Setting Retail Staffing Levels: A Methodology Validated with Implementation

Marshall Fisher
The Wharton School, The University of Pennsylvania, Philadelphia, PA 19104, fisher@wharton.upenn.edu

Santiago Gallino
The Wharton School, The University of Pennsylvania, Philadelphia, PA 19104, sgallino@wharton.upenn.edu

Serguei Netessine
The Wharton School, The University of Pennsylvania, Philadelphia, PA 19104, netessin@wharton.upenn.edu

We describe a three-step process that a retailer can use to set retail store sales staff levels. First, use historical data on revenue and planned and actual staffing levels by store to estimate how revenue varies with the staffing level at each store. We disentangle the endogeneity between revenue and staffing levels by focusing on randomly occurring deviations between planned and actual labor. Second, using historical analysis as a guide, we validate these results by changing the staffing levels in a few test stores. Finally, we implement the results chain-wide and measure the impact. We describe the successful deployment of this process with a large specialty retailer. We find that 1) the implementation validates predictions of the historical analysis, including the use of the variation between planned staffing and actual staffing as an exogenous shock, 2) implementation in 168 stores over a 6-month period produces a 4.5% revenue increase and a nearly $7.4 million annual profit increase, after accounting for the cost of the additional labor, and 3) the impact of staffing level on revenue varies greatly by store, and therefore staffing levels should also vary, with more sales staff relative to revenue assigned to those stores where sales staff have the greatest impact on revenue. Specifically, we found the largest impact of store labor in stores with the largest average basket sizes, located in regions with good growth potential, facing certain competitors (e.g., Wal-Mart), and run by long-serving managers

Key words: retail operations, business analytics, empirical operations management

History: Revised July 28, 2018

1. Introduction

Achieving profitability in retailing is more important than ever. In 2017 alone Payless, The Limited, RadioShack and 6 other retail chains filed for bankruptcy protection, while other nation-wide chains including J.C.Penney, Sears, Kmart and Macy's are closing hundreds of stores, on track to set a historical record (The Economist (2017), Loeb (2017)). Changing consumer habits and competition from online retailers are decreasing demand and, as noted in Fisher et al. (2017), making it imperative for retailers to pursue operations improvements that can increase sales in their existing stores. Arguably, the two most important factors of production for a retailer are the inventory and labor within a store. The Cost of Good Sold (COGS) (the cost of inventory) and store payroll are, respectively, the largest and second largest costs for most retailers. Moreover, several recent papers, including Fisher et al. (2006), Ton and Huckman (2008), Perdikaki et al. (2012), Kesavan et al. (2014a), Kesavan et al. (2014b), Ton (2011), and Mani et al. (2015), collectively identify a customer finding a store associate to help them and finding the product the customer wants to buy as the two most important factors impacting sales.

Logically, both the level of inventory in a store and the number of store associates should be set to tradeoff the cost of those resources with the profit earned on the revenue they generate. With inventory, there is a way to do this. For example, Cachon and Terwiesch (2009), Chapter 14, describes a process for first setting an optimal in-stock service level for an item based on its gross margin and inventory carrying cost, and then setting an optimal inventory level to achieve that in-stock, based on a probability distribution of demand.

However, to the best of our knowledge, neither retail practice nor the academic literature provides a similar method for setting store associate staffing levels. Current practice in retail store labor scheduling divides store labor needs into task-related (e.g., receiving a delivery) and selling (e.g., providing product-related information to customers). For task work, industrial engineering methods are used to estimate the time to perform a task. This estimate is then combined with the level of activity in a time period, such as number of deliveries to be received, to estimate the amount of labor needed. The level of sales labor is usually specified as the payroll budget for a store, which is set to a fixed percentage of a forecast of sales for the store for the ensuing period. Sometimes the payroll budget is instead set to a multiple of forecasted store traffic. In either case, the same multiple is used for all stores.

There are several problems with this approach. Intuitively, the level of store labor should impact sales, so setting store labor to a percentage of a sales forecast is circular and fails to take into account how store labor impacts sales. In the worst case, this approach creates a downward spiral: a low sales forecast leads to reducing labor which leads to lower sales and so on. Moreover, in practice the percentage of revenue that store selling labor is set to is somewhat arbitrary and it is usually the same percentage for all stores. The labor level in a store ought to be set to optimally tradeoff the cost of the labor vs. the benefit it generates in sales and gross margin, and as we shall see later in this paper, this optimal labor level varies by store in proportion to sales.

Finally, because setting the store staffing level involves a tradeoff between a known and immediate payroll cost vs. a benefit of increased sales and gross margin that is both uncertain and to some extent will happen in the future, there is a tendency to over-weight the known, immediate cost and therefore understaff stores. In our work with retailers we found that this is mainly a problem with publicly traded companies which, seeking to manage reported earnings, are tempted to temporarily reduce payroll near the end of a quarter (which is doable since much of store labor is part-time and the part-time hours can be easily reduced) to meet a profit target. However, these temporary reductions have a way of becoming permanent, leaving a retailer, over time, with a minimum number of workers earning a minimum wage.

One can envision a more rigorous approach to setting the level of sales labor in a store. Figure 1 suggests what the relationship between store revenue and selling labor should logically look like. Some minimum level of labor (e.g., at least one person) is needed to have the store open before we see any revenue. Then revenue increases as payroll increases, but at a diminishing rate, and eventually flattens out (e.g., when number of employees exceeds number of customers). Payroll should then be set to optimize a profit function equal to the gross margin earned on revenue, less any other variable costs, minus the cost of labor.

Of course, the challenge with this approach is estimating the function shown in Figure 1. Our approach is to use labor variations relative to the plan to estimate the first derivative of this curve at a store's current staffing level.

To motivate the approach, suppose the labor plan (schedule) is to have 30 people in a store and only 27 show up on a given day (due to last-minute illness, etc.). The question then is: what happens to sales? If unchanged from what had been forecasted, then the store is probably overstaffed. If sales decline by 10%, this suggests a linear relationship between staffing and sales, and, to a point, increasing the store sales staff by a certain percentage will add the same percentage increase to sales.

This paper describes the results of rigorously developing this approach at a U.S. specialty retailer with more than 700 stores. Section 2 summarizes related literature. Section 3 describes the data we received from the retailer and Section 4 presents the analysis of this data using the above approach to estimate the revenue increase per payroll dollar increase at each store. We find that the impact of adding store labor varies considerably across the stores. The stores that benefit most from additional labor are those with high potential demand, as indicated by average basket size, total households and annual household growth potential, strong competition for that demand—as indicated by the number and type of competitors within a 10 mile radius—and an ability to effectively use additional labor —as indicated by the years of service of the store manager. We then conducted a validation test, reported in Section 5, in which we increased labor in 16 stores that our analysis showed could benefit from additional labor. The test confirmed our results, so the retailer implemented our methodology chain-wide by increasing store labor in a total of 168 stores, as described in Section 6. We tracked the impact of this implementation using daily revenue data for the 168 treatment stores and 504 control stores over a 182 day period. The implementation is estimated to have added 4.5% to revenue and $7.4 million per year to profit, after accounting for the cost of the additional labor. This increase was significant relative to the retailer's current profit. Table 1 summarizes the timeline of the study, the unit of analysis of the different stages and the number of periods of each stage. Conclusions are offered in Section 7.

Our paper makes three contributions to the academic literature and to retail practice. First, although as described in the next section, several papers have used historical data to provide evidence that store labor significantly impacts revenue, there has been limited field testing to verify these results. Indeed, an excellent paper by Kesavan et al. (2014a) is the only one to report a field experiment which involved increasing store labor by one person in one store for a randomly chosen 4-hour peak traffic block over a 10-day period. Thus, to our knowledge, ours is the first paper to describe a process by which a retailer can systematically set the labor level in each store and to demonstrate the effectiveness of that process via a field test and then via chain-wide implementation over a six-month time period.

Second, the various papers that use historical data to estimate the impact of store labor on revenue face a major methodological challenge. Given that most retailers set store labor to be proportional to forecasted store revenue, store labor and revenue will always be correlated, and it is hard to observe to what extent labor affects the revenue vs. how much the sales forecast affects labor. A way to circumvent this issue is to focus on random deviations of actual labor payroll from the plan (e.g., due to last minute absenteeism, sickness, etc.), although absent field experiments, the validity of this approach is heretofore unconfirmed. We start with this approach in our analysis of historical data, then follow up with a field test, and finally a large-scale implementation, thus providing the first evidence of the validity of this previously used approach.

Finally, most retailers set store labor at the same level across stores, proportionate to revenue. We show that this is not the best approach because the revenue impact of store labor varies by store. The stores in our study that could benefit from relatively more labor were those with high potential demand, closely located competition for that demand, and experienced store managers. Overall, we provide the first simple but rigorous, field-tested approach that any retailer can use to increase revenue and profitability through better labor management.

2. Literature Review

The literature on retailing spans a variety of topics including store execution and labor planning, pricing, inventory management, branding and assortment planning (see Fisher and Raman (2010) for an overview). The literature that is most germane to our paper studies workforce planning and scheduling within the retail store. We refer to Kesavan and Mani (2015) for a recent detailed review of this literature along with current industry practices.

Several papers use historical data to estimate the impact of store labor on sales. One issue these papers face is endogeneity between sales and store labor, given that store labor levels are typically based on a forecast of sales. Hence, there is always a correlation between sales and store labor, but it is difficult to disentangle the causal impact of store labor on sales from the impact of sales on the amount of labor a retailer assigns to a store. To overcome this issue, most papers use various instruments such as planned labor or lagged labor to identify the causal effect, or they transform variables from labor (which is endogenous to sales) to labor deviations from plan (which is arguably exogenous to sales).

