Predicting Sales of Cold Drinks

Capstone Statistical Analysis

February 21, 2018

## Overview

In this statistical analysis submission for Springboard’s *Introduction to Data Science* workshop, I examine **the effect of weather variables on hourly sales of iced and chilled drinks at a cafe**.

After providing some background on my dataset, I motivate my choice of a linear model for the data.

Then, I build three models and choose the most parsimonious one after running a cross-validation procedure.

Finally, I take some steps to improve the model and discuss the results.

## The dataset

Now, let’s look at the dataset’s structure:

str(dataset)

## 'data.frame': 5667 obs. of 29 variables:  
## $ Day : Date, format: "2016-09-01" "2016-09-01" ...  
## $ Hour : int 10 11 12 13 14 15 17 18 19 7 ...  
## $ Season : Factor w/ 4 levels "Fall","Spring",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ ExamPeriod : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ LowSalesPeriod : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ DrinksSold : int 8 7 4 7 13 6 3 3 4 4 ...  
## $ Size.Mean : int 281 253 296 233 260 154 434 394 384 296 ...  
## $ Content.Espresso : int 6 5 2 6 9 3 0 1 1 4 ...  
## $ Content.Drip : int 0 0 1 1 4 1 0 1 1 0 ...  
## $ Content.Water : int 2 1 1 1 0 0 2 1 1 0 ...  
## $ Content.Tea : int 3 1 1 0 0 0 3 1 2 0 ...  
## $ Content.Milk : int 6 5 3 3 5 2 0 1 1 3 ...  
## $ Content.SpecialtyMilk: int 0 1 0 1 2 0 1 0 1 1 ...  
## $ Content.Chocolate : int 1 0 0 0 1 1 0 0 0 0 ...  
## $ Content.Seasonal : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Content.Juice : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Trait.Cold : int 0 0 1 0 1 0 1 0 1 0 ...  
## $ Trait.HighInSugar : int 1 0 0 0 0 1 1 0 1 0 ...  
## $ Trait.HighInCaffeine : int 0 0 1 1 4 1 0 1 1 0 ...  
## $ Trait.Froth : int 7 6 4 6 13 5 3 3 3 4 ...  
## $ Temperature : num 20.5 20.4 21.5 21 21.1 21.9 21.2 21.5 20.7 16.7 ...  
## $ DewPoint : num 12.6 12.2 12.2 14 13.2 13.1 13.2 13.3 14.5 11.9 ...  
## $ Humidity : int 60 59 55 64 60 57 60 59 67 73 ...  
## $ WindSpeed : int 16 24 16 29 27 7 14 12 25 20 ...  
## $ Pressure : num 101 101 101 101 101 ...  
## $ Clear : logi TRUE TRUE TRUE FALSE FALSE FALSE ...  
## $ Fog : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Rain : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Snow : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

head(dataset, 1)

## Day Hour Season ExamPeriod LowSalesPeriod DrinksSold Size.Mean  
## 1 2016-09-01 10 Fall FALSE FALSE 8 281  
## Content.Espresso Content.Drip Content.Water Content.Tea Content.Milk  
## 1 6 0 2 3 6  
## Content.SpecialtyMilk Content.Chocolate Content.Seasonal Content.Juice  
## 1 0 1 0 0  
## Trait.Cold Trait.HighInSugar Trait.HighInCaffeine Trait.Froth  
## 1 0 1 0 7  
## Temperature DewPoint Humidity WindSpeed Pressure Clear Fog Rain Snow  
## 1 20.5 12.6 60 16 100.94 TRUE FALSE FALSE FALSE

tail(dataset, 1)

## Day Hour Season ExamPeriod LowSalesPeriod DrinksSold Size.Mean  
## 5667 2017-11-22 9 Winter FALSE TRUE 21 297  
## Content.Espresso Content.Drip Content.Water Content.Tea Content.Milk  
## 5667 13 3 2 3 12  
## Content.SpecialtyMilk Content.Chocolate Content.Seasonal  
## 5667 2 1 0  
## Content.Juice Trait.Cold Trait.HighInSugar Trait.HighInCaffeine  
## 5667 0 1 2 3  
## Trait.Froth Temperature DewPoint Humidity WindSpeed Pressure Clear  
## 5667 16 1.6 0.7 94 15 101.08 FALSE  
## Fog Rain Snow  
## 5667 FALSE TRUE FALSE

Every observation is one hour of the cafe’s operations.

