**Probability Theory and Introductory Statistics**

**Final Project**

**Rambhupal Payyavula**

**NEU ID:  002108015**

**Date:12/12/2021**

**Dataset:** Walmart\_Store\_Sales.csv

**Source:** Kaggle

**About dataset:** Dataset represents the historical sales data of 45 Walmart retail stores for a time between 2010 and 2012. Dataset initially contains 8 variables and 6,435 observations.

**Description dataset attributes:**

**Store:** It is an integer value that lists the Walmart retail store numbers from 1 to 45

**Date:** it is a character variable having week dates of sales revenue

**Weekly sales:** it is a numerical value shows the Weekly sales revenue of Walmart retail stores

**Holiday Flag:** it is an integer value with 0s and 1s. 0 indicates that week does not have public holiday and 1 **indicates** the holiday week.

**Temperature:** It is a numerical value indicates the average temperature during that week

**Fuel Price:** It is a numerical value indicates the average Fuel price during that week

**CPI:** It is a numerical value indicates the consumer price index

**Unemployment:** It is a numerical value indicates the unemployment rate at time.

Below is the summary and structure of raw dataset.

A black screen with white text

Description automatically generated with low confidence

Text

Description automatically generated

Required cleaning has been performed, and below is the summary and structure of cleaned dataset.

Text

Description automatically generated

A black screen with white text

Description automatically generated with low confidence

**Milestone project 1:**

In milestone project 1, dataset representing the historical data of 45 Walmart retail stores for a time between 2010 and 2012 had been analyzed and had addressed the questions like weekly sales performance of store1 and store2 in three different years 2010, 2011, and 2012.’

**Summary of the milestone project 1:**

* Number of Walmart retail stores ranges from 1 to 45 and dataset covered the sales from 2010-02-05 to 2012-10-12.
* Maximum and minimum weekly sales reported over 45 stores is $2685352, $209989, and mean value is 1102767.
* Unemployment rate has negative impact on Walmart retail stores.
* For both store1 and store2, Sales had been increased from 2010 to 2011 and decreased from 2011 to 2012. We may expect the same pattern for other stores too.
* With increase in the store number and temperature, sales got decreased.
* Store 2 has recorded more sales in every financial year from 2010 to 2012 compared to store1

**Milestone project 2:**

In milestone project 2, The same dataset was analyzed, and inferential statistics and hypothesis testing has been employed to address the below questions.

1. Are average weekly sales of all stores in 2011 greater than average weekly sales of all stores in 2010?
2. Are average weekly sales of all stores in 2012 greater than average weekly sales of all stores in 2011?
3. Are average weekly sales of all stores in 2012 greater than average weekly sales of all stores in 2010?

With the help of inferential statistics and hypothesis testing, we have found that average Weekly sales of all stores are equal in2010, 2011, 2012.

**Summary of the milestone project 2:**

We have compared the three different samples representing the weekly sales in 2010, 2011, 2012 respectively. From the performed paired hypothesis testing, we can say that average sales of all 45 stores in there consecutive years,2010, 2011, 2012 were same and there was no increase in the sales average from 2010 to 2011 and 2011 to 2012.

We have analyzed the store1 and store2 performance and compared the sales of all stores among three years. its time know if the sales are influenced by the external factors fuel price, and unemployment rate. And comparing unemployment rate from 2010 to 2012.

In final project, we will analyze the dataset and address the below questions.

1. Did Fuel price show the effect on weekly sales of all stores in years 2010, 2011, 2012? If yes, how it affected?
2. Did unemployment rate show the effect on weekly sales of all stores in years 2010, 2011, 2012? If yes, how it affected?
3. Was unemployment rate increased from 2010 to 2012?

Above questions can be answered by employing inferential statistics and regression.

As dataset comprises of 6,435 observations, let take 28 random samples to perform the regression analysis.

Below is the summry, structure, and description of sample dataset.

Text

Description automatically generated

A black screen with white text

Description automatically generated with low confidence

A black screen with white text

Description automatically generated with low confidence

As we are concentrating on weekly sales, unemployment. Rate, Fuel. Price to address the above questions, box plots have been plotted for the respective data to know distribution of the sample data.

Chart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generated

Normal Q-Q plots have plotted to test normality of the sample.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

From the above normal Q-Q plots we can say that samples are normally distributed.

**Did Fuel price show the effect on weekly sales of all stores in years 2010, 2011, 2012? If yes, how it affected?**

**Independent Variable:** Fuel Price

**Dependent Variable:** Weekly Sales

Text

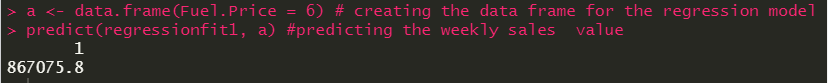
Description automatically generated

We can see from the summary that residuals range is very high, which is not good for the model. P-value is far greater than the significance level 0.1that shows that the covariance does not differ from zero and R-squared value is very near to zero, which shows the poor best line fit for the estimations. Overall p-value 0.9 which shows that it is not a good regression model.