Fisher et al. (2006) is probably the first paper to attempt to measure the impact of the level of labor in a store on sales. They use 17 months of historical data from a retailer with more than 500 stores to estimate the impact of store staffing level on sales, using the deviation of actual labor from planned labor as a shock. They find that the estimated sales impact of a \$1 increase in staffing payroll varies by store, from a low of \$4 to a high of \$28. Even at the stores with an estimated \$4 impact, increasing labor is profitable since this retailer's gross margin was about 50%. Moreover, the retailer could increase sales by 2.6% simply by reallocating labor from stores with a low revenue impact to those with a high revenue impact.

In another interesting working paper Ton (2009) finds that increasing labor at a store results in higher conformance quality and service quality, but only the former improves profitability. Their identification strategy is to use planned profit margins and planned sales calculated by headquarters as instruments. Ton and Huckman (2008) find that employee turnover is associated with decreased performance, as measured by profit margin and customer service. Ton (2014) further argues that, by paying higher wages a retailer, can achieve higher quality outcomes which will ultimately pay the higher wages. In a related stream of work, Lambert et al. (2012) argues that flexible schedules and unpredictable management practices impose negative effects on labor.

Netessine et al. (2010) make a fundamental observation that staffing levels should not be planned using a forecast of sales, because sales itself depends on the staffing level. Instead, sales should be planned using a forecast of traffic through the store. They study a grocery retailer, where few customers, if any, leave empty-handed, so checkout transactions can be used to measure traffic. They analyze the relationship between basket size and both labor planning (how well the labor plan matches store traffic) and labor execution (how well the deployed labor matches the plan). They find that poor labor planning and execution both lead to lower basket values, but labor execution is more impactful, with possible sales increases on the order of 3% through combined improvements in both planning and execution.

Subsequent studies advanced these results by obtaining detailed store traffic data from traffic counting sensors. Perdikaki et al. (2012) decompose sales into the conversion rate and basket size and use daily data to show that store sales have a concave relationship with traffic; conversion rate decreases non-linearly with increasing traffic and labor moderates the impact of traffic on sales. Overall, the paper demonstrates how store traffic impacts store performance, and in particular it shows that the marginal return on adding labor depends critically on the store traffic. Chuang et al. (2016) use data from an apparel retail chain to estimate the sales response to the ratio of labor to traffic. They show that when labor is insufficient, bringing additional traffic does not have a high impact on sales.

Mani et al. (2015) use traffic, sales and labor data to demonstrate systematic understaffing at a chain of retail stores (more than 40% of the time, 33% staffing shortfall on average). They show that understaffing is linked to lower conversion rates and hence a 7.02% decrease in profitability using perfect foresight as a benchmark and a 4.46% decrease with a week-ahead forecast (with no shift-length constraints).

Kesavan et al. (2014b) show that the use of temporary and part-time labor, which is very popular among retailers, can bring flexibility, but too aggressive use of part-time labor reduces profit. The identification strategy involves instruments (unemployment data and lagged labor mix). The latter instrument is commonly used: see, e.g., Siebert and Zubanov (2010) who show that management skills can account for up to 13.9% higher sales per worker in a retail setting.

Musalem et al. (2016) use short in-store videos to show that raising the customer assistance rate by sales associates from 50% to 60% increases conversion by 5% and transactions by 18.5%. They also point out that the assistance rate depends on staffing levels.

To summarize, there is ample empirical evidence that staffing levels should affect sales of a retail store, and there is some evidence that there might be systematic understaffing. There are also various estimates of the economic impact of labor increase on sales which rely on a variety of instrumental variable approaches. What is now clearly needed is a methodology to set staffing levels accordingly and validate these results in practice.

To our knowledge, only one paper thus far attempted to do this. Kesavan et al. (2014a) examine the impact of congestion in fitting rooms on the store performance for a retailer and demonstrate inverted-U relationship between fitting room traffic and sales, i.e., beyond a point, more traffic actually reduces sales because of congestion. They test the impact of labor on sales by increasing fitting room labor by one person in one store for a randomly chosen 4-hour peak traffic block over a 10-day period and show that this intervention has a statistically significant impact on sales. These results are consistent with our hypothesis introduced in Figure 1. We, on the other hand, experiment with the entire sales labor in stores, first through a limited test in 16 treatment stores over 204 days, and then through chain-wide implementation in 168 treatment stores over 182 days.

A number of authors have discussed that labor levels should vary by store, a hypothesis which our work supports on a large scale. For instance, Gauri et al. (2009) discuss that "standardizing performance expectations across different outlets within a chain, differing in their individual features, their consumers, and the nature of competition they face, can be an onerous task." The authors develop a model of store expectations that draws upon the existing trade area as well as store performance literature which is consistent with our approach to thinking of staffing decisions at the local level. Lusch and Serpkenci (1990), investigate the relationship between personal characteristics of the store manager and the store's performance outcomes and show evidence that store performance can vary, driven by the store managers ability and effective use of resources. Kumar and Karande (2000), argue that "Retail stores are segmented using socioeconomic characteristics of the trade area, and it is shown that the effects of store environment on store performance vary across segments" and measured performance by sales and productivity-based measures at the market level. They argue that the internal store environment has an impact on performance which can drive a different need for sales associates at the store level. Reinartz and Kumar (1999) argue that four factors have traditionally been identified in influencing store performance: store-, market-, and consumer characteristics and competition. The authors use structural modeling to assess the differential impact of store attractiveness, market potential, and socio-economic status on store performance measures. Once again, their results are consistent with the idea that retailers could benefit from staffing decisions made at the store level.

3. Empirical Setting

Our retail partner in this research is a U.S. specialty retailer (these type of retail businesses focus on specific product categories, such as office supplies, hardware, and tools, and household goods) with more than 700 stores and over $2 billion in annual revenue. They had been collecting data for some time on sales, staffing and a variety of factors influencing sales, but had not found any way to extract useful insights to improve their store execution. Management contacted us in 2012 for assistance, and it soon became apparent that the most impactful store execution lever was the level of staffing in each store at each point in time. They segmented store labor into selling labor (associates who directly interacted with customers to assist them and answer questions) and non-selling labor (associates who performed various tasks within the store, such as restocking shelves or receiving deliveries). There was a process to set a budget for each store for the next month for these two labor categories.

The non-selling labor budget was based on the planned tasks to be accomplished in the store next month and an estimate of the labor required for each task. The payroll budget for selling labor was set in each store to approximately 10% of forecasted revenue for that store for the next month. Allocation of the monthly payroll budget by day and department within the store was the responsibility of the store manager. The retailer was comfortable with their process for planning non-selling labor but wondered if there were other approaches to setting the selling labor budget that might result in higher sales and greater profit.

With this focus in mind, we sought to deploy the process described in the introduction. Table 2 shows those variables for which we received values for each week from August 2011 through September 2013. Table 3 contains store attributes that tended not to vary over the period under analysis and included physical attributes of the store, as well as its customers, competition and management. For these store-level variables we received single numbers giving their value for each store as of April 2014. Among other things, Table 3 includes the number of five specific competitors within 10 miles of a given store, Walmart and four specialty retailers that were direct competitors and who are not named so as to protect the identity of our subject retailer. The correlation of the variables described on Tables 2 and 3 are included in Tables 4 and 5 respectively.

Like a majority of retailers, our subject retailer did not compile store traffic data, so this information was not available to us. The retailer's management was actively involved throughout this project, reviewing our results for reasonableness and providing useful insights.

4. Analysis of Historical Data

In order to understand the impact of labor on sales and whether the staffing level was appropriate at the chain level, we first estimated the model below, controlling for marketing activity, type of transaction, seasonality, and a store fixed effect,

$$\begin{aligned} Log\,Revenue_{it} = \alpha_0 + \beta_1 \cdot \; Log\,MKT_{it} + \beta_2 \cdot \; Transaction\,Type_{it} + \\ + \beta_3 \cdot \; Payroll\,Deviation_{it} + \delta_t \cdot \; Week_t + \phi_i \cdot \; Store_i + \epsilon_{it}, \end{aligned} \tag{1}$$

where for all stores i and weeks t, $Revenue_{it}$ is dollar revenue for a store, MKT_{it} represent three different marketing variables sent from store i on week t: $Email$, $Mail$ and $Inserts$. Table 4 shows that $Email$ and $Mail$ are highly correlated, so to avoid collinearity issues we exclude the $Email$ variable from the analysis[1]. The results are almost identical if we include the $Mail$ variable. $Transaction\,Type_{it}$ captures the percentage of transactions that used any kind of discounting (which includes the sum of variables $PromoPer$, $CpPer$ and $PromoCpPer$) at store i on week t. Finally, $Payroll\,Deviation_{it}$ is defined as follows:

$$Payroll\,Deviation_{it} = \frac{ActSell_{it} - PlanSell_{it}}{PlanSell_{it}}.$$

The coefficient β_3 of this variable will essentially give us the first derivative of the Revenue-Labor curve in Figure 1. $Week_t$ and $Store_i$ denote seasonality and store fixed effects, and ϵ_{it} is an error term.