Each of the *Content.\** and *Trait.\** variables represents the number of drinks sold that had the noted ingredient or trait.

For example, an observation can be partially spelled out like this:

On September 1st, 2016, from 10:00 to 10:59AM, the cafe sold 6 espresso-based drinks,   
6 drinks containing milk, 1 drink containing chocolate, 0 drinks that were chilled or iced, and 7 frothed drinks.   
During that time, the sky was clear, there were no fog, rain, or snow,   
and the temperature was at 20.5 degrees C.

## Modeling

While my dataset features *time-series 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 are outside the scope of this workshop.** Given that I also lack prior academic training necessary to apply time series models, I will use a model that is within the learning parameters of the workshop.
* 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 a stronger effect.

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

For the purposes of modeling, I will also assume that this data *isn’t* count data.

### Hypothesis

In my dataset, I have many dependent variables to choose from:

## [1] "Content.Espresso" "Content.Drip"   
## [3] "Content.Water" "Content.Tea"   
## [5] "Content.Milk" "Content.SpecialtyMilk"  
## [7] "Content.Chocolate" "Content.Seasonal"   
## [9] "Content.Juice" "Trait.Cold"   
## [11] "Trait.HighInSugar" "Trait.HighInCaffeine"   
## [13] "Trait.Froth"

Here, I will analyze *Trait.Cold*—sales of drinks served on ice or served chilled from 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.

I picked the “cold” trait because, intuitively, this variable would likely have to be explained by higher temperatures and nice weather.

Examining this variable would also be a good trial test for my dataset: it should at least be possible to establish a link between cold drinks and warmer weather.

On the other hand, there is a real possibility of delivering value to the client. Normally, the client makes an emphasis on cold drinks (in particular, cold brew coffee) in the summer. What if the demand for cold drinks is maintained in colder seasons? In this case, the client could, for instance, start offerring cold brew coffee in the winter.

Therefore, my hypothesis is this:

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

### Independent variables

Let’s choose the maximum number of independent variables to use in the model explaining sales of cold drinks.

Here are all the independent (weather) variables available in the dataset:

## [1] "Temperature" "DewPoint" "Humidity" "WindSpeed" "Pressure"   
## [6] "Clear" "Fog" "Rain" "Snow"

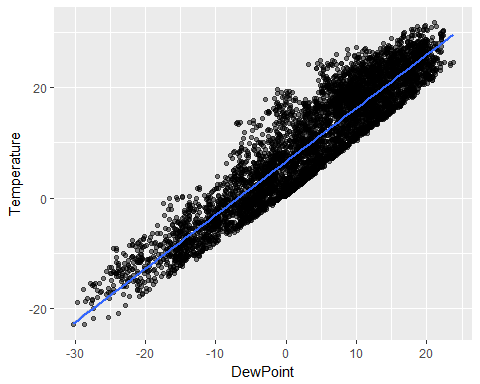
It will be helpful to see whether any of the weather variables correlate among each other, so we know what variables not to include in the models later to avoid multicollinearity.

round(cor(dataset[,21:length(colnames(dataset))]), 2)

## Temperature DewPoint Humidity WindSpeed Pressure Clear Fog  
## Temperature 1.00 0.93 -0.17 -0.10 -0.15 0.14 -0.08  
## DewPoint 0.93 1.00 0.20 -0.11 -0.28 0.00 0.05  
## Humidity -0.17 0.20 1.00 -0.06 -0.37 -0.36 0.42  
## WindSpeed -0.10 -0.11 -0.06 1.00 -0.34 -0.11 -0.08  
## Pressure -0.15 -0.28 -0.37 -0.34 1.00 0.30 -0.18  
## Clear 0.14 0.00 -0.36 -0.11 0.30 1.00 -0.18  
## Fog -0.08 0.05 0.42 -0.08 -0.18 -0.18 1.00  
## Rain 0.01 0.17 0.50 0.06 -0.28 -0.28 0.39  
## Snow -0.39 -0.32 0.20 0.16 -0.21 -0.20 0.00  
## Rain Snow  
## Temperature 0.01 -0.39  
## DewPoint 0.17 -0.32  
## Humidity 0.50 0.20  
## WindSpeed 0.06 0.16  
## Pressure -0.28 -0.21  
## Clear -0.28 -0.20  
## Fog 0.39 0.00  
## Rain 1.00 -0.09  
## Snow -0.09 1.00

*Temperature* and *DewPoint* are strongly correlated (0.93).

