Austin Dowd

ID: 001516270

MS Data Analytics

adowd12@wgu.edu

Advisor: Lea Yoakem

D214 Data Analytics Graduate Capstone | Task 3

Problem Statement

For most businesses being able to foresee when goods will be in demand and having the inventory on hand for consumers is challenging. When those goods are perishable, being able to forecast demand is a critical strategy in avoiding underestimating demand leading to stock-outs, or overestimating, which can lead to waste, both of which negatively impact revenue (Huber, Gossmann, & Stuckenschmidt, 2017)

Businesses that can forecast both the timing and amount of sales for goods can anticipate production, warehousing, and shipping needs and know what future constraints will be placed on labor and cash flow. (Milano, 2018)

Access to data beyond historical sales records has enabled companies to develop demand forecasting models based on various available data sources to predict demand (Brea, 2020).

The data set is available from Ke Peng and Nikhil Kaza on their study; Availability of neighbourhood supermarkets and convenience stores, broader built environment context, and the purchase of fruits and vegetables in US households (Peng & Kaza, 2019). It provides several variables beyond historical sales data that could provide value as predictors for the demand for perishable goods.

This study sought to discover: Can a regression model be built to predict a household's expenditures on perishable goods (fruits and vegetables) based on the available research data?

Hypothesis:

It is possible to build a regression model to statistically significantly predict household expenditures on perishable goods based on the available variables in the research data.

Data analysis process

To test this hypothesis, a multiple regression model was built using ordinary least squares, which included several predictor variables, both categorical and numerical data, to predict the household expenditure on fruits and vegetables.

The data set is available through public domain rights and can be downloaded directly from the University of North Carolina at Chapel Hill's

Dataverse site (Peng, 2022). The data set included 24 variables across 22,448 households, including fruit and vegetable expenditures, demographic data, and several variables related to the availability of stores, jobs, intersections, and other households in their area.

Before building the model, the data needed to be prepared with the following steps:

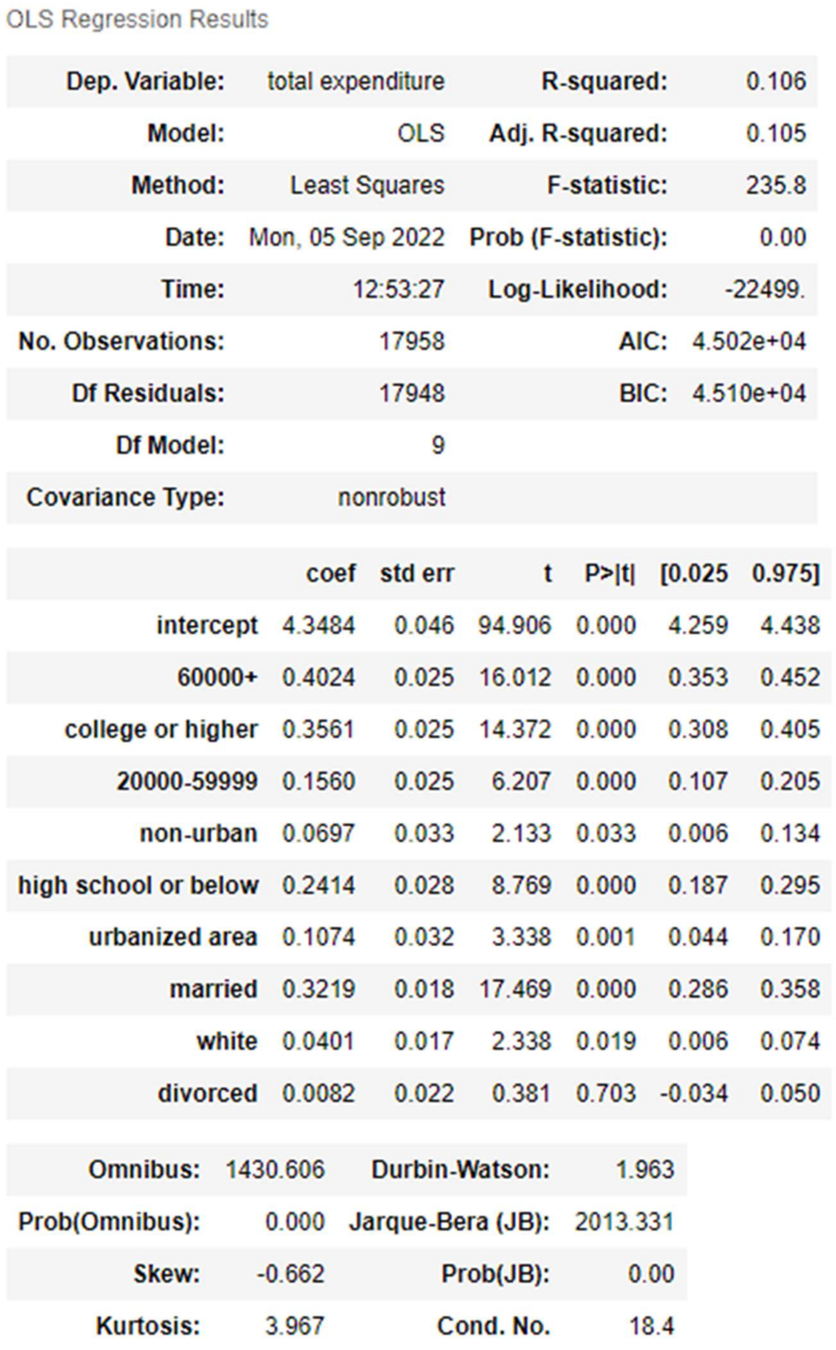
- Loading the data from the original .tab file into a pandas data frame.
- Examine the data for missing or poorly explained variables and remove them from the data frame.
- Creating dummy variables for categorical data in preparation for a regression model using one-hot encoding.
- Examine continuous variables' distribution and logarithmically transforming any skewed data for more accurate and reliable predictions.
- Develop an initial OLS model with all available variables to assess model performance and evaluate predictor variables for feature reduction.
- Reduce the number of predictor variables by assessing p-values and t-values to determine the strength of their relationship to the target variable. This was essential in reducing the likelihood of multicollinearity in the model and ensuring the selection of only values necessary to predict expenditures.