Chart, scatter chart

Description automatically generated

From the above graph, we can see that the regression line is almost horizontal. However, we can see a slight negative slope which is negligible. We can say that fuel price has no effect on the weekly sales.



**Did unemployment rate show the effect on weekly sales of all stores in years 2010, 2011, 2012? If yes, how it affected?**

**Independent Variable:** Unemployment rate

**Dependent Variable:** Weekly Sales

We can see from the below summary that residuals range is very high, which is not good for the model. P-value 0.289 is greater than the significance level 0.1 that shows that the covariance does not differ from zero and R-squared value is very near to zero, which shows the poor best line fit for the estimations. Overall p-value 0.289 which shows that it is not a good regression model.

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

From the above graph, we can see that the regression line has a positive slope where can say that weekly sales were increased with the unemployment rate.

Predicting the weekly sales unemployment rate of 8.

Text

Description automatically generated

We have seen that the fuel price has no influence on the weekly sales, but the unemployment rate has influence on weekly sales. Let’s compare the employment rate from 2010 to 2012.

**Was unemployment rate increased from 2010 to 2012?**

This question can be answered with the help of paired t-test. Checking if the assumptions are met to perform the hypothesis test. From the below plots, we can say that the sample data is normally distributed.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**Validating the homogeneity of two samples:**

Homogeneity in the variance of two samples can be tested using the F-test where null hypothesis is the ratio of variance of two samples is equal to 1 and alternative hypothesis is the ratio of variance of two samples is not equal to 1. If the p- value is greater than the 0.05 (level of significance), we can say that both the sample variances are equal.

**Null Hypothesis:** True ratio of variances of 2010 and 2012 samples is equal to 1

**Alternative Hypothesis:** True ratio of variances of 2010 and 2012 samples is not equal to 1

**Significance level:** 0.05

As p-value, 0.2149 is greater than 0.05, we can conclude that variance of 2010 and 2012 samples are same

Text

Description automatically generated

Since the variance is same, we will be performing the paired one tailed t-test with equal variance.

**Paired t-test with equal variance:**

**Null Hypothesis:** True difference in the means of 2010 and 2012 samples is equal to zero

**Alternative Hypothesis:** True difference in means is less than zero.

**Significance level:** 0.05

Text

Description automatically generated

From the t-test, we have obtained the p-value of 0.8224, which is greater than significance level 0.05. We don’t enough statistical evidence to reject the null hypothesis, true difference in means is equal to zero. So, we can conclude means of 2010 and 2012 are equal. Average unemployment rate remained the same in 2010 and 2012.

As average unemployment rate remained same in 2010 and 2012 and unemployment rate has positive correlation with the weekly sales, we can say that average weekly sales remained same in 2010 and 2012, which has been proved in milestone project 2 with the help of hypothesis testing.

**Multi Variable Regression with Unemployment. Rate, Fuel. Price as independent variables**

We can see from the below summary that residuals range is very high, which is not good for the model. P-value for all the independent variables is greater than the significance level 0.1, which shows that the covariance does not differ from zero and adjusted R-squared value is very near to zero, which shows the poor best line fit for the estimations. Overall p-value 0.725, which shows that it is not a good regression model.

Text

Description automatically generated

Added variable plots to check the fit of independent variables in the regression model

Chart, scatter chart

Description automatically generated

Predicting the weekly sales for fuel price $6 per gallon and unemployment rate of 8.

Text

Description automatically generated

**Summary:**

In conclusion, Store 2 has recorded more sales in every financial year from 2010 to 2012 compared to store1. For both store1 and store2, Sales had been increased from 2010 to 2011 and decreased from 2011 to 2012. Average sales of all 45 stores in there consecutive years,2010, 2011, 2012 were remained same. Weekly sales were not influenced by the fuel price, but surprisingly unemployment rate has positive affect on the weekly sales as weekly sales increased with the unemployment rate. And we can also see from hypothesis testing that average unemployment rate remained same in 2010 and 2012.

***References:***

* *Hypothesis Test: Difference Between Means*. (n.d.). Retrieved from stattrek: https://stattrek.com/hypothesis-test/difference-in-means.aspx
* Kabacoff, R. I. (2020). *R in Action Third Edition.* Manning Publications.
* *linear regression*. (n.d.). Retrieved from tutorialspoint.com: https://www.tutorialspoint.com/r/r\_linear\_regression.html
* *r multiple regression*. (n.d.). Retrieved from tutorialspoint.com: https://www.tutorialspoint.com/r/r\_multiple\_regression.html
* *walmart-dataset-retail*. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/rutuspatel/walmart-dataset-retail