Let’s plot the correlation:



Now, let’s run a correlation test:

##   
## 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

This correlation is significant, so I will refrain from using *DewPoint* as a predictor in my modeling.

### Multiple linear regression

Let’s run a multiple linear regression model on *Trait.Cold* using all independent variables besides *DewPoint*:

##   
## Call:  
## lm(formula = Trait.Cold ~ Temperature + Humidity + WindSpeed +   
## Pressure + Clear + Fog + Rain + Snow, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0084 -0.9671 -0.2532 0.5777 15.2659   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.1636025 3.1547541 5.441 5.53e-08 \*\*\*  
## Temperature 0.0740044 0.0022607 32.735 < 2e-16 \*\*\*  
## Humidity -0.0212901 0.0016853 -12.633 < 2e-16 \*\*\*  
## WindSpeed 0.0005653 0.0026461 0.214 0.830834   
## Pressure -0.1523836 0.0305989 -4.980 6.55e-07 \*\*\*  
## ClearTRUE 0.0759493 0.0517678 1.467 0.142401   
## FogTRUE 0.3579156 0.0977518 3.661 0.000253 \*\*\*  
## RainTRUE -0.2172513 0.0750010 -2.897 0.003786 \*\*   
## SnowTRUE 0.5297010 0.0931995 5.684 1.39e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.612 on 5658 degrees of freedom  
## Multiple R-squared: 0.2467, Adjusted R-squared: 0.2457   
## F-statistic: 231.6 on 8 and 5658 DF, p-value: < 2.2e-16

*Wind speed* and *clear sky* are insignificant. Between the two, *wind speed* is less significant, so let’s drop it first:

##   
## Call:  
## lm(formula = Trait.Cold ~ Temperature + Humidity + Pressure +   
## Clear + Fog + Rain + Snow, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0117 -0.9655 -0.2553 0.5737 15.2666   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.433355 2.890815 6.031 1.74e-09 \*\*\*  
## Temperature 0.073916 0.002223 33.256 < 2e-16 \*\*\*  
## Humidity -0.021378 0.001634 -13.080 < 2e-16 \*\*\*  
## Pressure -0.154887 0.028265 -5.480 4.44e-08 \*\*\*  
## ClearTRUE 0.075512 0.051723 1.460 0.144366   
## FogTRUE 0.355832 0.097256 3.659 0.000256 \*\*\*  
## RainTRUE -0.215482 0.074536 -2.891 0.003855 \*\*   
## SnowTRUE 0.530913 0.093019 5.708 1.20e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.612 on 5659 degrees of freedom  
## Multiple R-squared: 0.2467, Adjusted R-squared: 0.2458   
## F-statistic: 264.8 on 7 and 5659 DF, p-value: < 2.2e-16

In this model, *clear sky* is still insignificant, while all the other predictors remained significant.

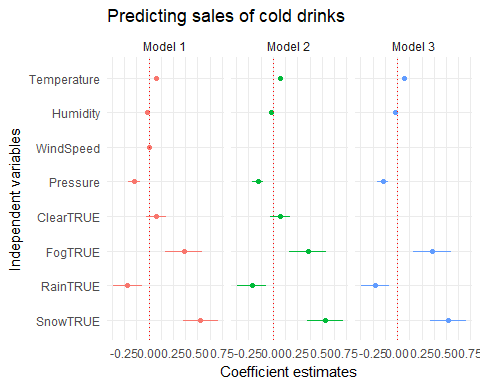
Getting rid of *clear sky* as a predictor and running the model again:

##   
## Call:  
## lm(formula = Trait.Cold ~ Temperature + Humidity + Pressure +   
## Fog + Rain + Snow, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0509 -0.9633 -0.2601 0.5759 15.3180   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.768985 2.855057 5.873 4.51e-09 \*\*\*  
## Temperature 0.074212 0.002214 33.524 < 2e-16 \*\*\*  
## Humidity -0.021767 0.001613 -13.498 < 2e-16 \*\*\*  
## Pressure -0.147832 0.027852 -5.308 1.15e-07 \*\*\*  
## FogTRUE 0.355421 0.097265 3.654 0.00026 \*\*\*  
## RainTRUE -0.229120 0.073956 -3.098 0.00196 \*\*   
## SnowTRUE 0.518636 0.092647 5.598 2.27e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.613 on 5660 degrees of freedom  
## Multiple R-squared: 0.2464, Adjusted R-squared: 0.2456   
## F-statistic: 308.5 on 6 and 5660 DF, p-value: < 2.2e-16

All remaining predictors are significant, and has remained stable.

Let’s plot coefficient estimates across the models:

library(dotwhisker)



Removing the insignificant predictors hasn’t affected the remaining predictors.

Confidence intervals stayed relatively stable and the direction of impact (positive or negative) on the DV didn’t change either.

It appears we have built the most parsimonious model.

### Cross-validation

Let’s run the following LOO cross-validation procedure:

library(snowfall)  
library(dplyr)

LeaveOneOut <- function(formula, mFrame) {  
 require(dplyr)  
   
 formula <- as.formula(formula)  
 DV <- all.vars(formula)[1]  
  
 pred <- numeric(dim(mFrame)[1])  
 for (i in 1:dim(mFrame)[1]) {  
 tR <- mFrame[i, ]  
 mF <- mFrame[-i, ]  
 lmModel <- lm(formula, data = mF)  
 pred[i] <- predict(lmModel, newdata = tR)  
 }  
 predError\_looModel <- sum((mFrame[,which(colnames(mFrame)==as.character(DV))] - pred)^2)/dim(mFrame)[1]  
  
 return(list("predictions"=pred, "error"=predError\_looModel))  
}

I’m [using *snowfall*](https://hernanresnizky.com/2014/01/10/quick-guide-to-parallel-r-with-snow/) to speed up the procedure:

clus <- makeCluster(4)  
clusterExport(clus, c("LeaveOneOut", "dataset"))  
  
dataset$Trait.Cold.Sqrt <- sqrt(dataset$Trait.Cold)  
ColdLOO1 <- LeaveOneOut("Trait.Cold ~ Temperature + Humidity + WindSpeed + Pressure + Clear + Fog + Rain + Snow", dataset)  
ColdLOO2 <- LeaveOneOut("Trait.Cold ~ Temperature + Humidity + Pressure + Clear + Fog + Rain + Snow", dataset)  
ColdLOO3 <- LeaveOneOut("Trait.Cold ~ Temperature + Humidity + Pressure + Fog + Rain + Snow", dataset)  
  
stopCluster(clus)

The error got a little bit smaller with each successive model:

ColdLOO1$error  
ColdLOO2$error  
ColdLOO3$error

It appears that model #3 is best, given the error estimates.

Plotting actual vs. predicted for LOO on model 3:

## Improving the model

We can try and improve the model a bit by transforming the dependent variable. Taking a log() doesn’t seem like a good idea, because many observations in the dataset had 0 cold drinks sold:

table(dataset$Trait.Cold)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14   
## 2904 1320 579 326 204 111 74 53 45 17 12 9 5 2 1   
## 15 17 18   
## 1 2 2

So, it would be better to take a square root of the dependent variable.

This shouldn’t interfere with my goals, since I’m not interested in interpreting the coefficients numerically.

Rather, I’m interested in the directionality of the relationships and how the coefficients stack up against each other.