- Develop a second OLS model with the reduced variables and assess performance.

The final model was developed with nine variables. The data was split into training and test sets along a random 80/20 split.

Project findings

Despite the availability of so many variables across thousands of households, the final model did not perform well enough to reject the null hypothesis.



OLS Regression Results

Dep. Variable:	total expenditure	**R-squared:**	0.106
Model:	OLS	**Adj. R-squared:**	0.105
Method:	Least Squares	**F-statistic:**	235.8
Date:	Mon, 05 Sep 2022	**Prob (F-statistic):**	0.00
Time:	12:53:27	**Log-Likelihood:**	-22499.
No. Observations:	17958	**AIC:**	4.502e+04
Df Residuals:	17948	**BIC:**	4.510e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	4.3484	0.046	94.906	0.000	4.259	4.438
60000+	0.4024	0.025	16.012	0.000	0.353	0.452
college or higher	0.3561	0.025	14.372	0.000	0.308	0.405
20000-59999	0.1560	0.025	6.207	0.000	0.107	0.205
non-urban	0.0697	0.033	2.133	0.033	0.006	0.134
high school or below	0.2414	0.028	8.769	0.000	0.187	0.295
urbanized area	0.1074	0.032	3.338	0.001	0.044	0.170
married	0.3219	0.018	17.469	0.000	0.286	0.358
white	0.0401	0.017	2.338	0.019	0.006	0.074
divorced	0.0082	0.022	0.381	0.703	-0.034	0.050

Omnibus:	1430.606	**Durbin-Watson:**	1.963
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2013.331
Skew:	-0.662	**Prob(JB):**	0.00
Kurtosis:	3.967	**Cond. No.**	18.4

With an r-squared of .106 the model can only explain variance of the predictor variable by 10.6%, which is not statistically significant to predict household expenditures on perishable goods.

The final model had a condition number of 18.4, indicating that multicollinearity was not likely in this final model.

The resulting assessment of model performance would conclude that it is not possible to build a regression model to predict household expenditures on perishable goods with this data set.

Project Limitations

Having a data set available from a peer-reviewed research study has the advantage of being open and accessible and assurance that the data is sound and trustworthy as multiple academic professionals have reviewed it.

However, acquiring data from a published academic study limits the ability to gather further data on each household.

Additionally, the preparation of this data for the original research meant variables that might be more valuable to this study in their original format were changed to address the needs of the original researchers. In this data set, for example, income was reported as one of three categorical distinctions: "below 20000", "20000-59999" or "60000+".

A continuous value representing each household's actual annual income may have been more helpful in predicting expenditures.

