# Stormy Weather Stormy Sales

*Jonathan Lu*
*Kara Wong*

*15 December, 2019*

## Linear Regression

Before we could build our first linear model, we had to stitch the weather data to the original dataset provided by Walmart. The main bulk of the work when it came to this process was cleaning the weather data to a point where it could be added together.

At an initial glance, we could see that the weather data was extremely incomplete (see Figure 8.) When aggregated to a station level, every single station had some sort of missing data. The variables that were missing the most data were depart, sunrise, sunset, and snowfall. Each of these variables had over 50% of their total data across all stations missing. With this many missing observations, we decided that it would be extremely unreasonable to try and impute data for that many missing observations as we were neither knowledgeable enough about weather, nor knowledgeable enough about the stations themselves to be able to classify these missing observations as anything other than "structurally missing". Therefore, since there was nothing we could do on our own to make these variables complete, we chose to remove them from the data. A variable which was close to the cusp in terms of being "un-imputable" was sea level. However, we quickly noted that sea level was pretty much identical to stnpressure (station pressure) as they both measured approximately the same thing. Since these two variables are pretty much perfectly collinear, we chose to remove sea level as there were far more observations for station pressure. Another variable we removed due to multicollinearity is tavg, or the average temperature of the day. When we looked into it, the value is calculated as the average of the minimum temperature and the maximum temperature. We chose to remove the average because we believed that the more extreme weather experienced, which would be captured better by min and max temperature, the more likely people are to buy weather related gear.

While removing a couple of variables definitely improved on the amount of missing data remaining, almost every single variable and station combination had some number of missing data. However, the amount of data missing at this point was extremely trivial ranging from $\sim 0.5\%$ of each station's data. Given that very few observations were missing, we felt relatively comfortable imputing a four day moving average for the missing observations. While this method cleaned up most of the variables for all of the stations except one. Once we aggregated the data to a station level, it became clear that station five had an un-imputable amount of data. Usually in these circumstances, we would remove the station completely, however, since we needed to have matching data for every store in Walmart's initial dataset we had to come up with a way to impute multiple values. To do this, we aggregated the data to a day level and imputed the mean of all the other station's observations at that day.

Before we finish with our weather data, we decided to create additional variables off of the weather data. When looking over the variables provided, one really stood out and that was date. We thought that it would be incredibly naive to think that the actual value of the date would have some predictive power over the number of units sold so instead; we constructed two categorical variables to suplant the use of the date. We decided that whether or not the day was a weekend and what season the day would fall under would be far more predictive as it narrows down the date to certain features related to it. While we thought that seasons might be really interesting as it sort of has to do with the weather, we were a bit worried about whether or not the change in seasons would already be reflected in the temperature data. We ultimately decided to leave it in as some locations such as California, barely have temperature change. We also changed codeSum from a set of strings representing weather events to a categorical variable representing whether or not some weather event happened on the day.

Joining the weather data to the Walmart data was extremely straightforward as all it required was two left

joins. The first was to match a station to each store using the provided key. The last join was on the station number and the date, thus giving us a fully merged data set.

When it comes to building the first model, we wanted to keep in as many variables as we could, however, we decided to remove date and station number from model. We chose to remove date because we believed that the actual date would have little to no impact on whether or not someone buys something and instead chose to use our created variables. We also chose to remove station number because the dummy variable that are created to represent each station is perfectly collinear with a certain set of store numbers that the station maps to. The code for the model can be found below.

```
md1 = lm(units ~ . - date - station_nbr , data = trn_data)
```

Since in this model, we treated store number and item number as a categorical variable, there are 170 total variables not including the intercept in this model. When looking at the summary output for this model, it becomes clear that matching store numbers is one of the best ways to predict the amount of unit sales. However, like all categorical variables with an extreme number of factors, not all of the levels are significant. Item number also follows a similar trend in that there are some levels which have coefficients indistinguishable from zero, but the m