The results of this regression are presented in Table 6. The first column of this Table presents the results for the model including only the $Week_t$ and $Store_i$ fixed effects and $Payroll\,Deviation_{it}$, our variable of interest. We validated that, in our context, $Payroll\,Deviation$ varies randomly by regressing $Payroll\,Deviation$ and our variables of interest. We did not find any statistically significant correlations from this analysis. It is possible to think of real-world circumstances under which the deviations between planned and actual labor are not random. This would happen if for example: employees decide to be absent in anticipation of slow traffic and/or sales, a store manager discouraged sales staff to come to work (or encourages them to leave early) when sales are slow, or store managers can quickly increase (decrease) staffing levels when traffic increases (decreases) during a day.

[1] We log transform the main variables in the analysis to correct the skewness. In addition, this gives us a better model fit and better capture the multiplicative effect of these factors.

However, our approach does not require that this variation is entirely random but that it is a useful shock which helps us estimate the impact of sales staff on revenues while avoiding a substantial endogeneity bias. In other words, even when this variation might not be entirely random, it can still be effectively used to optimize staffing levels at individual stores. For this reason, validating the hypothesis (that the variation is likely exogenous) with a large-scale implementation in the field is relevant.

We can observe that the estimated coefficient for this variable is 0.141 and it is statistically significant. We can interpret the coefficient of 0.141 on *Payroll Deviation* as implying that a 10% increase in payroll for the chain would produce a 1.41% increase in revenue.

The second column of Table 6 shows the results when we include the additional controls corresponding to the different marketing actions. The overall impact of the marketing actions is positive. Column 3 present a model where, in addition to the marketing variables, we include the proportion of transactions in the store on a particular week that used a promotion, a coupon, or both. The estimate of *Payroll Deviation* is consistent across all three models.

Finally, in column 4 we estimate a model where we include *Payroll Deviation* squared as an additional variable since our hypothesized relationship is concave increasing. As expected, this analysis seems to indicate that the increase in payroll has a concave increasing relationship with revenue which further confirms that our model is a good representation of the dynamics between payroll and revenue. Figure 2 presents the predicted values and the 95% confidence for standardized sales at different Payroll Deviation levels when we consider this last model. We note that the curvature is minimal, and for all practical purposes we can ignore quadratic terms.

4.1. Sales Variation vs Payroll Deviation

We did not have access to a sales forecast from the retailer. However, to further validate our approach we considered two alternative specifications where we generated a sales forecast to compare it with the actual sales observed.

We generated the forecast by considering the data from the fiscal year 2012 and from that point we generated a rolling forecast for two weeks out at the store-week level. By generating this forecast, we replicate what the retailer does. Specifically, we considered the following two models to generate forecasts.

Model 1:

$$\begin{aligned} Log\,Revenue_{it} = \alpha_0 + \beta_1 \cdot \; Log\,MKT_{it_{-2}} + \beta_2 \cdot \; Transaction\,Type_{it_{-2}} + \\ + \beta_3 \cdot \; Payroll\,Deviation_{it_{-2}} + \delta_t \cdot \; Week_t + \phi_i \cdot \; Store_i + \epsilon_{it}, \end{aligned} \tag{2}$$

Model 2:

$$\begin{aligned} Log\,Revenue_{it} = \alpha_0 + \alpha_1 \cdot \; Log\,Revenue_{it_{-2}} + \alpha_2 \cdot \; Log\,Revenue_{it_{-3}} + \\ + \beta_1 \cdot \; Log\,MKT_{it_{-2}} + \beta_2 \cdot \; Transaction\,Type_{it_{-2}} + \\ \beta_3 \cdot \; Payroll\,Deviation_{it_{-2}} + \delta_t \cdot \; Week_t + \phi_i \cdot \; Store_i + \epsilon_{it}, \end{aligned} \tag{3}$$

The first model includes all the dependent variables we used in our analysis with a lag of two weeks. The second model, in addition to the dependent variables used in the first model, includes the two- and three-week lagged revenue.

In the estimation we considered the actual sales data from the fiscal year 2012 to predict the second fiscal week of 2013 and subsequently we add one more week of actual historical sales to generate the forecast two weeks out. We then estimate the model presented in equation 1 of the paper where we considered the weekly sales forecast error in percentage terms with respect to payroll deviation, and the additional variables, for each of the two models. The result of this analysis is presented in Tables 7 and 8. Note that the sign change for the variables presented in Table 6, Tables 7, and 8 is driven by the high correlation between the variables $Log(Mail)$ and $Log(Insert)$, and $PromoPer$ and $CpPer$. The different columns correspond to the same variations presented in Table 6.

Although the results from this analysis are not directly comparable to the ones presented on Table 6, they show a high level of consistency. This is not a surprise since the sales forecast two-weeks out is a very good predictor of the actual sales. The average R^2 when implementing the first and second models are 0.96 and 0.97, respectively.

4.2. Individual Store Variation

As existing literature argues (see papers in the literature survey section), payroll deviations are typically random, driven by weather, home emergencies, sickness of employees, traffic disruptions, overbooking by store manager, inflexibility of temporary employees, among others. and therefore it is plausible that the relationship between payroll deviations and sales is causal.

We can then consider whether it is profitable to increase store selling labor at the chain level. For this retailer, the gross margin earned on sales was 50% and the selling labor budget was set at about 10% of revenue across the chain. Consider for illustration purposes a store with $100,000 revenue in a month. The payroll budget would be $10,000 so a 10% increase in payroll costs is $1,000 and—according to our results of column 3 in Table 6 (we considered this value because the function is pretty flat)—produces incremental revenue and gross margin respectively of $1,380 and $690 respectively. Since the incremental margin of $690 is less than the incremental payroll cost of $1,000, this is not profitable, for an average store (in other words, at the chain level stores do not appear to be understaffed, on average).

However, this retailer has more than 700 stores, and we suspected that the impact would vary considerably by store. To assess this variation, we extended the model to allow for store-specific coefficients on labor deviations:

$$\begin{aligned} Log\,Revenue_{it} = \alpha_0 + \beta_1 \cdot\ Log\,MKT_{it} + \beta_2 \cdot\ Transaction\,Type_{it} + \\ + \beta_{3i} \cdot\ Payroll\,Deviation_{it} + \delta_t \cdot\ Week_t + \epsilon_{it}. \end{aligned} \tag{4}$$

Note that in this new specification we did not include the individual store fixed effects ($Store_i$). With two years of historical weekly data, we have more than 100 observations of $Payroll\,Deviation_{it}$ and associated sales applicable to estimating β_{3i}, the impact of payroll deviation at the store level.

There are two main alternatives to estimate the model in equation 4. The first alternative is to evaluate the model n times where each estimation includes only the observations corresponding to store i. By doing this we would obtain n different coefficients for β_{3i}, each one corresponding to a different store i. Because we are interested in comparing the individual coefficients β_{3i} with respect to each other, we use an alternative approach and estimate one model where we obtain the individual coefficients β_{3i} by including the interaction term between $Payroll\,Deviation_{it}$ and the specific store i ($Payroll\,Deviation_{it} \cdot Store_i$).

Figure 3 illustrates the results of this analysis, where the horizontal axis gives values of β_{3i} and the vertical axis is the percentage of stores with that value. The dots in Figure 3 correspond to the count of stores within that estimate range where β_{3i} is statistically significant. Note that a few of the β_{3i} coefficients are negative. While a store may be so highly staffed that additional labor would have minimal impact on sales, it is not plausible that adding labor would decrease sales. In fact, for the stores showing negative β_{3i} values, the actual β_{3i} value is at or near 0 given the standard error of the estimate, so we are probably observing negative values as a result of statistical noise. In support of this conclusion, most of the 95% confidence intervals for the negative coefficients include positive values. Only 3 out of 710 stores have a negative upper 95% confidence limits. We can expect 5%, or 35 stores out of 710, to be outside of 95% confidence limits, and we think these 3 stores are part of those 35 stores.

4.3. Factors Explaining the Store's Variation

Figure 3 shows large variation in the impact that additional labor has on store revenue. As a post-hoc analysis and to understand the store attributes that might explain this variation, we regressed the β_{3i} values against the store attributes in Tables 2 and 3 using lasso regression, step-wise regression, logit regression and discriminatory analysis.

4.3.1. Lasso Regression The least absolute shrinkage and selection operator (lasso) model, proposed by Tibshirani (1996), is a variable selection regression. This process selects variables by penalizing the absolute size of the regression coefficients so that the weak parameter estimates are shrunk towards zero, effectively dropping them. For a linear regression $y = \alpha_0 + \sum_j \beta_j x_j$, the objective is to minimize $\sum_i \left(y_i - \alpha_0 + \sum_j \beta_j x_i j\right)^2 + \lambda \sum_j |\beta_j|$, where λ is the hyperparameter that determines how much the estimates are penalized. The value of λ is estimated in the cross-validation step—see Friedman et al. (2001, p.68) for details. For the time-varying attributes in Table 2 we used their average value over the period under analysis.

The results of this regression are presented in Table 9, Column 1, where the estimation was performed following Reid et al. (2016) to ensure valid post-selection inference. The statistically significant correlates of β_{3i} values are the percentage of transactions with coupons, the average basket size, the years since the store has been remodeled, the Household score[2], the years of service of the store manager, and the number of Competitor 3 and Walmart stores within 10 miles. Competitor 3 is a large specialty retailer that competes head to head with our subject retailer.

[2] The variable $HouseholdScore_{100}$ is House Score scaled by dividing by 100 to make its coefficient of reasonable size.

