Predicting Sales of Cold Drinks

Capstone Statistical Analysis

February 21, 2018

## A note on modeling

Although my dataset features *time-series count data*, I will be using a linear model to assess the effects of weather variables.

There are two reasons for this:

* **The appropriate statistical analysis tools—time series analysis, as well as Poisson and Negative Binomial Regression—are outside the scope of this workshop.** Given that I lack prior academic training necessary to apply the more complex models, I will have to ignore the time-series and count nature of the data and use linear regression.
* More importantly, **I am interested in the *directionality* of the relationships between the variables, rather than exact predictions.** Further, I want to know which *among* the weather variables have stronger effects.

Given this, *I will treat each observation in my dataset as an independent snapshot, rather than a part of a time series*.

### Hypothesis

I will analyze *Trait.Cold*—sales of drinks served on ice or chilled in the refrigerator.

The drinks that were coded as “cold” included iced lattes and americanos, iced teas, cold brew coffee, as well as juice and water bottles and cans from the fridge.

My hypothesis is this:

Better weather (manifested at least partly by higher temperatures, better visibility, and lack of precipitation) coincides with higher numbers of cold drinks sold.

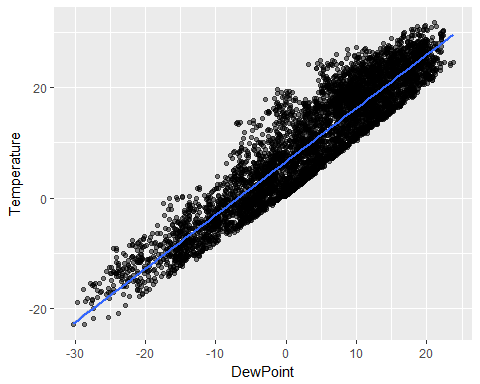
### Independent variables

First, I will choose the maximum number of independent variables to use in the model explaining sales of cold drinks.

It will be helpful to see whether any of the weather variables correlate among each other, so that I know what variables not to include in modeling to avoid multicollinearity.

The output of a correlation matrix shows that none of the variables’ correlation coefficients are exceptionally large, except for one pair: *Temperature* and *DewPoint* (0.93):

## Warning: package 'ggplot2' was built under R version 3.4.3



Pearson’s correlation test confirms that the correlation is significant:

##   
## Pearson's product-moment correlation  
##   
## data: dataset$Temperature and dataset$DewPoint  
## t = 193.57, df = 5665, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9285201 0.9353630  
## sample estimates:  
## cor   
## 0.9320246

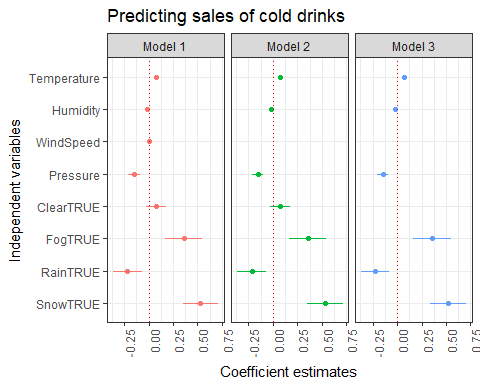
Given this, I will refrain from using *DewPoint* as a predictor in my modeling.

### Multiple linear regression

I will start building a model using all independent variables besides *DewPoint*.

Over two iterations of removing insignificant predictors (*wind speed* and *clear sky*), I arrive at three models:

library(dotwhisker)



Over each iteration, removing the insignificant predictors did not the remaining predictors, and remained stable:

summary(coldModel1)$r.squared

## [1] 0.2467212

summary(coldModel2)$r.squared

## [1] 0.2467151

summary(coldModel3)$r.squared

## [1] 0.2464314

### Cross-validation

I will need to compare the three models using cross-validation—especially since their are almost equal.

library(snowfall)  
library(dplyr)

Running a leave-one-out cross-validation procedure on the three models gives the following result:

ColdLOO1$error

## [1] 2.60314

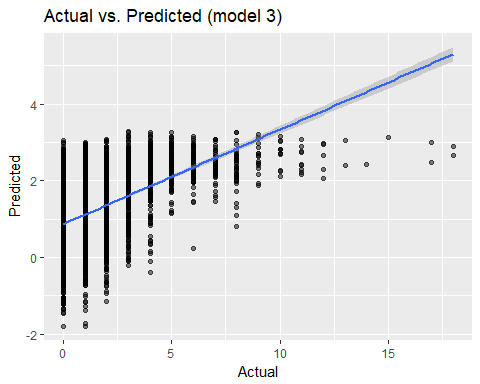
ColdLOO2$error

## [1] 2.602215

ColdLOO3$error

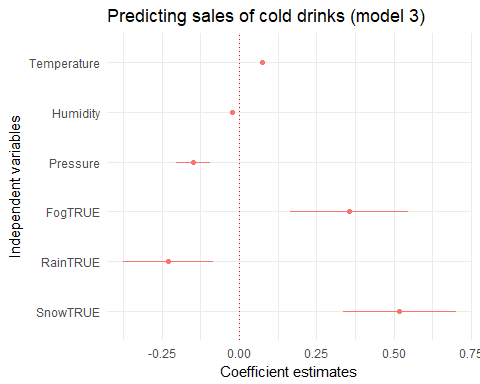
## [1] 2.602058

Error estimates get smaller for each successive model, confirming that model number 3 is the best available.



## Assessing the model

I will now examine the model more closely:



**There’s room for improvement** in my model, because:

* *The impact of the binary variables, as measured by coefficient estimates, seems unreasonably strong as measured by coefficient estimates.* This is particularly the case for *Snow*, the coefficient estimate for which is far greater than for temperature or any other variable in the model.
* *The three binary variables (Fog, Rain, and Snow) have suspiciously large confidence intervals, compared to the continous variables.*

I suppose that this may be due to **two possibilities**:

* the binary variables have *outlier observations* that influence their effects, or
* *a lack of precision in measurement* of the binary variables (Snow, Rain, and Fog) amplifies their effects compared to the continuous variables (Temperature, Humidity, and Pressure).

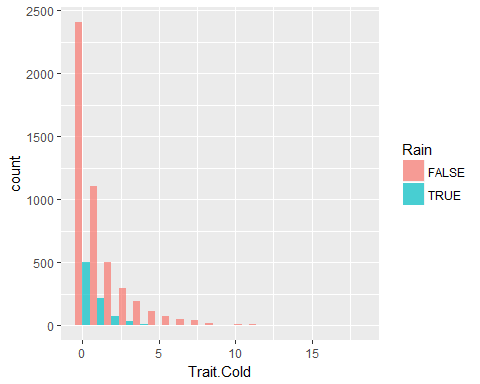
I explore these possibilities in the next section.

### Outliers

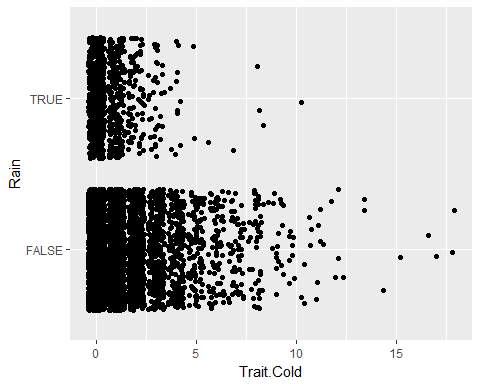
#### Rain

In the model, *Rain* is a negative predictor, with a large confidence interval and

It seems reasonable that *Rain* is a negative predictor for cold drink sales because during most of the rainy hours, zero cold drinks were purchased:

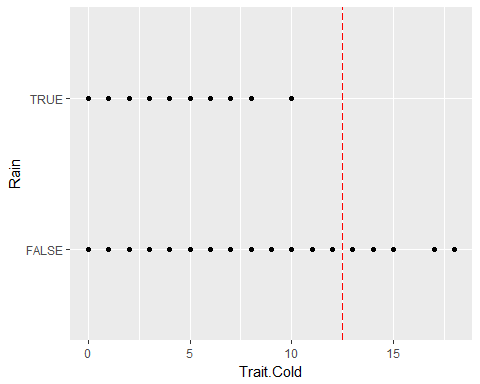


However, the following plot suggests there are some extreme outliers for hours when it did not rain:

