# ECON 203 Empirical Modeling Research Paper:
# Employment Rate within Retail Trade Companies

Katherine Leyonmark
Allegheny College

**Introduction:**

I have a double major with Economics and Integrative Informatics. One day I hope to work as a Data Analyst for a digital marketing firm. This not only requires an immense amount of statistics, but the knowledge to interpret what that data means in the world of marketing. For my question of interest, I will be looking at how certain explanatory variables result in the number of employees within retail trade companies. I know a large background on retailing practices due to my marketing education, which involves knowing the difference between retailing and retail. Retail is when the company who makes the product sells it directly to the customer, whereas retailing is the last intermediary in the marketing channel that sells directly to the customer. It is not the company that makes the product. This question of interest will also tie into the current state of our economy today surrounding employees being laid off due to the pandemic. I will look at different factors that would dictate the amount of employees needed to maintain stability in different types of retailers. This fills a gap in our knowledge because it will show what companies would need to focus their attention on during the pandemic to keep their employment rate up, while still making money. Many retailers today have most likely not prepared for a worldwide crisis like this, so knowing where to focus their attention will be helpful in addition to the knowledge that they already have.

**Background/Literature Review:**

I looked at three different, relevant literature studies to justify my research question of interest. Because there is so much that goes into employment, I will look at it from some different angles that still correlate with the context of my regression model. One of my explanatory variables is “Number of Establishments”. Because of this, I looked at an article in which the question of interest was “Whether regions in the United States in which there is high unemployment show increases in the number of establishments in retail and consumer service industries in a subsequent four-year period”. The thought process behind this question was that unemployed workers seek self-employment as an escape out of unemployment (Carree, 2002). They had two parts to their model. In the first part, the number of establishments in a state is predicted using data on population, income, urbanization and age distribution (Carree, 2002). The residual of this regression was considered as an indicator of the opportunities present for new enterprises to enter (Carree, 2002). The second part of their model is the change in the number of establishments explained from changes in population and income (Carree, 2002). Their data consisted of sixteen industries in retailing and consumer services with the number of

establishments, average number of employees in that same year, the percentage increase in the number of establishments and the ratio of the number of establishments over the number of (employer) firms (Carree, 2002). Some limitations were that they did not include explanatory variables like population growth, change of disposable incomes and market room, which have to be taken into account as well to arrive at an adequate test (Carree, 2002). They concluded that even when unemployed workers would be more inclined to become self-employed than employed workers, they are less successful in terms of growth and survival (Carree, 2002). This is helpful to my model because it shows that there is a correlation to some degree between employment and number of establishments. Unemployed people are wanting to start establishments, but they more often than not fail to survive, meaning they probably will go back to the larger establishments when they can. The next article I looked at dives deeper into issues surrounding sales and employment. I wanted to look at this because of my variable “Sales, Value of Shipments, or Revenue”. Some parts of retail have seen large declines in sales and employment due to two different factors: a prediction that retail sales will migrate online and physical retail will be virtually extinguished, and a prediction that future shoppers will almost all be heading to giant physical stores like warehouse clubs and supercenters (Hortaçsu & Syverson, 2015). These are important factors that could possibly be error term residuals in my data. The data that this literature looks into uses the NAICS codes that I also use in my data. The article gives statistics such as the retail sector's share of total employment is near 11 percent currently and the retail sector has seen little change in its share of economic activity since the onset of the Great Recession (Hortaçsu et al., 2015). Something that I found important in this study was that the productivity levels and growth rates of retail establishments were correlated with their rates of investment in information technology (Hortaçsu et al., 2015). In conclusion, within the NAICS codes they studied, total retail employment has grown 17 percent since 1990 in every NAICS code industry they studied except one (Hortaçsu et al., 2015). Also, nonstore retailers, the industry in which the vast majority of online retail occurs, saw 27 percent employment growth over the period (Hortaçsu et al., 2015). A substantial statistic for my model that they found was that in 2014, the value added per employee in the nonfarm economy was $124,000, while in retail it was roughly half this level at $66,000 (Hortaçsu et al., 2015). This could affect how many employees are hired in retail due to how it would affect revenue of the industry. The third article I looked at examines the relationship between the size distribution of establishments and the level of regulation on the industry from 1992 to 2004 (Calcagno & Sobel, 2014). They concluded that higher levels of spending on regulation at the state level result in a

reduction in the proportion of zero and 1–4 employee establishments and an increase in the proportion of relatively larger establishments (Calcagno et al., 2014) This makes it harder for the smallest establishments to compete in the marketplace, thus lowering the number of establishments in the market and giving a relative cost advantage to larger establishments (Calcagno et al., 2014). This relates to my model because costly regulations could affect the number of employees hired or establishments in business, creating another variable that could be included in my error term.