We can consider why these store attributes make sense as drivers of high potential for a store to increase labor. Average basket size and Household score are indicators of market size, so we can interpret this as the larger the market potential, the greater the value of having sufficient labor to provide excellent customer service. This result is consistent with Perdikaki et al. (2012), who found that the marginal return of adding labor depended on the level of store traffic. More competition for that high market potential would also increase the need for a high level of customer service. If a store is in an isolated area with no competitors, then customers have no alternatives to shopping at that store, and are forced to tolerate what might be a less than ideal level of customer service. But if there are one or more competitors, then customers have choices, and if they find themselves waiting a long time for a store sales associate to help, they will be inclined to exit the store and go to a competitor store. Finally, the more years of service a store manager has, the more likely he or she will be to make effective use of additional labor.

4.3.2. Step-wise Regression To ensure robustness of our results, in addition to the lasso regression we apply a step-wise regression approach with the same set of variables. The results are presented in column 2 of Table 9 and they are largely consistent with the outcome of the lasso estimation.

4.3.3. Logit Regression A closely related approach to the Discriminatory Analysis and a more traditional regression approach is to use a logit model where the dependent variable is a dummy that indicates whether the store is above the threshold or not. We implemented this analysis as an additional validation. Columns 3 and 4 of Table 9 present the results when we implement the logit model with the variables selected with the lasso and the step wise approach respectively.

4.3.4. Discriminatory Analysis To further explore the factors explaining the store variation we implemented discriminant analysis. Discriminant analysis can be used for two main purposes: group separation and prediction or allocation of observations into groups (Rencher and Christensen (2012)). In our case, we implement the analysis with the first purpose in mind since in our data we know which group each store belongs to so we are not concerned about a situation in which some stores membership is unknown and needs to be classified. One of the challenges with discriminant analysis is that the interpretation of the resulting function is very much limited in determining the contribution of each variable.

Hence, when this method is used the concern is to understand the relative contribution of the different variables rather than interpreting the magnitude of the effect of any specific variable (Rencher and Christensen (2012)).

Columns 5 and 6 of Table 9 show the results when we implement the discriminant analysis and we report the standardized canonical discriminant coefficients.

4.3.5. Summary The implications and relevance, both from the magnitude and significance perspective, is similar across the different models. We followed Rencher and Christensen (2012) to determine the classification rate matrix for the discriminant analysis. The results of these analyses are available upon request. The analysis of this matrix shows that there is no one model that dominates the others. These results emphasize the challenge to correctly identify understaffed stores from observable characteristics.

It is interesting to note the relatively low R^2 of this analysis. This indicates that, although there are variables that are correlated with stores being understaffed, determining which stores need to adjust their staffing levels is hard based on observable store characteristics, making our methodology even more relevant.

4.4. Profit impact of Adding Labor

Note that a β_{3i} value of 0.2 is approximately the threshold of profitability for adding labor to a store, since adding 10% more selling labor to our illustrative store with \$100,000 in revenue and a coefficient of 0.2 would cost \$1,000 and it would add \$2,000 and \$1,000, respectively, to revenue and gross margin, thus producing a net profit increase of zero. Figure 3 suggests an opportunity for improving performance by increasing labor for those stores with β_{3i} greater than 0.2, thereby profitably increasing revenue, and reducing labor for those stores with β_{3i} less than 0.2, saving more labor cost than the revenue and margin lost. In the next section we describe a test to explore this idea.

5. Store Test to Validate Results

The analysis above shows potential improvements from increasing labor in some stores and reducing it in others. Our subject retailer was much more interested in growing revenue than in saving labor cost and hence we designed with them a test to validate the potential for increasing revenue using additional labor.

Management hand-selected 16 stores that, according to the analysis presented in the previous section, should have a substantial revenue and profit increase as a result of labor increase.

To evaluate the impact of this intervention, we generated a matched control group intended to resemble the characteristics of the 16 test stores. We implemented a nearest neighbor algorithm with full Mahalanobis matching which is a type of propensity score matching. We match each one of the selected stores with the best 3 neighbors. Following the retailer's recommendation, we matched stores on four main features: total sales, total sales staff hours, number of competitors in the stores area of influence and store square footage. The matching was done with the average weekly data from the previous fiscal year. We allow a particular matching store to be a neighbor of more than one test store so the matching process resulted in, not 48, but 45 matched control stores and a total sample of 61 stores. Table 10 shows the values of the four matching attributes for the 16 "treatment" stores, all other stores in the chain (in rows labeled U) and the 45 matched control stores. As can be seen, the matching process has created a control sample that is statistically similar in these four attributes to the treatment stores.

The test ran for the 26-week period lasting from October 11, 2013 to May 3, 2014. Each test store was allocated 32 extra selling hours per week, in addition to the number of selling hours assigned by the system. The 32 additional hours was an amount selected by the retailer based on a discussion of our results with them, and it was intended to balance the cost of the test, due to the extra payroll, with the need for an increase large enough to detect a revenue impact. Moreover, as our estimates in the previous section indicate, the revenue-to-labor relationship is close to linear in the regions stores were operating so the 32 hour labor increase was small enough that our regression model could be expected to accurately predict the change in revenue. Following the retailer's standard procedure, the store managers at the test stores received the information on how many hours they were allowed to allocate and assigned hours to days within the week at their discretion. They neither knew that the hours assigned were more than the ones the retailer's system suggested nor that their stores were part of a test. In particular, the scheduling system used by the retailer has the possibility of labeling different types of hours assigned to the budget for a particular store. During both the test and the implementation we discuss in the following section, the selected stores were receiving the additional hours under the corporate hours label. This labeling has been used by the retailer before and it was not specifically created for our test. Using this labeling, we could track the level of compliance with the test at the daily level.

During the test we tracked the number of hours assigned each day in each store, as well as the level of revenues. When the test was completed, we used these results to fit the model below for all stores i and days t.

$$Log\,Revenue_{it} = \alpha_0 + \beta_1 \cdot\; Test\,Mag_{it} + \delta_t \cdot\; Day_t + \phi_i \cdot\; Store_i + \epsilon_{it}, \tag{5}$$

where for all i and t, $Test\,Mag_{it}$ equaled:

$$Test\,Mag_{it} = \frac{Actual\,Test\,Hours_{it}}{Planned\,Hours_{it}}.$$

Day_t is a seasonality control, $Store_i$ is a store fixed effect and ϵ_{it} is an error term. The model was fit on data from 16 test stores and the 45 control stores, with $TestMag_{it}$ equal to zero for the control stores. A total of 9,938 additional hours were used by the store managers during the test, which corresponds to 75% of the total test hours available to them.

Regression results are provided in Table 11. The coefficient of 1.346 on $TestMag$ implies that, for example, a 10% selling labor increase would produce a 13.46% revenue increase.

To further validate the results of the test, following the retailer's request we implemented an additional analysis where we include the data from the previous year for the test and control group. This analysis resembles what is a standard metric in the industry by looking at comparable sales across stores. From the modelling perspective, by including the same period from the previous year for the test and control stores, the analysis resembles a difference in difference approach. The results of this alternative analysis is presented in the column 2 of Table 11, indicating a smaller sales lift of about 10%.

We would like to emphasize the fact that this test was not a randomized experiment since the 16 test stores were hand-picked by the company's management and therefore they did not resemble a true randomized experiment even after matching them to comparable stores. This is, however, precisely what we wanted to obtain in our context. We wanted to validate that our model can accurately detect understaffed stores that can generate a profit when assigned additional labor hours.

The store test was a success in confirming that our methodology can identify stores that can benefit from additional staffing. It also showed the retailer that for these stores, increasing labor was highly profitable. As a result, the retailer decided to implement our methodology and expand the labor increase to a total of 168 stores. Although there were a total of 197 stores that, according to our model, were above the break-even threshold of 0.2, the retailer cautiously decided to add labor only to those stores that presented the opportunity to substantially increase profits and not simply to break-even.

6. Company-wide Implementation

For the set of 168 stores with high β_{3i} coefficients the retailer committed to adding 10% more selling staff hours to the store labor budget for each week. According to our model estimation, when selling staff hours were incremented by 10%, the predicted average revenue increase for the 168 test stores was 4.07%.

We monitored results of the implementation for a total of 26 weeks (182 days) for the period from November 2, 2014 to May 3, 2015. Similar to what we described for the test, store managers determined when within a week to use the additional hours, and indeed also how much of the allowed additional hours to use. For each day and store we knew the number of additional hours used, denoted $AdditonalHours_{it}$.

To evaluate impact, we generated a matched control group to resemble the characteristics of the 168 stores. As before, we implemented Mahalanobis nearest neighbor algorithm to match each one of the selected stores with the best 3 neighbors and we allow a particular matching store to be a neighbor of more than one test store. Following the retailer's recommendation, we matched stores on the same four features considered during the test: total sales, total sales staff hours, number of competitors in the stores area of influence and store square footage. The Matching was done with data from fiscal week 19 to fiscal week 39 of 2015. The matching process resulted in a total sample of 672 stores.

Table 12 shows the values of the four matching attributes for the 168 "treatment" stores, all other stores in the chain (in rows labeled U) and the 504 matched control stores. As can be seen, once again, the matching process has resulted in a control sample that is statistically similar along these four attributes to the treatment stores.

