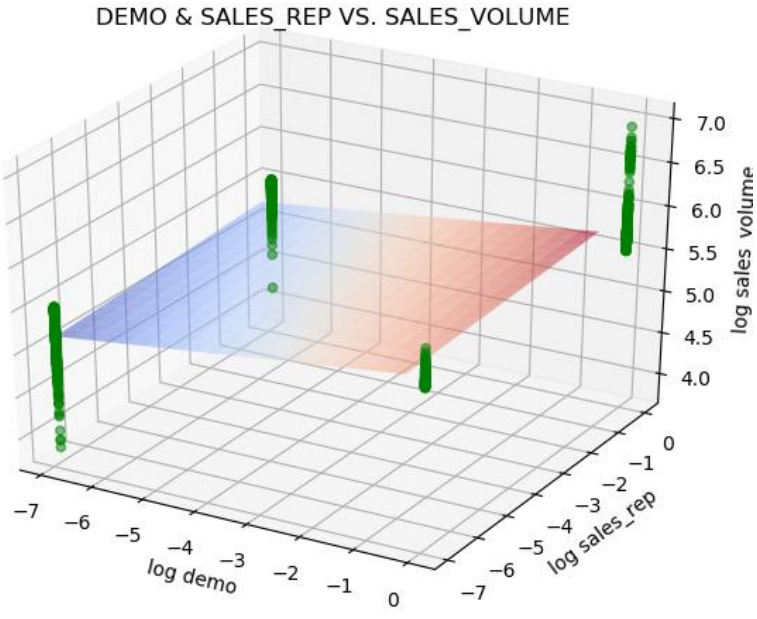
In stage 1, substage 1, we considered only the following IVs:

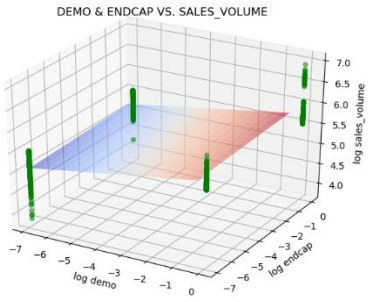
- **Demo**: whether or not there was a product demonstration at the store (binary 0/1)
- **Sales_Rep**: whether or not a sales representative was present at the store (binary 0/1)
- **Endcap**: whether or not there was an endcap display of the product (binary 0/1)

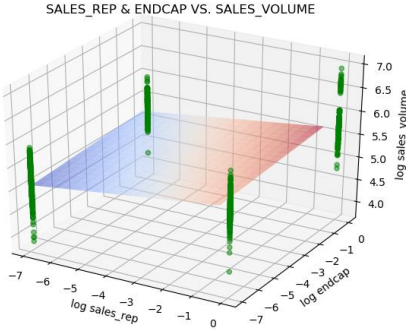
We reasoned that these IVs have the strongest impact on sales – in theory, the presence of a product demonstration, a sales representative, and/or an endcap display should separately and jointly have a meaningful positive impact on sales (i.e. increase sales).

Task: evaluate the effect, if any, of a combination of Demo, Sales_Rep, and Endcap on Sales_Volume. First, we looked at the combination of all three, and found a medium-strength positive correlation (R2 = 0.614) with Sales_Volume.

The plots below show each combination of IVs we used in the model. Only two IVs can be shown in any given plot because we are limited to 3D plots.