**Data:**

For my data, I am using data from the United State Census Bureau's public database. This database is free and open for anyone to use and study. The table is titled "Retail Trade: Summary Statistics for the U.S., States, and Selected Geographies" and is a panel data table. Due to this fact, I chose to look at the most recent year's data, which happened to be 2017, because I only wanted to work with a cross sectional data set. This data was collected from the Economic Census survey. There are 163 observations within the table all from the United States. The table looks at different types of retailers such as Boat Dealers, Household Appliance Stores, Meat Markets and Shoe Stores. These are types of retailers that are the last intermediary selling to the customer in the channel, not necessarily the companies that make the product. I will be using the "Number of Employees" column as my dependent variable or my "y" value. Other explanatory variables that I will be using are "Number of Establishments", "Sales, Value of Shipments, or Revenue", "Annual Payroll" and "Meaning of NAICS Code". This last category will be in the form of a binary variable. The NAICS code shows the type of retailer that it is, which will most likely change the number of employees that a retailer will need. Therefore, I chose to look at "Food Retailers", "Clothing/Shoe Retailers", " Home/Garden Retailers", "Transportation Retailers", "Gasoline/Fuel Retailers" and "Other Retailers", which I believe will be some of the retailers most affected after the pandemic.

**Methodology:**

My regression model to start with has the variable "nmbremploy" which is the number of employees as the "y" or the dependent variable. The explanatory variables are "nmbrest" or number of establishments, "annpay" or annual payroll in thousands, "saleshiprev" or sales, value of shipments and revenue in thousands, "food" if the company is identified as a food or drink retailer by NAICS code, "clothshoe" if it is identified as a clothing or shoe retailer, "gasfuel",

if it is identified as a gasoline or fuel retailer, “home” if it is identified as a home or garden retailer, “trans” if it is identified as a transportation retailer, or “other” if it is identified as some other type of retailer. In order to avoid perfect collinearity, I am using dummy variables for six of my variables. These are the six categories of retailers that I chose from the NAICS codes in the data. The variables “food”, “clothshoe”, “home”, “trans”, “gasfuel” and “other” are binary and if they equal one then the data is from a NAICS retailer type within that category, and a zero if not. Because each of these variables can either be a value of one or zero, they do not need to have a log applied to them. These dummy variables will give a small glimpse into the employment within different categories of retail trade companies. I applied a log to the variables “saleshiprev” and “annpay” because they are continuous variables and for my dataset, they are never negative or zero. These variables will show me how a one percentage change in the monetary values will increase employment. I also applied a log transformation to “nmbrest” and “nmremploy” because even though they are discrete variables, they are significantly large enough values that a log transformation is needed. I will be using the adjusted R square value because I have more than one explanatory variable and I want to avoid overfitting, which can decrease the residual sum of squares, even if the explanatory variable does not have a great amount of influence. However, they could be highly correlated by chance and there is not an explanation for the relationship, so I am always skeptical. In order to answer my question of interest, I needed to find the statistical significance of my explanatory variables to see just how much they each affected the number of employees within retail trade companies. I set up a regression hypothesis test to check the significance of my variables with each of the three tests, the T statistic, the P value and the Confidence Interval test. For all of my hypothesis testing, I will be using a significance level of 5% and a critical value of 1.96. I do not need to run a F statistic joint hypothesis test because none of my explanatory variables are in similar categories. In order to fully make sure my regression was not biased, I ran two homoskedasticity tests: the White Test and the Breusch-Pagan Test. These will allow me to test for heteroskedasticity, and if my standard errors are biased. If at least one of these tests provides evidence against the null hypothesis, there is evidence in favor of heteroskedasticity and I will compute the robust standard errors using Stata.

**Empirical Results:**

My results of these configurations are as follows and are rounded to two decimal places. When my regression model is run, the intercept parameter is -2.73. This means that when all my