At the end of the evaluation period, we used the results to fit the model below for all stores i and days t:

$$Log\,Revenue_{it} = \alpha_0 + \beta_1 \cdot \ Mag_{it} + \delta_t \cdot \ Day_t + \phi_i \cdot \ Store_i + \epsilon_{it}, \tag{6}$$

where for all i and t, Mag_{it} equaled:

$$Mag_{it} = \frac{AdditonalHours_{it}}{PlanHours_{it}}.$$

Day_t is a seasonality control, $Store_i$ a store fixed effect and ϵ_{it} an error term. The model was fit on data from the 168 treatment stores and the 504 control stores, with Mag_{it} equal to zero for the control stores. The result for the implementation analysis is included on Table 13. Column 1 present the results when we include only the $TestMag$ as a variable without any additional controls. We can observe that the Test stores had a positive revenue impact from the additional labor hours, and the analysis shows a statistically significant coefficient of 0.582 for the $TestMag$ coefficient. The second and third column in Table 13 show the results when we include the seasonal controls and the seasonal controls and stores controls respectively. Finally, column 4 shows the results when we include $TestMag$ and $TestMag^2$ as independent variables. This analysis indicated that, as we have hypothesized, the relationship between labor hours and sales is concave. The linear model suggests that the retailer can enjoy a 4.5% sales increase when the sales staff is increased by 10%, a number close to our predicted 4.07%.

Figure 4 presents the predicted values, and confidence intervals, for the models presented in columns 3 and 4 of Table 13. We can observe that the linear and quadratic models present similar outcomes when the labor change is small. However, considering a quadratic specification would become more relevant when trying to explain the impact of larger labor variations.

It is important to note that one underlying assumption in our analysis is that SUTVA "stable unit treatment value assumptions" holds. In our context, this is a reasonable assumption since the execution of the field validation was done in such a way that other stores were not aware of the changes being implemented. This means that we have no reason to believe that the additional labor assigned to store i would impact other treated stores. In other words, the effect observed at store i is driven only by the additional labor at store i_1 and not affected by what occurred at a different store i_2. We assume that in our case the assumption of a non-contamination scenario is plausible. We base this assumption on the fact that our setting is similar to those settings described in the literature where it is reasonable to expect non-contamination. We refer the reader to Rosenbaum (2002) for a more detail discussion on SUTVA.

6.1. Store Level Analysis of the Implementation

Similarly to what we saw in Section 4 when we analyzed historical data, we observe store-specific variations in the implementation. The first histogram in Figure 5(a) summarizes the result of this analysis. As expected, the impact of the implementation is centered around a 5% increase at the store level. In addition, we then calculated the difference between the estimates from our historical model and the estimates from the implementation for the 168 test stores. The second histogram on Figure 5(b) shows the individual differences (Implementation Estimate minus Model Estimate). A successful outcome for the retailer is when the actual revenue increase from adding labor is at least as great as we had predicted from our analysis of historical data. This happens in 155 out of the 168 stores. The distribution of the variation of the estimates is remarkably tight. We can see that, consistent with the results of the individual estimates, these differences tend to be slightly positive. This indicates that, if anything, the model underestimated the impact of additional labor.

6.2. Profitability Analysis

We now evaluate the impact on profits of the implemented labor increases. The implementation involved a total of 201,512 additional hours at a cost of \$2.24 million, an 8.27% increase over standard practice. Given our analysis presented in column 4 of Table 13, we can estimate that the added labor increased sales by 4.5%. Total sales for the 168 stores during the evaluation period was \$276.84 million. A 4.5% sales increase implies that sales without the additional labor hours would have been 276.84/1.045 = \$264.92 million, and that the additional labor added \$11.92 million to sales. The retailer's gross margin is 50%, so these incremental sales generated \$5.96 million incremental margin, for a \$3.72 million profit after deducting the \$2.24 million cost of the incremental labor. The \$3.72 million profit over the 26-week test extrapolates to an \$7.44 million annual benefit. This number was significant for the retailer and, not surprisingly, they are continuing to use the additional hours in the 168 stores while considering other stores to which these results might be extended.

6.3. Additional Evidence: Benchmark with an Arbitrary Policy

As noted, the results presented so far were not based on a randomized experiment. We have a chance to at least partially redress this shortcoming by analyzing an action our retail partner took that was incidental to our work with them. The management team of our retail partner, starting on February 2 2013, had implemented a policy that no store could be assigned fewer than 250 hours per week. The details of the policy were as follows: every time the staffing system allocated fewer than 250 hours for a store-week, the number of hours assigned was increased to a minimum of 250. There were a total of 246 stores affected by this policy.

This policy, although not entirely random, provided us with the opportunity to measure the impact of a somewhat arbitrary policy that was not based on our analysis of historical data. Moreover, the average coefficient β_{3i} for the 246 stores was 0.123, a value below the threshold of profitability for increasing labor, and below the average coefficient for all stores. This allows us to answer the question: we have shown that adding labor to stores with high β_{3i} values produces the predicted revenue increase, but if we add labor to stores with low β_{3i} values, do we similarly obtain a low revenue increase?

To answer this question we follow a similar process to the one we presented in Section 4. We first considered the actual hours assigned by the system and the additional hours that were required to reach the 250 hours threshold. Then we calculate the percentage change that these additional hours implied for each store-week combination (we called this the $TestMagnitude$). The period of analysis we considered was the same we used for the analysis of our historical data. This analysis was done at the weekly level since this was the level of aggregation available to us. From this analysis we obtained an estimate of β_1 equal to 0.136 with a standard error of 0.019. This result compares well with the predicted value of 0.123. Also, if we compare this result with the results obtained in the first part of this section we can see that the implication of following our analysis when choosing which stores to increase the sales labor gave an estimate for β_1 of 0.451, more than three time larger than in the case when the stores where selected arbitrarily. Moreover, when we look at the individual estimates of the 246 stores, following the same approach we used earlier, we found that only 48% of the stores actually had a revenue increase that justified the investment. As mentioned earlier, this number was 92% for the 168 stores where we tested our model.

7. Conclusions

We have described a process by which a retailer can increase revenue by refining their allocation of sales staff by store. The process starts with analysis of historical data, then continues with testing to confirm results and concludes with eventual chain-wide implementation. We demonstrated the effectiveness of this approach through deployment with a large specialty retailer and found that our approach produced a 4.5% revenue increase at the impacted stores and a nearly $7.4 million profit increase.

Clearly, our results indicate only a lower-bound on what can be achieved through this methodology because we were limited in our interventions by what management of the company was prepared to do. For a more impactful implementation we would want to vary the amount of labor added by store, potentially subtracting labor in some stores. Moreover, if finer grained data was available, we would have wanted to also consider labor allocation by hour or by department rather than relying on store managers to do so. Nevertheless, given the dire state of the retail industry, even our current results appear to be economically important.

Our process revealed significant variation across stores in the impact that staffing had on revenue. The stores that could benefit from relatively more labor were those with the highest potential demand, as indicated by average basket size, number of households and household growth, the greatest competition, and the most experienced store managers.

An implication of this finding is that, contrary to common practice, store staffing levels should not be set to the same level across all stores, proportional to revenue, but to varying levels dependent on the impact store sales associates can have on revenue at each store.

Our results further demonstrate that the relationship between revenue and labor is concave increasing and, since different stores are located at different points on this curve, it is important to estimate the response that sales has to changing labor, so the correct tradeoff can be established.

In this paper we rely on very basic data that is available to perhaps any retailer. Of course, as technology penetrates retail companies, one can add data from traffic counters as in Kesavan and Mani (2015), movement sensors, or video cameras as in Musalem et al. (2016) etc. All of this data can be potentially used to further improve recommendations regarding labor allocation and utilization. Another important aspect of this equation is making sure that labor is adequately trained, which is not easy to achieve in retail setting where turnover is high and salaries are low, so cost of training is significant, see Fisher et al. (2016) for related analysis.

References

Cachon, Gerard, Christian Terwiesch. 2009. *Matching Supply with Demand: An Introduction to Operations Management*. McGraw-Hill Irwin, NY.

Chuang, Howard Hao-Chun, Rogelio Oliva, Olga Perdikaki. 2016. Traffic-based labor planning in retail stores. *Production and Operations Management* **25**(1) 96–113.

Fisher, Marshall, Santiago Gallino, Serguei Netessine. 2016. Does online learning work in retail? *Working Paper - Available on SSRN* .

Fisher, Marshall, Vishal Gaur, Herb Kleinberger. 2017. Curing the addiction to growth. *Harvard Business Review* **95**(1) 66–74.

Fisher, Marshall, Jayanth Krishnan, Serguei Netessine. 2006. Retail store execution: An empirical study. *Working Paper - Available at SSRN* .

Fisher, Marshall, Ananth Raman. 2010. *The New Science of Retailing: How Analytics are Transforming the Supply Chain and Improving Performance*. Harvard Business Review Press.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani. 2001. *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin.

Gauri, Dinesh Kumar, Janos Gabor Pauler, Minakshi Trivedi. 2009. Benchmarking performance in retail chains: an integrated approach. *Marketing Science* **28**(3) 502–515.

Kesavan, Saravanan, Vinayak Deshpande, Hyun Seok Lee. 2014a. Increasing sales by managing congestion in self-service environments: Evidence from a field experiment. *Working Paper - Available at SSRN* .

Kesavan, Saravanan, Vidya Mani. 2015. An overview of industry practice and empirical research in retail workforce management. *Retail Supply Chain Management*. Springer, 113–145.

Kesavan, Saravanan, Bradley R. Staats, Wendell Gilland. 2014b. Volume flexibility in services: The costs and benefits of flexible labor resources. *Management Science* **60**(8) 1884–1906.

Kumar, Vipin, Kiran Karande. 2000. The effect of retail store environment on retailer performance. *Journal of business research* **49**(2) 167–181.

Lambert, Susan J, Anna Haley-Lock, Julia R Henly. 2012. Schedule flexibility in hourly jobs: Unanticipated consequences and promising directions. *Community, Work & Family* **15**(3) 293–315.