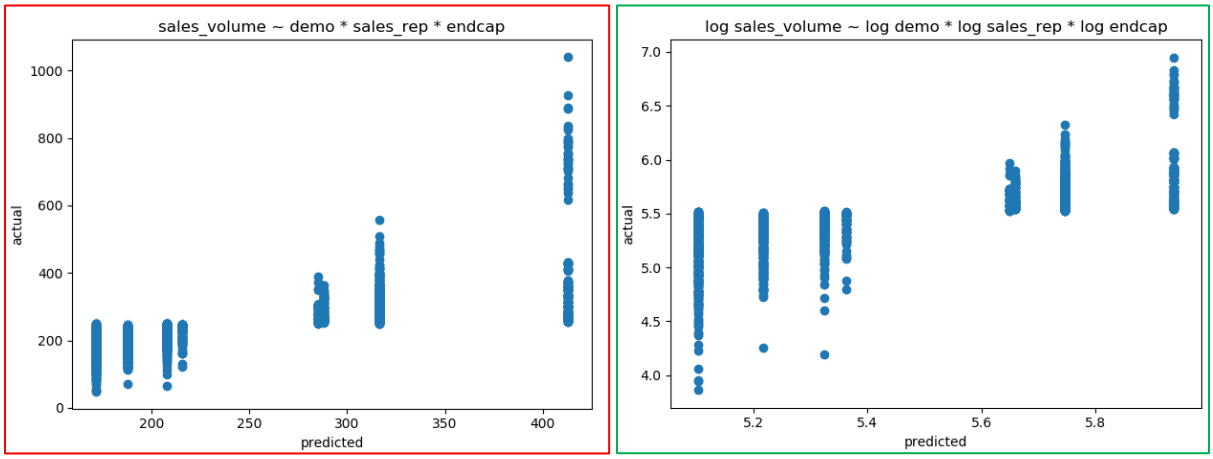






The coefficients for all pure terms and all interaction terms were statistically significant at p = 0.05. Thus, the expected effect is observed.

We used a log-log transformation in which all three IVs as well as Sales were transformed to their natural logarithm. Our initial predicted vs. actual plots with predicted Sales_Volume on the horizontal axis and actual Sales_Volume on the vertical axis showed unsuitability for a linear model. The true relationship between these IVs in combination and Sales is likely nonlinear. Since nonlinear models are outside the scope of this project, we decided to use a linear regression model with a log-log transformation. All models in this project use log-log transformed data. We added 0.001 to both 0 and 1 prior to transformation (`ln(0.001) = -6.9`).



Untransformed & log plots of predicted vs. actual Sales_Volume with Demo, Sales_Rep, and Endcap as IVs.

Based on the above plots, it can be argued that the log-log plot is *still* nonlinear, but it is more plausibly linear, while the untransformed plot shows a clear inflection point near the coordinate (250, 200). Note that these plots give the appearance of 8 levels because we are working with 3 variables, each of which has 2 levels, for a total of 8 combinations:

demo	sales_rep	endcap
1	1	1
1	1	0
1	0	1
1	0	0
0	1	1
0	1	0
0	0	1
0	0	0

Since Endcap's coefficient was the lowest, we decided to try a secondary model dropping only Endcap. We found a lower correlation with only the combination of Demo and Sales_Rep as IVs vs. Sales than with Endcap included. The R^2 was 0.586, which represents a meaningful negative ΔR^2 from the first model. Thus, the first model with the combination of Demo, Sales_Rep, and Endcap on Sales_Volume is the one that we will carry forward from this substage. All three IVs' relationships to Sales are backed by strong theory and all three IVs have statistically significant coefficients.

Weaker Theories (Hypotheses): Stage 1, Substage 2

At this stage, we looked at the variables that had weaker theoretical explanations:

- **Demo1.3**: whether or not a store has had a demonstration within the past 1-3 weeks
- **Demo4.5**: whether or not a store has had a demonstration within the past 4-5 weeks
- **Natural and Fitness**: the number of natural-products stores and fitness centers, respectively, within a 5-mile radius

We ran multiple models to see how these IVs impact the sales individually and via interaction effect.

Adding Demo1.3: **Modeling Demo, Sales_Rep, Endcap, and Demo1.3** vs. Sales_Volume, we found an R^2 of 0.619, only 0.005 higher than our previous model (only Demo, Sales_Rep, and Endcap with R^2 of 0.614). Thus, the ΔR^2 from adding Demo1.3 is small. Also, unlike the previous model, this model has several interaction terms (though no pure terms) with statistically insignificant coefficients. The highest coefficients are still held by Demo, Sales_Rep, and Endcap, along with Demo1.3; the interaction terms have lower, even if statistically significant, coefficients. Because the coefficient for the Demo1.3 term is statistically significant, and Demo1.3 has a plausible theoretical explanation above for at least a weak relationship with Sales, we can overlook the low ΔR^2 and consider this part of our secondary models for now.