##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature + Humidity + Pressure +   
## Fog + Rain + Snow, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5477 -0.5227 -0.1224 0.5389 2.8798   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.9851839 1.2504274 5.586 2.43e-08 \*\*\*  
## Temperature 0.0327024 0.0009695 33.731 < 2e-16 \*\*\*  
## Humidity -0.0101929 0.0007063 -14.432 < 2e-16 \*\*\*  
## Pressure -0.0588972 0.0121981 -4.828 1.41e-06 \*\*\*  
## FogTRUE 0.1691919 0.0425991 3.972 7.22e-05 \*\*\*  
## RainTRUE -0.0710859 0.0323904 -2.195 0.0282 \*   
## SnowTRUE 0.2124757 0.0405766 5.236 1.70e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7062 on 5660 degrees of freedom  
## Multiple R-squared: 0.2511, Adjusted R-squared: 0.2503   
## F-statistic: 316.2 on 6 and 5660 DF, p-value: < 2.2e-16

The directions of all relationships stayed the same, and improved a little bit:

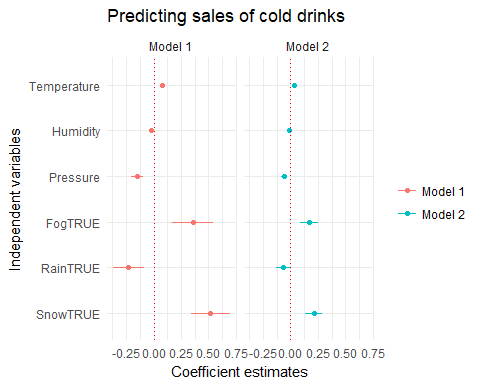
summary(coldModel3)$r.squared

## [1] 0.2464314

summary(coldModel4)$r.squared

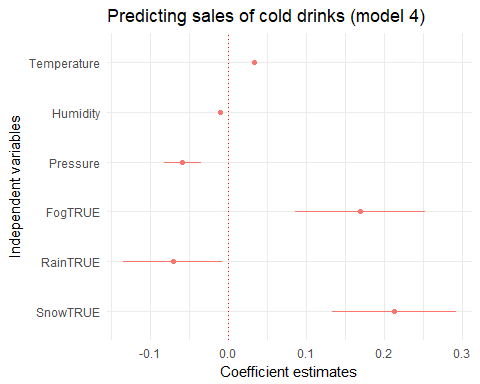
## [1] 0.251056

Using *dotwhisker* package, let’s plot regression coefficient estimates for the two models (DV as is vs. DV as square root) and compare:



The final model (with square root of Trait.Cold as DV) seems like an improvement, since confidence intervals for all variables are smaller and the effects for Fog, Rain, and Snow seem smaller, which is reasonable. (Snow in particular moved from ~ 0.5 estimate to ~ 0.25, and Rain still is a negative effect on cold drink sales, but not as much.)

Let’s now plot only the fourth model’s coefficients:



There’s room for improvement in this model.

We have tiny confidence intervals for Temperature (a positive predictor), and Humidity (a negative predictor). Compared to all other variables, they have the smallest effects.

We also see a larger effect of Fog, Rain, and Snow.

This is particularly strange: how is Snow a stronger predictor for the sale of cold drinks than Temperature?

**This may be due to outliers in the binary variables, or due to the very fact that they’re binary, as the other ones (Temperature, Humidity, Pressure) are continous.**

Let’s explore the two possibilities.

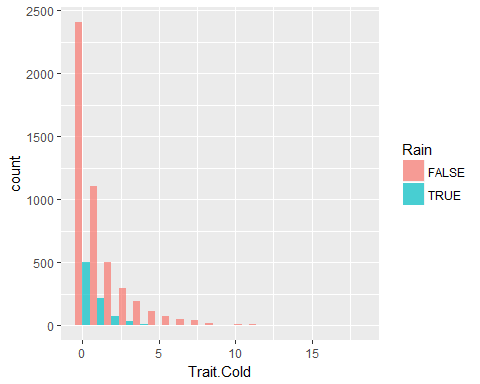
### Outliers

Could it be that outliers are impacting the results?

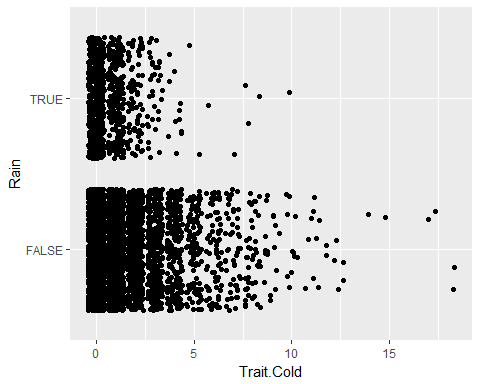
#### Rain

Rain has a large confidence interval, but it’s still definitely a negative predictor.