Loeb, Walter. 2017. These 21 retailers are closing 3,591 stores—who is next? *Forbes* .

Lusch, Robert F, Ray R Serpkenci. 1990. Personal differences, job tension, job outcomes, and store performance: A study of retail store managers. *The Journal of Marketing* 85–101.

Mani, Vidya, Saravanan Kesavan, Jayashankar M. Swaminathan. 2015. Estimating the impact of understaffing on sales and profitability in retail stores. *Production and Operations Management* **24**(2) 201–218.

Musalem, Andres, Marcelo Olivares, Ariel Schilkrut. 2016. Retail in high definition: Monitoring customer assistance through video analytics. *Columbia Business School Research Paper No. 15-73. Available at SSRN* .

Netessine, Serguei, Marshall Fisher, Jayanth Krishnan. 2010. Labor planning, execution, and retail store performance: An exploratory investigation. *Working Paper, available at SSRN* .

Perdikaki, Olga, Saravanan Kesavan, Jayashankar M. Swaminathan. 2012. Effect of traffic on sales and conversion rates of retail stores. *Manufacturing & Service Operations Management* **14**(1) 145–162.

Reid, Stephen, Robert Tibshirani, Jerome Friedman. 2016. A study of error variance estimation in lasso regression. *Statistica Sinica* 35–67.

Reinartz, Werner J, V Kumar. 1999. Store-, market-, and consumer-characteristics: The drivers of store performance. *Marketing Letters* **10**(1) 5–23.

Rencher, Alvin C, William F Christensen. 2012. Chapter 10, multivariate regression–section 10.1, introduction. *Methods of Multivariate Analysis, Wiley Series in Probability and Statistics* **709** 19.

Rosenbaum, Paul R. 2002. Observational studies. *Observational Studies*. Springer, 1–17.

Siebert, W Stanley, Nikolay Zubanov. 2010. Management economics in a large retail company. *Management Science* **56**(8) 1398–1414.

The Economist. 2017. Sorry, we're closed. *The Economist* 61–63.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Ton, Zeynep. 2009. The effect of labor on profitability: The role of quality. *Working Paper, Harvard Business School* .

Ton, Zeynep. 2011. Why "good jobs" are good for retailers. *Harvard Business Review* **90**(1-2) 124–31.

Ton, Zeynep. 2014. *The good jobs strategy: How the smartest companies invest in employees to lower costs and boost profits*. Houghton Mifflin Harcourt.

Ton, Zeynep, Robert S. Huckman. 2008. Managing the impact of employee turnover on performance: The role of process conformance. *Organization Science* **19**(1) 56–68.