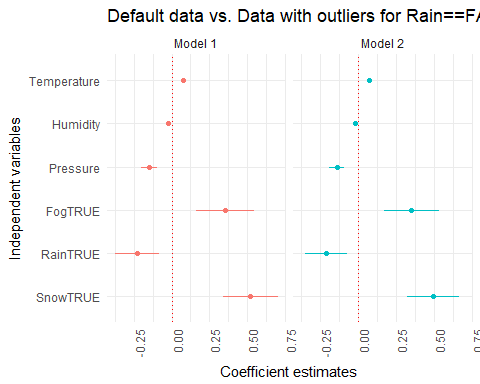


I see four-five extreme outliers. In order to remove them, I need to find an arbitrary cutoff point.

The following dot plot confirms that *Trait.Cold*=12.5 can serve as this point:



library(dplyr)  
lessOutliers <- dataset %>% filter(Trait.Cold<=12.5)



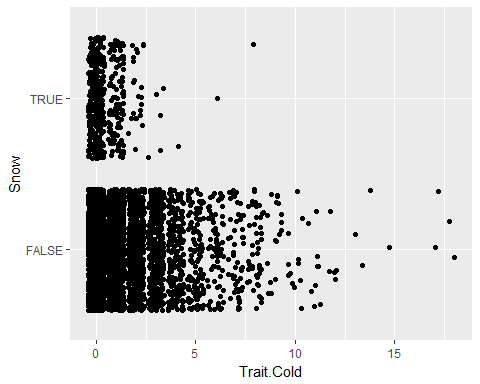
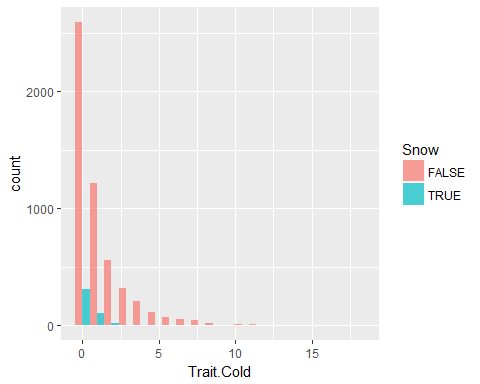
After the outlier observations were taken out, *Rain*’s coefficient estimate remains almost the same.

This suggests that **outliers are not influencing the relationship between *Trait.Cold* and *Rain***.

#### Snow

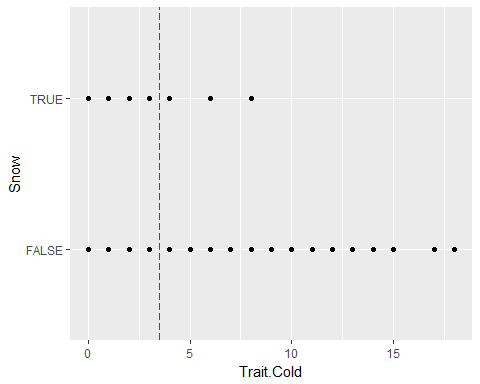
The relatioship between *Snow* and *Trait.Cold* is particularly alarming.

Commonsensically, snow should not be a significant, let alone strong, predictor for sales of cold drinks—even though this is Canada.

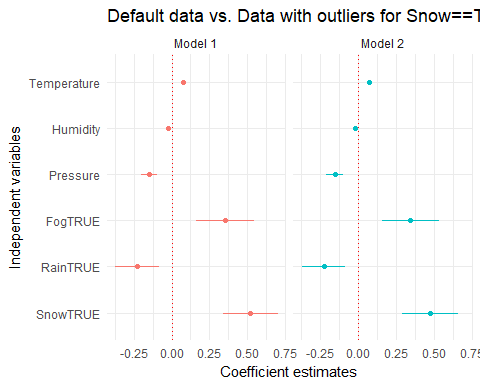


There are three outliers for hours when it snowed.

