Practical Exercise 1

Student ID: 11484265

2024-05-26

`Briefly present the problem you are addressing (5%);`

To address Walmart's sales forecasting challenges, a statistical model is proposed. Leveraging historical sales data and key variables like store, holidays, temperature, fuel prices, and unemployment rates, and employing generalized linear regression techniques, the model aims to predict weekly sales more accurately to help avoid stock shortages and revenue losses.

`Explore the dataset, perform variable transformations/recoding where necessary, and choose one suitable`

The variable "Store" is converted into factor type, considering the numerical values as representations of distinct locations rather than quantitative measures. This adjustment ensures that the categorical nature of the data is appropriately accounted for in subsequent analyses.

Variable "Date" is transformed into "Week", the corresponding week of the year, since the variable is the Friday of each week rather than representing the specific day sales happened.

Additionally, an extra factor variable "near_Hol" is introduced to determine if the next nearest holiday falls within a 2-week period. This variable takes on values of 0 if the week is not near a holiday, 1 if it is, and 2 if the week contains a holiday. This allows the assessment of the impact of holiday.

Finally, "SalesStM" represents the standardized value derived from "Weekly_Sales." Recognizing that each store's expected sales differ due to their geographical locations, this variable standardizes weekly sales to a fixed range between 0 and 1. This approach is chosen over normalization to maintain the original distribution of the data for comparison between holiday and non-holiday periods.

```
## $corr
##              Weekly_Sales      SalesStM   Temperature   Fuel_Price          CPI
## Weekly_Sales  1.000000000 -0.1025046707 -0.0638100132  0.009463786 -0.072634162
## SalesStM     -0.102504671  1.0000000000 -0.0001998065 -0.009367500 -0.030126121
## Temperature  -0.063810013 -0.0001998065  1.0000000000  0.144981806  0.176887676
## Fuel_Price    0.009463786 -0.0093675004  0.1449818060  1.000000000 -0.170641795
## CPI          -0.072634162 -0.0301261205  0.1768876763 -0.170641795  1.000000000
## Unemployment -0.106176090  0.0799092088  0.1011578573 -0.034683745 -0.302020064
## Year         -0.018377543 -0.0126154949  0.0642692289  0.779470302  0.074795731
## Week          0.076139018  0.1756230287  0.2339406339 -0.042829710  0.005032562
##              Unemployment        Year         Week
## Weekly_Sales  -0.10617609 -0.01837754  0.076139018
## SalesStM       0.07990921 -0.01261549  0.175623029
## Temperature    0.10115786  0.06426923  0.233940634
## Fuel_Price    -0.03468374  0.77947030 -0.042829710
## CPI           -0.30202006  0.07479573  0.005032562
## Unemployment   1.00000000 -0.24181349 -0.012858814
## Year          -0.24181349  1.00000000 -0.193649046
```

```
## Week          -0.01285881 -0.19364905  1.000000000
##
## $corrPos
##           xName        yName x y          corr
## 1  Weekly_Sales Weekly_Sales 1 8  1.0000000000
## 2      SalesStM Weekly_Sales 2 8 -0.1025046707
## 3      SalesStM     SalesStM 2 7  1.0000000000
## 4   Temperature Weekly_Sales 3 8 -0.0638100132
## 5   Temperature     SalesStM 3 7 -0.0001998065
## 6   Temperature  Temperature 3 6  1.0000000000
## 7    Fuel_Price Weekly_Sales 4 8  0.0094637863
## 8    Fuel_Price     SalesStM 4 7 -0.0093675004
## 9    Fuel_Price  Temperature 4 6  0.1449818060
## 10   Fuel_Price   Fuel_Price 4 5  1.0000000000
## 11          CPI Weekly_Sales 5 8 -0.0726341620
## 12          CPI     SalesStM 5 7 -0.0301261205
## 13          CPI  Temperature 5 6  0.1768876763
## 14          CPI   Fuel_Price 5 5 -0.1706417952
## 15          CPI          CPI 5 4  1.0000000000
## 16 Unemployment Weekly_Sales 6 8 -0.1061760897
## 17 Unemployment     SalesStM 6 7  0.0799092088
## 18 Unemployment  Temperature 6 6  0.1011578573
## 19 Unemployment   Fuel_Price 6 5 -0.0346837449
## 20 Unemployment          CPI 6 4 -0.3020200637
## 21 Unemployment Unemployment 6 3  1.0000000000
## 22         Year Weekly_Sales 7 8 -0.0183775426
## 23         Year     SalesStM 7 7 -0.0126154949
## 24         Year  Temperature 7 6  0.0642692289
## 25         Year   Fuel_Price 7 5  0.7794703019
## 26         Year          CPI 7 4  0.0747957315
## 27         Year Unemployment 7 3 -0.2418134940
## 28         Year         Year 7 2  1.0000000000
## 29         Week Weekly_Sales 8 8  0.0761390178
## 30         Week     SalesStM 8 7  0.1756230287
## 31         Week  Temperature 8 6  0.2339406339
## 32         Week   Fuel_Price 8 5 -0.0428297095
## 33         Week          CPI 8 4  0.0050325625
## 34         Week Unemployment 8 3 -0.0128588137
## 35         Week         Year 8 2 -0.1936490460
## 36         Week         Week 8 1  1.0000000000
##
## $arg
## $arg$type
## [1] "upper"
```

The correlation matrix illustrates a weak correlation across various variables with sales. Nevertheless, a weak correlation coefficient does not necessarily imply the absence of a meaningful relationship. Examination of statistical significance is still necessary to uncover potential relationships.

```
Develop a suitable model that predicts sales at Walmart stores, incorporating significant influencing fa
```

The data is first split into training (70%) and test data (30%).