Appendix. Tables and Figures

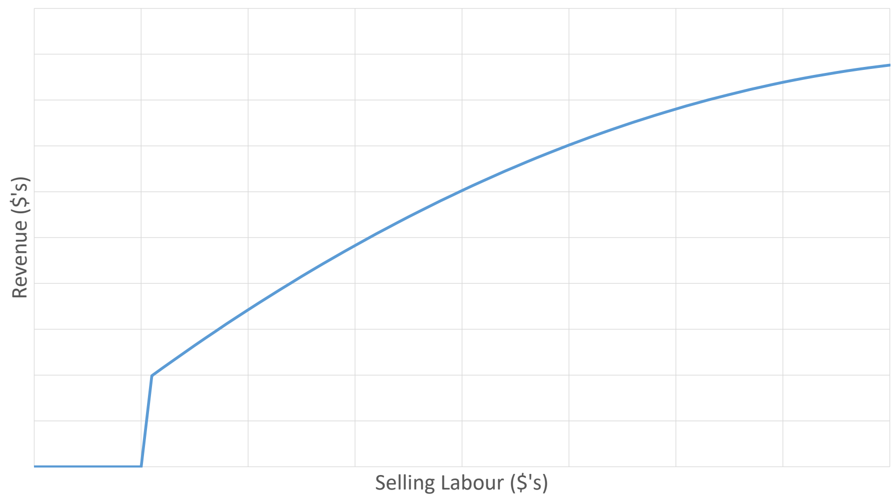


Figure 1 Hypothesized Relationship Between Revenue and Staffing Level

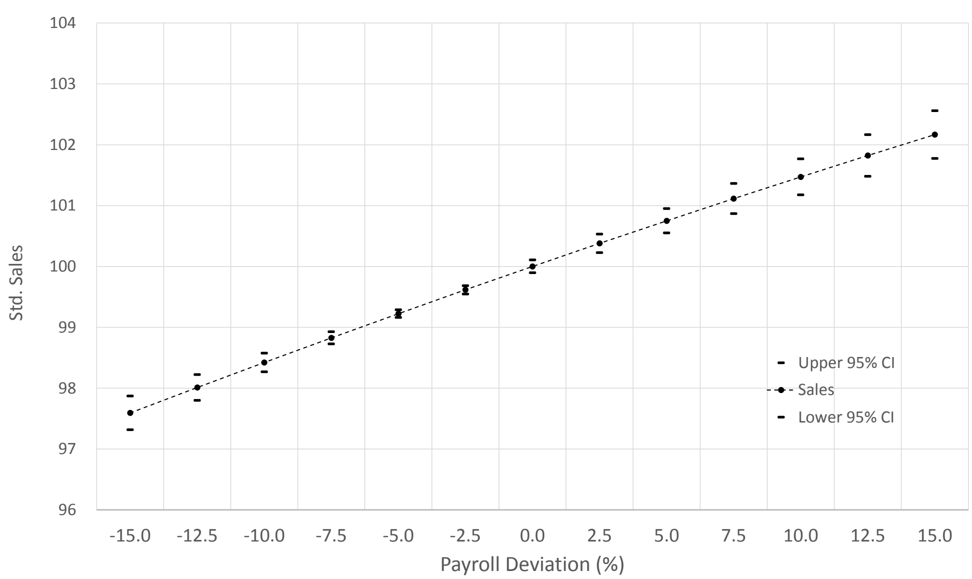


Figure 2 Predicted Values for Standardized Sales wrt Labor Change

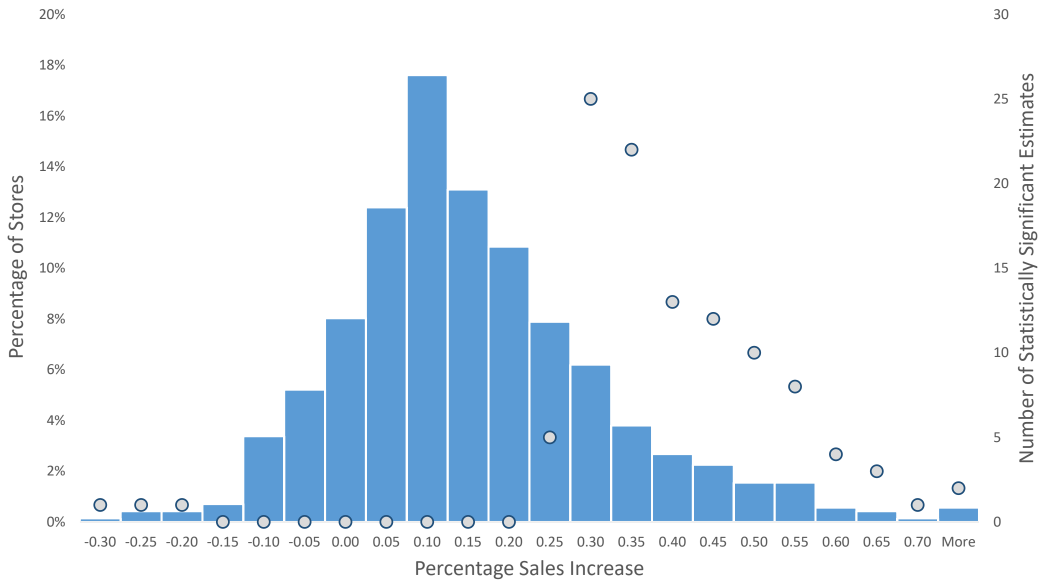


Figure 3 **Distribution of Sales Lift by Store**

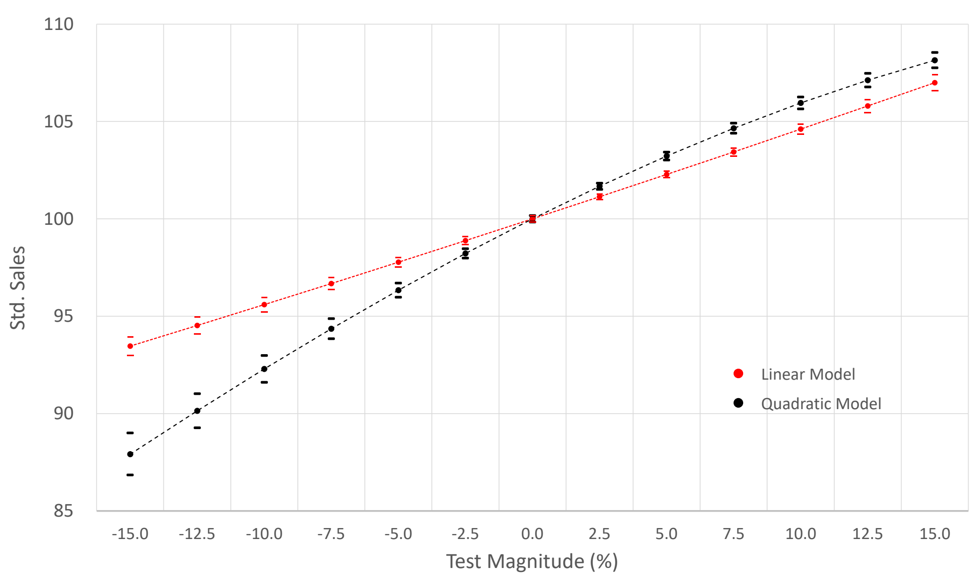
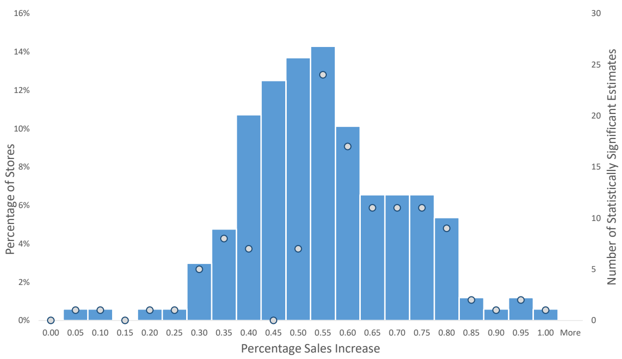
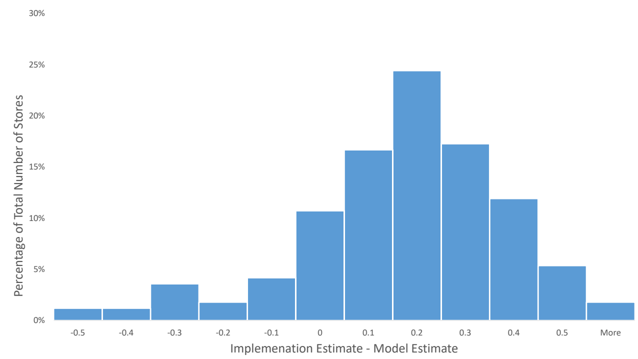


Figure 4 **Implementation Predicted Values for Linear and Quadratic Model**



(a) Historical Data



(b) Model Deviation

Figure 5 168 Implementation Validation

Table 1 Analyses Timeline

	Stores	Unit of Analysis	Periods	Date Range
Historical Data	710	Store - Week	112 Weeks	08/07/2011 to 09/22/2013
Test	61	Store - Day	204 Days	10/11/2013 to 05/03/2014
Implementation	672	Store - Day	182 Days	11/02/2014 to 05/03/2015

Table 2 Summary Statistics for Store-Week Variables

Name	Description	Units	Mean	St. Dev.	Median
Revenue	Revenue	dollars	55,872	34,375	45,489
Units	Units sold	count	25,017	16,880	20,464
GM	Total gross margin	dollars	28,159	19,953	22,203
PlanSell	Planned selling labor budget	dollars	5,895	3,012	5,026
ActSell	Actual selling labor	dollars	5,697	2,921	4,835
PayrollDev	Payroll Deviation between Actual and Planned	%	-2.72	12.34	-3.91
Transactions	Customer transactions	count	4,225	2,328	3,582
Basket	Basket size: equals average number of units purchased in a given customer transaction	count	5.90	2.19	5.79
PromoTrans	Transactions with a promotion, where a promotion is across all items in a store e.g. 10% off if you spend more than $100	count	1,372	859	1,138
PromoPer	Transactions with a promotion	%	31.90	4.99	30.97
CpTrans	Transactions with a coupon, where a coupon is specific to an item, e.g. 10% off of a particular item	count	941	559	780
CpPer	Transactions with a coupon	%	22.23	4.57	22.00
PromoCpPer	Transactions with a promotion and a coupon	%	9.04	4.25	8.04
NonDiscTrans	Transactions with neither a promotion or a coupon	%	36.81	6.03	36.96
Email	Advertising Emails sent	count	15,441	26,678	90
Mail	Physical mailings sent	count	2,776	3,694	332
Inserts	Newspaper advertising inserts	count	16,231	26,360	0

Total number of observations = 78,231

Table 3 Summary Store Attributes

Name	**Description**	**Units**	**Mean**	**St. Dev.**	**Median**
LP Members	Number of loyalty program members	count	35,262	16,859	31,215
Sq.ft.	Store square footage	square feet	22,264	11,154	18,359
DateOpen	Date the store was opened	date	5/29/1940		3/9/1998
DateRemod	Date of latest remodel, if any	date	5/18/1996		0
YearOpen	Years since Opened	Years	16.72	12.38	14.91
YearRemod	Years since Remodeled	Years	10.52	11.16	6.23
TotHH*	Total households	count	102,450	46,895	99,960
HHGrowth*	Annual household growth	count	1,489	12,574	0.01
MedHHInc*	Median Household income	dollars	56,064	14,447	53,357
HHScore*	Index based on total households and annual household growth	index	439	2,643	100
HHInc.*	Households with income $60K - $120K	percent	34.70	10.80	34.00
ZIPDC*	ZIP Density Class	index	5.08	1.02	5.00
ZIPMC*	ZIP Market Class	index	2.85	1.73	2.00
$Comp_1$	Competitor 1 stores within 10 miles	count	1.78	1.68	1
$Comp_2$	Competitor 2 stores within 10 miles	count	0.68	1.04	0
$Comp_3$	Competitor 3 stores within 10 miles	count	0.23	0.61	0
$Comp_4$	Competitor 4 stores within 10 miles	count	0.35	0.65	0
Walmart	Walmart stores within 10 miles	count	3.32	2.48	3
TotComp	Total competitor stores within 10 miles	count	6.36	4.72	5
SMTenure	Store manager time in position	years	3.19	3.93	1.77
SMYears	Store manager total years of service	years	9.08	8.15	7.07

Total number of observations = 710.

*These variables correspond to the ZIP code where the store is located.

Table 4 Correlation for Store-Week Variables

	Units	GM	PlanSell	ActualSell	Transactions	Basket	PromoTrans
Units	1.000						
GM	0.699	1.000					
PlanSell	0.764	0.821	1.000				
ActualSell	0.769	0.825	0.973	1.000			
Transactions	0.818	0.854	0.934	0.939	1.000		
Basket	0.453	-0.003	-0.006	-0.004	-0.004	1.000	
PromoTrans	0.784	0.810	0.854	0.850	0.958	-0.003	1.000
PromoPer	0.168	0.152	0.074	0.057	0.206	0.000	0.438
CpTrans	0.764	0.798	0.882	0.891	0.934	-0.004	0.836
CpPer	0.018	0.015	0.032	0.038	0.022	0.000	-0.106
PromoCpPer	0.033	0.092	0.071	0.081	0.038	0.004	-0.037
NonDiscTrans	-0.175	-0.201	-0.135	-0.133	-0.213	-0.003	-0.255
Email	0.214	0.239	0.263	0.261	0.264	-0.002	0.249
Mail	0.269	0.299	0.328	0.340	0.327	0.002	0.289
Inserts	0.560	0.616	0.712	0.712	0.684	-0.003	0.607

	PromoPer	CpTrans	CpPer	PromoCpPer	NonDiscTrans	Email	Mail	Inserts
PromoPer	1.000							
CpTrans	0.014	1.000						
CpPer	-0.513	0.327	1.000					
PromoCpPer	-0.260	0.095	0.178	1.000				
NonDiscTrans	-0.253	-0.326	-0.461	-0.626	1.000			
Email	0.075	0.251	0.000	0.124	-0.149	1.000		
Mail	0.046	0.329	0.030	0.091	-0.125	0.759	1.000	
Inserts	0.044	0.700	0.130	0.103	-0.207	0.217	0.272	1.000

Table 5 Correlation for Store Attributes Variables

	LP Members	Sq.ft.	DateOpen	DateRemod	YearOpen	YearRemod	TotHH	HHGrowth	MedHHInc	HHScore
LP Members	1									
Sq.ft.	0.6941	1								
DateOpen	0.249	0.3894	1							
DateRemod	0.1892	0.0893	-0.2324	1						
YearOpen	-0.249	-0.3894	-1	0.2324	1					
YearRemod	-0.2663	-0.3144	-0.5471	-0.4797	0.5471	1				
TotHH	0.5783	0.4442	0.2329	0.2176	-0.2329	-0.3145	1			
HHGrowth	-0.0213	-0.0066	0.138	-0.0905	-0.138	-0.0871	0.0132	1		
MedHHInc	0.1658	0.2438	0.1166	0.1459	-0.1166	-0.181	0.3478	-0.0416	1	
HHScore	-0.0399	-0.0125	0.1511	-0.0981	-0.1511	-0.0959	-0.0182	0.94	-0.0483	1
HHInc	0.2037	0.2974	0.1457	0.1102	-0.1457	-0.1683	0.3265	-0.0399	0.8804	-0.0414
ZIPDC	-0.3341	-0.1274	-0.0791	-0.2439	0.0791	0.2115	-0.4699	0.0136	-0.2458	0.0419
ZIPMC	-0.395	-0.359	-0.2072	-0.188	0.2072	0.2609	-0.5655	0.0239	-0.6148	0.0529
$Comp_1$	0.3348	0.2995	0.1894	0.1647	-0.1894	-0.2297	0.5291	-0.0156	0.5307	-0.0368
$Comp_2$	0.1647	0.2367	0.1804	-0.0621	-0.1804	-0.1094	0.3177	-0.009	0.0983	-0.0162
$Comp_4$	0.0061	-0.0496	-0.0105	0.0508	0.0105	-0.0086	0.1737	0.023	0.2023	0.0155
$Comp_4$	0.1578	0.1225	0.1974	-0.0168	-0.1974	-0.1736	0.3119	0.043	0.1101	0.0251
Walmart	0.3832	0.335	0.2435	0.0765	-0.2435	-0.2149	0.5883	0.0228	0.2561	-0.0017
TotComp	0.3795	0.3454	0.261	0.0895	-0.261	-0.2439	0.6329	0.0133	0.3865	-0.0121
SMTenure	0.0375	-0.0048	-0.1489	0.0324	0.1489	0.193	-0.0833	-0.0689	-0.108	-0.0761
SMYears	0.1008	0.1039	-0.0849	0.0256	0.0849	0.1216	-0.0375	-0.0446	-0.0898	-0.0524