The following point plot suggests that *Trait.Cold*=3.5 is a good cutoff point to capture the three outliers.



outlierSubset <- dataset %>% filter(Snow==TRUE, Trait.Cold >=3.5)  
rm(lessOutliers)  
lessOutliers <- anti\_join(dataset, outlierSubset)



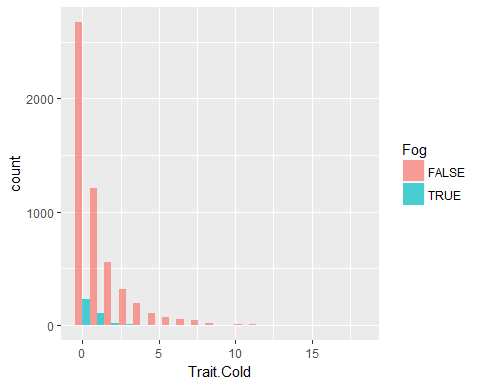
After the outlier observations were taken out, *Snow*’s coefficient estimate decreased slightly by around 0.04: far from a drastic change in the relationship with *Trait.Cold*.

This suggests that **outliers are not influencing the relationship between *Trait.Cold* and *Rain***.

#### Fog

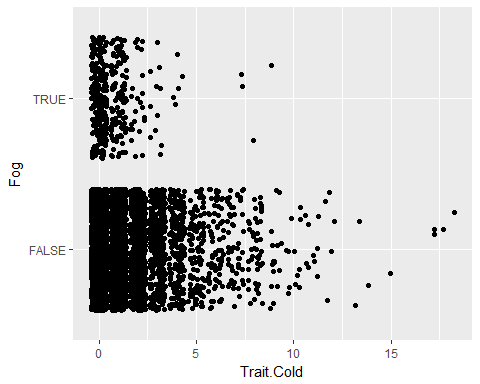
Commonsensically, *Fog* should be a *negative* predictor for cold drink sales.

However, in the model, it is halfway between Temperature and Snow as a *positive* predictor.

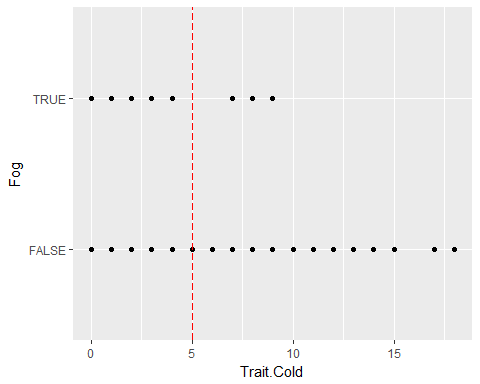


The following jitter plot shows four distinct outliers for foggy hours:

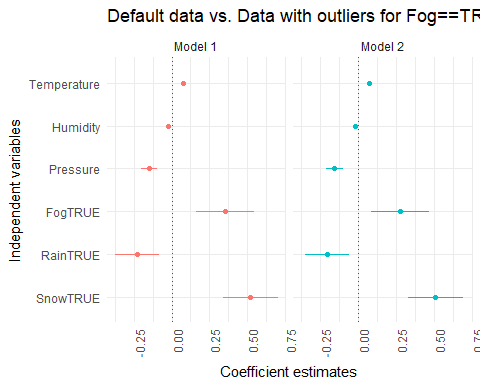
ggplot(dataset, aes(x=Trait.Cold, y=Fog)) + geom\_jitter()



The point plot confirms *Trait.Cold*=5 as the cutoff for these four outliers:



outlierSubset <- dataset %>% filter(Fog==TRUE, Trait.Cold >=5)  
rm(lessOutliers)  
lessOutliers <- anti\_join(dataset, outlierSubset)



*Fog* is still a positive significant predictor for *Trait.Cold*, although its coefficient estimate decreased by around 0.07.

### Difference in precision

Altogether, removing outliers for each of the binary variables changed coefficient estimates neither for the variables for which outliers were removed, nor for other variables in the model.

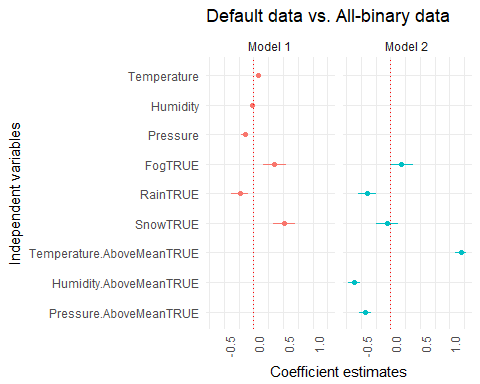
It is clear that **extreme outliers in the binary variables are not producing their stronger effects on *Trait.Cold* compared to the continuous variables**.