A Gaussian generalized linear model was selected due to the continuous nature of the response variable, expected to follow a normal distribution.

Initially, all predictor variables, along with interaction terms, were included in the model on training data. Subsequently, non-significant variables ($p > 0.05$) were removed through a backward, iterative process. Remaining significant predictors include fuel price, temperature and unemployment interaction, fuel price and CPI interaction, and CPI and Holiday proximity interaction.

Overall model significance was confirmed by a significant F-value, indicating a relationship between predictors and the response variable, thereby rejecting the null hypothesis.

Fuel_Price -6.995e+04 2.336e+04 -2.995 0.002762 ** Temperature:Unemployment -1.407e+02 1.867e+01 -7.536 5.85e-14 ***Fuel_Price:CPI 3.019e+02 1.202e+02 2.512 0.012037***
CPI:near_Hol1 -1.009e+02 8.419e+01 -1.198 0.230903
CPI:near_Hol2 3.638e+02 5.427e+01 6.703 2.30e-11 *

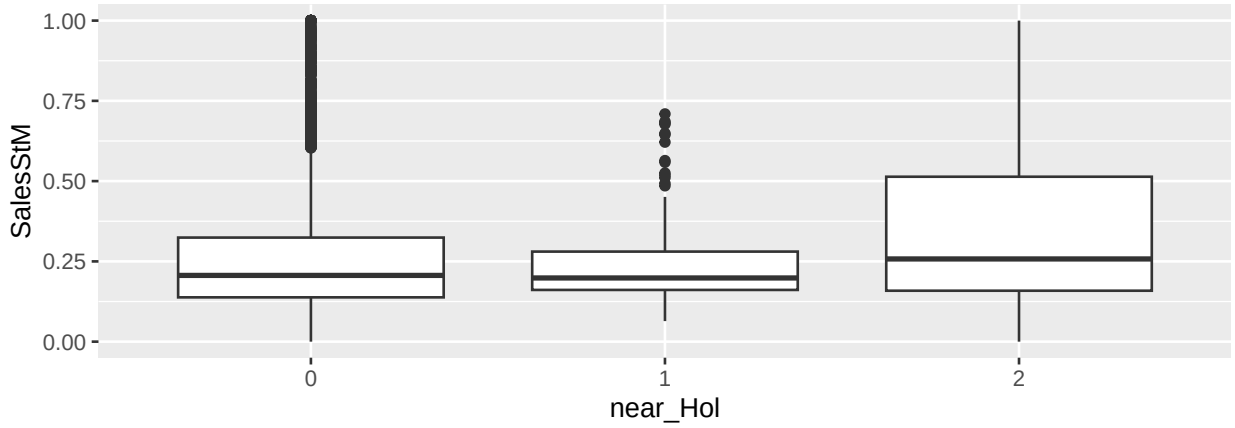
```
Perform suitable diagnostics, assess fit,  and justify your final choice of model (30%);
```

The model's R-squared values for both training and test datasets are notably high, at 91.8% and 92.1% respectively. This indicates that a substantial portion of the variation in the data is explained by the model.

```
## [1] 0.91895
```

```
## [1] 0.9190735
```

Although the holiday variable alone does not exhibit significance in this model, its interaction with the Consumer Price Index (CPI) proves to be statistically significant. The boxplot illustrates a noticeable increase in sales during holiday periods, when sales are disaggregated by store location.



Although the residual plot shows a slightly higher frequency of positive errors, implying a systematic underestimation of the response variable by the model, there is no discernible pattern. This suggests that the errors are randomly distributed around zero, indicating adherence to the assumptions of the regression model.

```
## NULL
```

```
Present your main findings and briefly comment on the limitations of your approach in the context of sta
```

Store explains the majority of variation in sales, other variables maintain significance, albeit to a lesser extent.

Fuel_Price -6.995e+04 2.336e+04 -2.995 0.002762 ** Temperature:Unemployment -1.407e+02 1.867e+01 -7.536 5.85e-14 ***Fuel_Price:CPI 3.019e+02 1.202e+02 2.512 0.012037***
CPI:near_Hol1 -1.009e+02 8.419e+01 -1.198 0.230903
CPI:near_Hol2 3.638e+02 5.427e+01 6.703 2.30e-11 *

This analysis identified a significant interaction effect between holiday and the Consumer Price Index (CPI) on sales. Although the holiday variable alone lacked significance, its interaction with CPI proved influential.

The model's generalizability may be limited to the specific dataset used. Omitted variable bias and the assumption of linear relationships between predictors and the response variable pose potential limitations. Additionally, outliers and influential observations may impact the robustness of our findings.