	HHInc	ZIPDC	ZIPMC	$Comp_1$	$Comp_2$	$Comp_3$	$Comp_4$	Walmart	TotComp	SMTenure	SMYears
HHInc	1										
ZIPDC	-0.1869	1									
ZIPMC	-0.6143	0.491	1								
$Comp_1$	0.4804	-0.5363	-0.6291	1							
$Comp_2$	0.1378	-0.1319	-0.1855	0.2421	1						
$Comp_3$	0.2163	-0.0788	-0.1862	0.1033	-0.186	1					
$Comp_4$	0.1231	-0.2566	-0.1879	0.2657	0.4055	-0.0711	1				
Walmart	0.316	-0.4054	-0.5169	0.6215	0.5494	0.0828	0.3937	1			
TotComp	0.4123	-0.4787	-0.5865	0.7862	0.6269	0.1579	0.5191	0.9327	1		
SMTenure	-0.0565	0.0973	0.1319	-0.0906	-0.1003	-0.0371	-0.0867	-0.1088	-0.1282	1	
SMYears	-0.0546	0.0554	0.0873	-0.0314	-0.0419	-0.0953	-0.0677	-0.0241	-0.0546	0.5299	1

Table 6 Payroll Deviations vs Sales

	(1)	(2)	(3)	(4)
PayrollDev	0.141***	0.138***	0.138***	0.153***
	(0.011)	(0.011)	(0.010)	(0.011)
$PayrollDev^2$				-0.065**
				(0.022)
Log(Mail)		0.016***	0.016***	0.017***
		(0.001)	(0.001)	(0.001)
Log(Insert)		0.004***	0.004***	0.004***
		(0.000)	(0.000)	(0.000)
PromoPer			1.106***	1.111***
			(0.056)	(0.056)
CpPer			-0.053	-0.048
			(0.057)	(0.058)
PromoCpPer			1.218***	1.217***
			(0.049)	(0.049)
Seasonal Controls	Yes	Yes	Yes	Yes
Store Controls	Yes	Yes	Yes	Yes
R^2	0.787	0.791	0.789	0.798
Observations	78,231	78,231	78,231	78,230
Groups	710	710	710	710

Robust Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7 Payroll Deviations vs Forecasted Sales - Model One

	(1)	(2)	(3)	(4)
PayrollDev	0.121***	0.121***	0.123***	0.135***
	(0.017)	(0.017)	(0.017)	(0.014)
$PayrollDev^2$				-0.053
				(0.030)
Log(Mail)		0.001	0.001	0.001
		(0.002)	(0.002)	(0.002)
Log(Insert)		-0.002***	-0.001***	-0.001***
		(0.000)	(0.000)	(0.000)
PromoPer			-0.491***	-0.490***
			(0.066)	(0.066)
CpPer			-0.523***	-0.522***
			(0.073)	(0.072)
PromoCpPer			0.028	0.028
			(0.054)	(0.054)
Seasonal Controls	Yes	Yes	Yes	Yes
Store Controls	Yes	Yes	Yes	Yes
R^2	0.423	0.424	0.429	0.430
Observations	41,884	41,884	41,884	41,884

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8 Payroll Deviations vs Forecasted Sales - Model Two

	(1)	(2)	(3)	(4)
$PayrollDev$	0.100***	0.100***	0.102***	0.111***
	(0.014)	(0.013)	(0.013)	(0.009)
$PayrollDev^2$				-0.042
				(0.032)
$Log(Mail)$		0.002	0.003*	0.003*
		(0.001)	(0.001)	(0.001)
$Log(Insert)$		-0.002***	-0.002***	-0.002***
		(0.000)	(0.000)	(0.000)
$PromoPer$			-0.784***	-0.783***
			(0.049)	(0.049)
$CpPer$			-0.522***	-0.521***
			(0.043)	(0.043)
$PromoCpPer$			-0.250***	-0.250***
			(0.042)	(0.042)
Seasonal Controls	Yes	Yes	Yes	Yes
Store Controls	Yes	Yes	Yes	Yes
R^2	0.494	0.497	0.506	0.506
Observations	41,883	41,883	41,883	41,883

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9 Drivers of Payroll Lift

	(1)	(2)	(3)	(4)	(5)	(6)
CpTrans	0.827*		12.099*		0.504	
	(0.318)		(4.954)			
PromoCp	-1.018		-17.518		-0.351	
	(0.600)		(9.724)			
PromoTrans	1.130		14.036		0.309	
	(0.628)		(9.868)			
Basket	0.044**	0.049***	0.654**	0.635***	0.551	-0.592
	(0.014)	(0.011)	(0.209)	(0.178)		
PlanSell	0.000		-0.000		-0.126	
	(0.000)		(0.000)			
#Loyalty Program Members	-0.000	0.000	-0.000	0.000	-0.368	-0.047
	(0.000)	(0.000)	(0.000)	(0.000)		
Years Since Remodel	-0.001*		-0.016		-0.220	
	(0.000)		(0.010)			
Mail	0.000		0.000		0.425	
	(0.000)		(0.000)			
ZIP Density Class	0.011		0.132		0.177	
	(0.009)		(0.126)			
ZIP Market Class	-0.003		-0.049		-0.123	
	(0.005)		(0.076)			
EffHH_100	0.001*	0.001*	0.007*	0.007*	0.371	-0.412
	(0.000)	(0.000)	(0.003)	(0.003)		
Store Manager Total Years of Service	0.002**	0.003***	0.035***	0.039***	0.434	-0.557
	(0.001)	(0.001)	(0.011)	(0.011)		
$Competitor_1$	0.017	0.019	0.308	0.224	-0.172	-0.155
	(0.020)	(0.018)	(0.306)	(0.276)		
$Competitor_2$	-0.028*	-0.023	-0.579**	-0.520**	-0.393	0.412
	(0.013)	(0.012)	(0.207)	(0.199)		
$Competitor_3$	0.052**	0.051**	0.382	0.310	0.216	-0.192
	(0.019)	(0.018)	(0.251)	(0.241)		
$Competitor_4$	-0.008	-0.010	-0.084	-0.179	-0.052	0.130
	(0.015)	(0.015)	(0.252)	(0.237)		
$Walmart$	0.068*	0.075*	0.965	1.025	0.180	-0.225
	(0.033)	(0.033)	(0.775)	(0.765)		
TotHH		-0.000		-0.000		
		(0.000)		(0.000)		
Cons.	-0.731**	-0.241	-12.505***	-6.163***		
	(0.268)	(0.071)	(4.332)	(1.268)		
R^2	0.092	0.078	0.074[a]	0.058[a]		
Observations	710	710	710	710	710	710

Robust Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
[a] Pseudo R^2

Table 10 Test Control Stores Validation

Variable	Unmatched Matched	Mean Treated	Mean Control	Bias %	Bias Reduction %	t-test t	t-test p
Revenue	U	84,318	53,753	104.6		5.03	0.00
	M		81,498	9.7	90.8	0.35	0.73
ActSell	U	8,249	5,543	105.4		4.93	0.00
	M		7,997	9.8	90.7	0.37	0.72
TotComp	U	6.44	6.35	2.1		0.09	0.93
	M		6.47	-0.7	68.9	-0.03	0.99
Sq.Ft.	U	29,123	21,556	70.0		3.45	0.00
	M		28,663	4.3	93.9	0.15	0.88

Total number of Stores = 61

Table 11 16 Store Validation Test

	(1)	(2)
TestMag	1.346***	1.059*
	(0.234)	(0.494)
Seasonal Controls	Yes	Yes
Store Controls	Yes	Yes
R^2	0.514	0.434
Observations	10,472	20,946
Groups	61	61

Robust Standard Errors in Parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12 Implementation Control Stores Validation

Variable	Unmatched	Mean		Bias	Bias	t-test	
	Matched	Treated	Control	%	Reduction %	t	p
Revenue	U	1,413,207	1,213,618	30.0		3.52	0.00
	M		1,357,378	8.4	72.0	0.77	0.44
ActSell	U	156,030	133,809	35.9		4.24	0.00
	M		146,327	12.7	64.6	1.16	0.25
TotComp	U	3.09	3.21	-4.4		-0.51	0.61
	M		2.99	3.6	18.0	0.35	0.73
Sq.Ft.	U	24,572	21,293	32.7		3.86	0.00
	M		23,708	8.6	73.6	0.78	0.44

Total number of Stores = 672

Table 13 168 Store Implementation

	(1)	(2)	(3)	(4)
$TestMag$	0.582***	0.532***	0.451***	0.690***
	(0.037)	(0.033)	(0.014)	(0.025)
$TestMag^2$				-1.121***
				(0.111)
Seasonal Controls	No	Yes	Yes	Yes
Store Controls	No	No	Yes	Yes
R^2	0.002	0.272	0.831	0.832
Observations	120,474	120,474	120,474	120,474
Groups	672	672	672	672

Robust Standard Errors in Parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$