I supposed that **the lack of precision in measurement of the binary variables could be what amplifies the effect of *Snow*, *Rain*, and *Fog* compared to the continuous variables (*Temperature*, *Humidity*, and *Pressure*)**.

To see if this is the case, I will restructure the continuous variables, making them binary.

Then, I will run a model using binary predictors only. For now, I will only use the same variables that had made it into model #3.

dataset$Temperature.AboveMean <- ifelse(dataset$Temperature>mean(dataset$Temperature), TRUE, FALSE)  
dataset$Humidity.AboveMean <- ifelse(dataset$Humidity>mean(dataset$Humidity), TRUE, FALSE)  
dataset$Pressure.AboveMean <- ifelse(dataset$Pressure>mean(dataset$Pressure), TRUE, FALSE)  
dataset$WindSpeed.AboveMean <- ifelse(dataset$WindSpeed>mean(dataset$WindSpeed), TRUE, FALSE)



After continuous variables have recoded as binaries, their coefficient estimates are bigger compared to the variables that were previously binary—*Fog*, *Rain*, and *Snow*.

In fact, neither *Snow* nor *Fog* survived as predictors: they are both insignificant.

Confidence intervals are smallest for both *Temperature* and *Pressure*.

In this preliminary test model with all predictors formatted as binaries,

* *Temperature* has a very high coefficient estimate compared to other variables and, together with *Pressure*, has the smallest confidence intervals (intuitively),
* Both *Snow* and *Fog* do not predict sales of cold drinks (intuitively), and
* *Rain* has a negative effect on sales of cold drinks (intuitively).

The results of this model make more sense than what I saw previously when modeling using both binary and continuous variables.

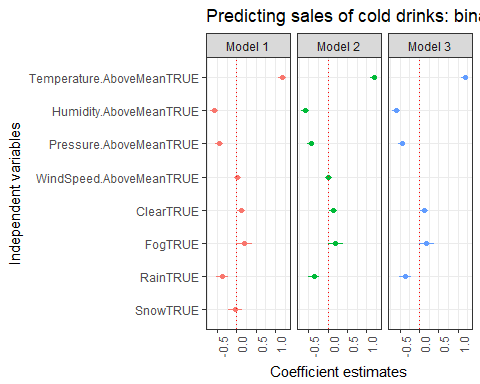
This confirms that **the lack of precision in measurement of the binary variables amplified the effect of *Snow*, *Rain*, and *Fog* compared to the continuous variables (*Temperature*, *Humidity*, and *Pressure*), since the effects were essentially reversed after converting continuous variables to binaries**.

## Final model

Given the above, I will need to create a model explaining *Trait.Cold* from scratch.

However, this time I will run the first model exclusively with binary independent variables.

After two iterations of removing insignificant predictors (*Snow* first, *Wind speed* second) produce the final model:



In this final model (model 3 on the above graph), coefficient estimates suggest that

* *above-mean temperature* is a positive predictor for the sales of cold drinks, with an effect farthest from zero compared to the other variables,
* *above-mean humidity* follows as a negative predictor,
* *above-mean pressure* and *rain* are negative predictors, both roughly at the same level, and
* both *clear sky* and *fog* are somewhat significant\* positive predictors, with their effects closest to zero compared to the other variables.

Judging by these results, I will be able to confidently report on the effects of temperature, humidity, pressure, and rain on sales of cold drinks.

Visibility, on the other hand, does not appear to make a difference for sales of cold drinks. Clear skies and fog are less significant than the other variables. Plus, they are opposites, and hence can’t both be positive predictors.

**Regarding the hypothesis for *Trait.Cold*, I can conclude that**

Higher numbers of cold drinks sold coincide with above-average temperatures.

However, above-average pressure and humidity are both associated with lower numbers of cold drinks sold.

Better visibility—includingClear skies or fog—does not appear to be a factor in sales of cold drinks.