explanatory variables are equal to zero, the predicted number of employees is -2.73%. This is not very helpful, because there can not be a negative percentage of employees hired. With each additional increase of 1% in the number of establishments, the number of employees increased by 0.13%, with all other explanatory variables held fixed. With each additional increase of 1% in sales, value of shipments or revenue, the number of employees decreased by 0.16%, with all other explanatory variables held fixed. With each additional increase of 1% in annual payroll, the number of employees increased by 1.06%, with all other explanatory variables held fixed. For the dummy variables, when the retail company is in the food and drink category, the number of employees increased by 0.32%, with all other explanatory variables held fixed. When the retail company is in the clothing and shoe category, the number of employees increased by 0.43%, with all other explanatory variables held fixed. When the retail company is in the gasoline and fuel category, the number of employees increased by 0.19%, with all other explanatory variables held fixed. When the retail company is in the home and garden category, the number of employees decreased by 0.08%, with all other explanatory variables held fixed. When the retail company is in the transportation category, the number of employees decreased by 0.26%, with all other explanatory variables held fixed. Lastly, when the retail company is in the other category, the number of employees increased by 0.09%, with all other explanatory variables held fixed. The adjusted R squared value for this model is 0.9595, which means that all of these explanatory variables combined explain about 96% of the total variation in the number of employees. The results of my hypothesis tests are as follows for both the T-statistic test and the P value test. At the 5% significance level, we reject the null hypothesis in favor of the alternative hypothesis for the intercept, lnmbrest, lsaleshiprev, and lannpay. Thus, the effects of the intercept, Number of Establishments, Sales, Value of Shipment, or Revenue, and Annual Payroll are statistically different from zero at the 5% significance level. Yet, at the 5% significance level, we fail to reject the null hypothesis in favor of the alternative hypothesis for food, clothshoe, gasfuel, home, trans and other. Therefore, the effect of Food Retailers, Clothing/Shoe Retailers, Gasoline/Fuel Retailers, Home/Garden Retailers, Transportation Retailers and Other Retailers are not statistically different from zero at the 5% significance level. I also ran a confidence interval test for all the explanatory variables. Like it should, the results were all the same as the t-statistic and p value test. The results of the White homoskedasticity test are as follows. The Lagrange Multiplier Test Statistic is greater than the Chi-square distribution. The calculated F-statistic is greater than the F-distribution from the table. The p-value is also less than the significance level. Therefore, at the 5% significance level, we reject

the null hypothesis in favor of the alternative. The LM, F-statistic and p-value provide evidence against the null hypothesis of homoskedasticity in favor of heteroskedasticity. The results of the Breusch-Pagan test are as follows. The Lagrange Multiplier Test Statistic is less than the Chi-square distribution. The calculated F-statistic is less than the F-distribution from the table. The p-value is also greater than the significance level. Therefore, at the 5% significance level, we fail to reject the null hypothesis in favor of the alternative hypothesis. The p-value, LM and F-statisitc provide little evidence against the null hypothesis, or homoskedasticity. Because at least one of these tests provides evidence against the null, or heteroskedasticity, the robust standard errors must be computed. These standard errors are very different from the original standard errors. This makes sense because heteroskedasticity is violated in the White test. These results relate to my literature findings. In the first article, it was concluded that there was some correlation between number of establishments and employment rate. My results show that the number of establishments is a statistically significant explanatory variable for employment. The second article I looked at showed that there was an increase in retail employment growth since 1990 in all NAICS categories, except for gasoline stations. This would be interesting to look at against my 2017 results that state the number of employees increases by 0.19% when the company falls under the gasoline/fuel category. The third article talks about how more fixed costs affect the number of establishments being able to hire. My results showed that annual payroll is statistically significant, and that is related to cost of hiring employees like with the article.

**Conclusion:**

Based on the coefficients of my results, the annual payroll had the largest increase in percentage on the number of employees. This makes sense because annual payroll and number of employees are directly related, the more employees you have the higher your payroll will be. The NAICS categories of home/garden retailers and transportation retailers actually decreased the percentage of the number of employees as it increased. I am unsure as to why this would be and this would be a cause for further study. The variable sales, value of shipments or revenue also decreased the percentage of number of employees as it increased. This is also interesting to me because I would have assumed that more employees increases production and therefore sales. This could also be a situation of diminishing marginal returns or due to technological advances. The adjusted R squared value gives insight that my chosen variables explain a lot about the number of employees, which is promising. But again, with each of these

results, I am remaining skeptical. In regards to hypothesis testing, number of establishments, sales, value of shipment, or revenue, and annual payroll are all statistically significant, meaning that these variables most likely are what effect the number of employees the most. And therefore, in these trying times, should be focused on the most. The categories of NAICS codes were not found to be statistically significant and therefore, other factors besides for what type of retail trade company they were affects employment on a much greater scale.

**Further Issues:**

The main challenge I faced with my data availability was that in the table I found, I really only wanted to focus on four of the explanatory variables. Having a few more would have most likely strengthened my results and decreased the zero conditional mean assumption. Going through my literature review, I learned of a few explanatory variables that would have been beneficial. One of the issues talked about in the article by Hortaçsu et al. was that retail sales might migrate online and physical retail will be diminished. In our current worldly situation, this prediction is very true. During this pandemic, virtually everything is bought online. Having an explanatory variable that showed how much sales are online would most likely affect how many employees are hired by a company. In that same article, it talked about how the productivity of establishments also depended on the technology that the establishment invested in. This would be another interesting explanatory variable to include due to the fact that technology is being discussed as a threat to human laborers. In the article by Calcagno et al., it talked about how regulation affects the size of establishments. Having an explanatory variable for regulation costs would have been beneficial to see how many employees are hired based on how fixed costs affect an establishment. Lastly, possibly having some sort of explanatory variable on minimum wage in the area might have been helpful as well. This way I could see again how more costs of maintaining these employees affects the establishments in the different retail sectors.