Adding Demo4.5: **Modeling Demo, Sales_Rep, Endcap, and Demo4.5** vs. Sales_Volume, we found an R^2 of 0.618. With respect to this model, we share many of the same concerns as we did with the model just above. The ΔR^2 from adding Demo4.5 is even smaller, and several interaction terms (though no pure terms) have coefficients that are statistically insignificant. More importantly, though, the theory for adding Demo4.5 is even more implausible than that for Demo1.3; demos conducted nearly a month before are less likely to meaningfully impact sales that occur much later in time. The impact is statistically significant, but not large enough to be meaningful. The small ΔR^2 from adding Demo4.5, combined with the weak theoretical justification for adding Demo4.5, leads us to reject this model.

Adding Natural: **Modeling Demo, Sales_Rep, Endcap, and Natural** vs. Sales_Volume, we found an R^2 of 0.616. This is an even smaller ΔR^2 compared to the above two models. On top of this, our theoretical justification for adding Natural is weak to begin with, leading us to reject this model. Lastly, Natural itself does not have a statistically significant coefficient in this model.

Adding Fitness: **Modeling Demo, Sales_Rep, Endcap, and Fitness** vs. Sales_Volume, we found an R^2 of 0.620, which is 0.006 higher than our first model in substage 1. Thus, the ΔR^2 from adding Fitness is still small, but more meaningful. Also, all of the pure term coefficients in the model are statistically significant, while some interaction terms are not. The only reason to reject this model would be the aforementioned dubious (but plausible) theory, so we will consider this part of our secondary models for now.

Adding both Demo1.3 and Fitness: **Modeling Demo, Sales_Rep, Endcap, Demo1.3, and Fitness** vs. Sales_Volume, we found an R^2 of 0.627, which is a healthy ΔR^2 from our first model in substage 1. Also, all the pure terms have statistically significant coefficients, with Demo, Sales_Rep, Endcap, and Demo1.3 having the highest coefficients. After that, some interaction terms have high coefficients that are also statistically significant. *(The details are available in the outputs folder.)* Unfortunately, Fitness was found to have a negative coefficient (-0.029) that is nevertheless also statistically significant ($p < 0.001$), indicating that in the context of this model, Fitness has a negative relationship with Sales. (A model of simply Fitness vs. Sales shows that Fitness has a weak, but statistically significant, positive relationship with Sales.) While the theoretical justification for a positive relationship between Fitness and Sales is weak, the theoretical justification for a negative relationship between Fitness and Sales is nonexistent – we could not figure out a reason why, even in the context of this model, the presence of fitness centers nearby would drive customers away from GoodBelly products. In the absence of any theoretical justification whatsoever, we decided to reject this model.

Light "Data Mining": Stage 2

Looking for relationships in the data that are not necessarily apparent based on theory

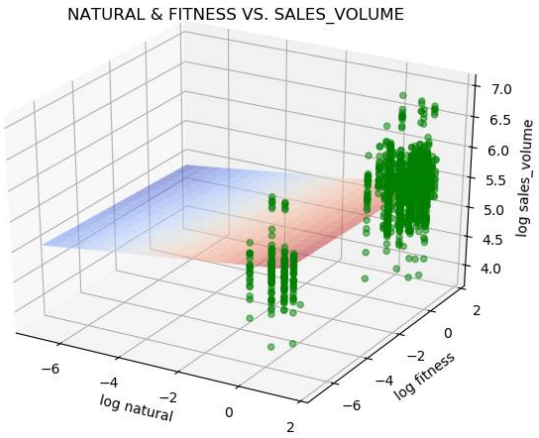
Modeling Demo, Sales_Rep, Endcap, and ARP vs. Sales_Volume, we found an R^2 of 0.628, which represents an increase of 0.008 from our secondary model above with an R^2 of 0.620. In this model, all pure terms have statistically significant coefficients. In the context of this model, ARP has a statistically significant negative correlation with sales (coefficient is -0.602); however, this is actually to be expected based on theory. According to the price-volume curve, as average retail price (ARP) increases, sales volume decreases because customers buy fewer items at the higher price than they would at a lower price. (This is not a bad thing because it does not mean sales revenue is lower, just that sales volume is lower.) The other pure coefficients are all positive and statistically significant, and thus easily explainable (as explained above), so this model is a good candidate for our final model.

