**ISM 6137 - Statistical Data Mining**

**Assignment 5**

**Sahil Shah – 19895141**

1. Predictor Table

|  |  |  |
| --- | --- | --- |
| **Variables:** | **Sign** | **Rationale** |
| Item\_ID | NONE | ID itself would not affect sales |
| Item\_Weight | NONE | Item weight would not affect sales |
| Item\_Fat\_Content | +/- | Items with high fat content would sell less if consumers are health conscious |
| Item\_Visibility | + | Higher visbility should create more sales |
| Item\_Type | +/- | Item type will affect sales |
| Item\_MRP | +/- | higher mrp likely lower sales volume |
| Outlet\_ID | NONE | The ID itself would not affect sales |
| Outlet\_Year | + | Newer the outlet, the higher the sales would probably be |
| Outlet\_Size | + | As outlet size increases, sales should also increase |
| City\_Type | + | Bigger cities should have more traffic |
| Outlet\_Type | +/- | Type of outlet, supermarket vs grocery etc. would affect inventory and thus sales |
| Store\_Age | - | younger the store likely higher the sales |

**Data Exploration:**

1. **Target Variable: Item\_Sales**

**Chart, histogram

Description automatically generated** **Chart, histogram

Description automatically generated**

We see that the target variable seems to follow a poisson type of curve, a histogram of log(Item\_Sales) is closer to normal

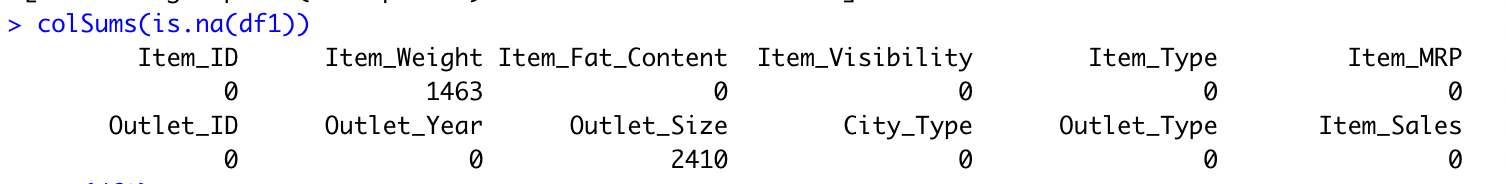
Chart

Description automatically generated Chart, histogram

Description automatically generated Timeline

Description automatically generated with medium confidence

Many missing values in Item\_Weight and Outlet Size:

****

**However, we are not going to consider weight, and outlet size is somewhat captured by outlet type (supermarkets being bigger than grocery stores)**

**MODELS:**

**Because target variable is non-linear, makes sense to use either log transform or poisson**

fe2 <- glm(Item\_Sales ~ Item\_Type+ City\_Type + Item\_MRP + Item\_Fat\_Content +

Item\_Visibility + Store\_Age + Outlet\_Type +Outlet\_ID, data=df1, family=poisson(link=log))

re2 <- lmer(log(Item\_Sales) ~ Item\_Type+ City\_Type + log(Item\_MRP) + Item\_Fat\_Content

Item\_Visibility + Outlet\_Type +(1 | Outlet\_ID), data=df1)

re3 <- lmer(log(Item\_Sales) ~ Item\_Type+ log(Item\_MRP) + Item\_Fat\_Content Item\_Visibility + Store\_Age + Outlet\_ID + Outlet\_Type +(1 | City\_Type ), data=df1)

stargazer(fe2, re2, re3, out = "star\_out\_mlm.txt",align = TRUE, title = "MLM Regression Results", single.row = T, type = "text")

(OUTPUT IS ON NEXT PAGE)

Based on the output, model 2 (re2) seemed best to answer the given questions. Assumption testing results for

model 2 are below:

Text, letter

Description automatically generated**Table

Description automatically generated with low confidence**

**Model does not seem to have major multicollinearity and DW test shows no issue with independence**

**QUESTIONS:**

Your client is a business entrepreneur considering franchising one or more stores of this retail chain and is looking for the following answers, with adequate justification:

* What type of outlet will return him the best sales: Grocery store or Supermarket Type 1, 2, or 3.

Based on the output, all the supermarket types had a positive effect on overall sales. Supermarket type 3 had the greatest effect, having a β-Coefficient value of 2.50 indicating a 250% increase in sales with that outlet type compared to the base case of grocery.

* What type of city will return him the best sales: Tier 1, 2 or 3.

Model two suggests that Tier 2 Cities have the greatest positive effect on sales, but the effect is very small at 0.2% (β = 0.002).

* What are the top 3 highest performing and lowest performing stores in the sample.

The top 3 performing outlets can be ascertained from the ranef output of outletID shown below. According to this output, the top performing outlets were 35, 49 and 10 with ranef output values of 0.0307, 0.0178 and 0.0035 respectively indicating 3%, 1.7% and 0.3% sales above the mean

The lowest 3 performing stores were 45, 46 and 13 with output values of -0.045, -0.018, & -0.0035 respectively, indicating -4.5%, -1.8% and -0.3% sales below the mean

**A picture containing text

Description automatically generatedText

Description automatically generated**

Figure Ranef(re2) output

**Recommendation:** The owner should open supermarket type stores, primarily in tier 2 cities. However, since tier 2 city benefit is quite small, should also consider cost differences of opening outlets in tier 2 cities vs other options.