Regression before applied logs:

```
. reg nmbremploy nmbrest saleshiprev annpay food clothshoe gasfuel home trans other
```

| Source | SS | df | MS | Number of obs | = | 163 |
|---|---|---|---|---|---|---|
| | | | | F(9, 153) | = | 1478.86 |
| Model | 2.9899e+14 | 9 | 3.3221e+13 | Prob > F | = | 0.0000 |
| Residual | 3.4370e+12 | 153 | 2.2464e+10 | R-squared | = | 0.9886 |
| | | | | Adj R-squared | = | 0.9880 |
| Total | 3.0242e+14 | 162 | 1.8668e+12 | Root MSE | = | 1.5e+05 |

| nmbremploy | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| nmbrest | 1.924134 | .4292191 | 4.48 | 0.000 | 1.076173 | 2.772095 |
| saleshiprev | -.002026 | .0002097 | -9.66 | 0.000 | -.0024404 | -.0016117 |
| annpay | .0543811 | .0027307 | 19.91 | 0.000 | .0489864 | .0597758 |
| food | 491473.5 | 155227.7 | 3.17 | 0.002 | 184807.1 | 798139.8 |
| clothshoe | 497200.7 | 163500.1 | 3.04 | 0.003 | 174191.5 | 820209.9 |
| gasfuel | 714798.9 | 168730.7 | 4.24 | 0.000 | 381456.1 | 1048142 |
| home | 347484.6 | 156519 | 2.22 | 0.028 | 38267.23 | 656702 |
| trans | 221548.3 | 163560.8 | 1.35 | 0.178 | -101580.9 | 544677.4 |
| other | 400772.8 | 160279 | 2.50 | 0.013 | 84127.29 | 717418.4 |
| _cons | -413320.4 | 158737.8 | -2.60 | 0.010 | -726921.3 | -99719.43 |

Regression after logs are applied:

```
. reg lnmbremploy lnmbrest lsaleshiprev lannpay food clothshoe gasfuel home trans other
```

| Source | SS | df | MS | Number of obs | = | 163 |
|---|---|---|---|---|---|---|
| | | | | F(9, 153) | = | 427.87 |
| Model | 319.638655 | 9 | 35.5154061 | Prob > F | = | 0.0000 |
| Residual | 12.6999276 | 153 | .083006063 | R-squared | = | 0.9618 |
| | | | | Adj R-squared | = | 0.9595 |
| Total | 332.338582 | 162 | 2.05147273 | Root MSE | = | .28811 |

| lnmbremploy | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnmbrest | .1290661 | .0391378 | 3.30 | 0.001 | .0517458 | .2063864 |
| lsaleshiprev | -.1608155 | .076742 | -2.10 | 0.038 | -.3124263 | -.0092047 |
| lannpay | 1.060831 | .0914225 | 11.60 | 0.000 | .8802177 | 1.241444 |
| food | .3234226 | .2981369 | 1.08 | 0.280 | -.2655737 | .9124189 |
| clothshoe | .4320449 | .3171929 | 1.36 | 0.175 | -.1945983 | 1.058688 |
| gasfuel | .1940649 | .3343688 | 0.58 | 0.563 | -.4665109 | .8546407 |
| home | -.0754085 | .3032782 | -0.25 | 0.804 | -.674562 | .5237449 |
| trans | -.2561405 | .3190256 | -0.80 | 0.423 | -.8864044 | .3741235 |
| other | .0948449 | .3105631 | 0.31 | 0.760 | -.5187006 | .7083904 |
| _cons | -2.730443 | .3649013 | -7.48 | 0.000 | -3.451338 | -2.009547 |

White homoskedasticity test results:

```
. imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(30)     =     69.89
         Prob > chi2  =    0.0001

Cameron & Trivedi's decomposition of IM-test
```

| Source | chi2 | df | p |
|---|---|---|---|
| Heteroskedasticity | 69.89 | 30 | 0.0001 |
| Skewness | 16.35 | 9 | 0.0599 |
| Kurtosis | 0.69 | 1 | 0.4060 |
| Total | 86.93 | 40 | 0.0000 |

Breusch-Pagan homoskedasticity test results:

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnmbremploy

         chi2(1)      =     3.12
         Prob > chi2  =   0.0773
```

# Works Cited

Calcagno, P. T., & Sobel, R. S. (2014). Regulatory costs on entrepreneurship and establishment employment size. Small Business Economics, 42(3), 541-559.

Carree, M. A. (2002). Does unemployment affect the number of establishments? A regional analysis for US states. Regional Studies, 36(4), 389-398.

Hortaçsu, A., & Syverson, C. (2015). The ongoing evolution of US retail: A format tug-of-war. Journal of Economic Perspectives, 29(4), 89-112.