Modeling Demo, Sales_Rep, Endcap, ARP, and Demo1.3 vs. Sales_Volume, we found an R^2 of 0.632. We rejected this model because Sales_Rep and Demo1.3 pure terms no longer have statistically significant coefficients and there is no good reason why Sales_Rep would no longer have a positive impact on sales.

Modeling Demo, Sales_Rep, Endcap, ARP, and Demo4.5 vs. Sales_Volume, we found an R^2 of 0.637. We rejected this model because Demo, Sales_Rep, and Endcap no longer have statistically significant coefficients and there is no theoretical reason that these three variables should no longer have an impact on sales. Also, Demo4.5 as well as the interaction term Demo4.5*Endcap, now both have negative coefficients (both statistically significant), for which there is no theoretical explanation. For these reasons, we rejected this model.

Modeling Demo, Sales_Rep, Endcap, ARP, Demo1.3 and Demo4.5 vs. Sales_Volume, we found an R^2 of 0.642. We rejected this model because Demo, Sales_Rep, and Endcap no longer have statistically significant coefficients and there is no theoretical reason that these three variables should no longer have an impact on sales. Also, Demo4.5 as well as the interaction term Demo4.5*Endcap, now both have negative coefficients (both statistically significant), for which there is no theoretical explanation. For these reasons, we rejected this model.

Modeling Demo, Sales_Rep, Endcap, ARP, Natural and Fitness vs. Sales_Volume, we found an R^2 of 0.646. We rejected this model because Demo and Sales_Rep both no longer have statistically significant coefficients in this model, and there is no theoretical explanation for why both Demo and Sales_Rep would no longer have a positive impact on sales. Additionally, Fitness has a statistically significant, but negative, impact on sales. There is no theoretical reason why the presence of nearby fitness centers would negatively impact sales. The explanation for a positive correlation might be weak, but there is no reason for a negative correlation. For these reasons, we rejected this model.



Plot showing weak positive relationship between Natural, Fitness, and Sales_Volume.

Modeling Demo, Sales_Rep, Endcap, Demo1.3, Demo4.5, Natural and Fitness vs. Sales_Volume, we found an R^2 of 0.649. We rejected this model because of similar reasons to the above: Sales_Rep had no correlation to sales (statistically insignificant coefficient), for which there is no theoretical reason.

Our last model had the highest R^2 at 0.690. By modeling Demo, Sales_Rep, Endcap, ARP, Demo1.3, Demo4.5, Natural and Fitness vs. Sales_Volume, we found an R^2 of 0.690. Despite this comparatively high R^2, we rejected this model for many of the same reasons as above: key terms such as Demo, Sales_Rep, and Endcap had coefficients that were not statistically significant, and other coefficients such as that belonging to the pure term for Fitness were statistically significant but negative, which again has no theoretical explanation. For these reasons, we rejected this model.

Model Comparison and Model Choice

Ultimately, we chose the model of Demo, Sales_Rep, Endcap, and ARP vs. Sales_Volume because it was the best explanatory power (R^2 of 0.628) we could get without compromising theoretical strength or "explainability". As discussed earlier, the other contenders for our final model all had lower explanatory power, and all the models with a higher R^2 than this one had theoretical problems (coefficients that were negative that should have been positive based on theory and/or coefficients that were not statistically significant that should have been statistically significant based on theory).

The plots below show each new combination of IVs we used in the final model. Only two IVs can be shown in any given plot because we are limited to 3D plots.