Gathering additional variables that might impact consumer behavior and decisions on perishable goods purchases is unavailable through this data source and limits the ability of this study to explore other predictors to improve model performance.

Proposed actions

Despite the performance of the regression model with this data set, it is clear that the need for demand forecast of perishable goods is a highly valuable tool for many businesses.

With many companies already forecasting demand, the goal to further study consumer behavior to better predict expenditures on perishable goods is likely attainable.

The proposed course of action based on the results of this model are:

- Investigate the source of Peng & Kaza's original data from the Nielsen Homescan Consumer Panel Dataset for 2010(Peng & Kaza, 2019) to look for additional variables for modeling. Developing a richer data set with a larger number of possible predictor variables for each household should provide better insight into whether a regression model can be built to answer the proposed research question.
- Examine the data grouped by zip code and determine if a regression model might better fit demand forecasting by zip code rather than the individual household. Area-wide demand forecasting is still a valuable

tool for retailers whose markets might include multiple neighborhoods across their geographic area.

- Evaluate other predictive modeling options with the available data set. Despite the model's inability to predict a continuous variable like household expenditure, other options exist to predict consumer behavior and demand for perishable goods. Classification modeling might provide valuable insights into determining which households or zip codes will spend more on perishable goods.

Expected benefits of the study

The benefits of demand forecasting are spread across any business partner connected to the supply chain of perishable goods from the producer to the retailer.

For producers, knowing how demand for goods will rise and fall over time helps to plan the planting and harvesting of goods like fruit, vegetables, and flowers.

For suppliers and distributors, demand forecasting can aid in planning logistics for getting perishable goods from producers to sellers while minimizing the time needed to store goods along the supply chain. This can reduce waste from spoilage and reduce the need for large and expensive storage along the way, like industrial refrigerators.

At the end of the supply chain, retailers like grocers and food markets can benefit from demand forecasts by utilizing them as a strategy to anticipate the demands of consumers and planning orders and inventory management around those predicted demands. Correctly stocked perishable goods mean less waste to spoilage through overstocking and loss of revenue to stock-outs.

Additionally, every partner along the supply chain can benefit from demand forecasting when planning their business expenses. Knowing when demand will be at its highest ahead of time means companies can plan additional labor needs and ensure a well-trained and properly staffed team. Predicting when demand will ebb gives businesses the insights to better plan maintenance and equipment downtime. (Milano, 2018)

Sources

Bruce, P., Bruce, A. G., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using r and Python. O'Reilly.

Milano, S. (2018, December 12). *The advantages of demand forecasting*. Small Business - Chron.com. Retrieved September 2022, from https://smallbusiness.chron.com/advantages-demand-forecasting-60405.html

Chen, C., Wang, Y., Huang, G., & Xiong, H. (2019). Hierarchical demand forecasting for factory production of Perishable Goods. 2019 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata47090.2019.9006161

Huber, J., Gossmann, A., & Stuckenschmidt, H. (2017). Cluster-based hierarchical demand forecasting for perishable goods. Expert Systems with Applications, 76, 140–151. https://doi.org/10.1016/j.eswa.2017.01.022

Peng, Ke, 2022, "Availability of neighbourhood supermarkets and convenience stores, broader built environment context, and the purchase of fruits and vegetables in US households", https://doi.org/10.15139/S3/U9NMA9, UNC Dataverse, V1, UNF:6:21gunr4KSHmfLrBDyeu9ig== [fileUNF]

Peng, K., & Kaza, N. (2019). Availability of neighbourhood supermarkets and convenience stores, broader built environment context, and the purchase of fruits and vegetables in US households. Public Health Nutrition, 22(13), 2436–2447. https://doi.org/10.1017/s1368980019000910

Traasdahl , A. (2020, June 11). How AI is Changing Perishable Food Forecasting. SupplyChainBrain RSS. Retrieved August 25, 2022, from https://www.supplychainbrain.com/blogs/1-think-tank/post/31411-how-ai-and-analytics-are-changing-fresh-and-perishable-food-forecasting

Lu, C.-J., & Chang, C.-C. (2014). A hybrid sales forecasting scheme by combining independent component analysis with K-means clustering and support vector regression. *The Scientific World Journal*, *2014*, 1–8. https://doi.org/10.1155/2014/624017

Brea, C. (2020, November 26). *Predicting consumer demand in an unpredictable world*. Harvard Business Review. Retrieved September 2022, from https://hbr.org/2020/11/predicting-consumer-demand-in-an-unpredictable-world