It seems reasonable that Rain is a negative predictor, because most observations when it rained concentrate around 0 cold drinks purchased:

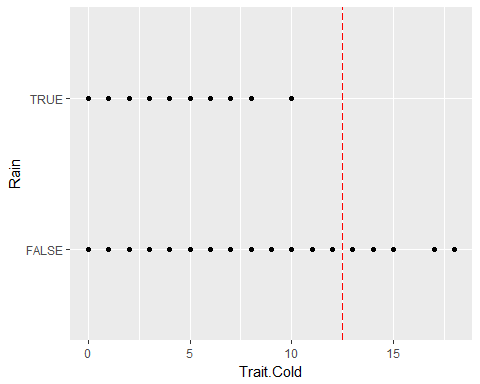


We can see there are some outliers for Rain==FALSE, maybe they’re influencing the relationship?



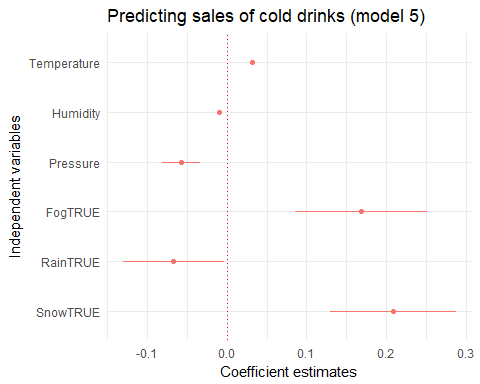
Let’s try and run the model on the dataset with some of these outliers exluded. Will Rain still be a negative predictor?

Let’s eyeball 12.5 as an arbitrary cutoff:



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

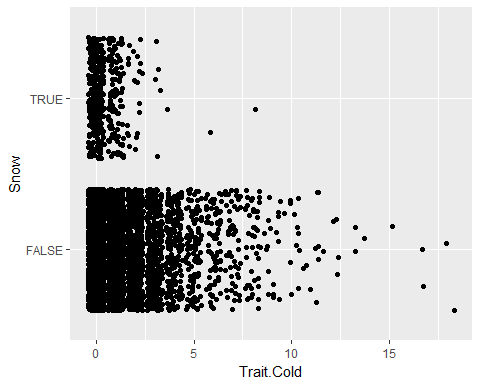
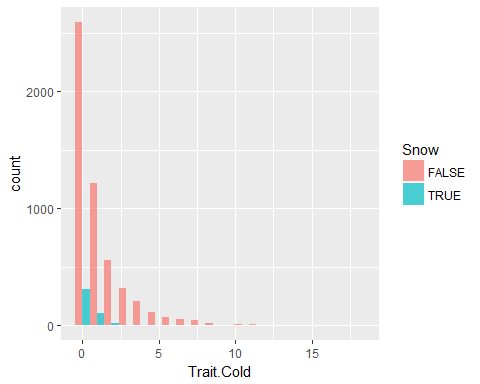
##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature + Humidity + Pressure +   
## Fog + Rain + Snow, data = lessOutliers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5333 -0.5221 -0.1243 0.5411 2.3480   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.8619802 1.2402498 5.533 3.29e-08 \*\*\*  
## Temperature 0.0321360 0.0009628 33.376 < 2e-16 \*\*\*  
## Humidity -0.0101534 0.0007005 -14.495 < 2e-16 \*\*\*  
## Pressure -0.0576861 0.0120987 -4.768 1.91e-06 \*\*\*  
## FogTRUE 0.1683087 0.0422338 3.985 6.83e-05 \*\*\*  
## RainTRUE -0.0668876 0.0321156 -2.083 0.0373 \*   
## SnowTRUE 0.2082115 0.0402317 5.175 2.35e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7002 on 5652 degrees of freedom  
## Multiple R-squared: 0.248, Adjusted R-squared: 0.2472   
## F-statistic: 310.7 on 6 and 5652 DF, p-value: < 2.2e-16



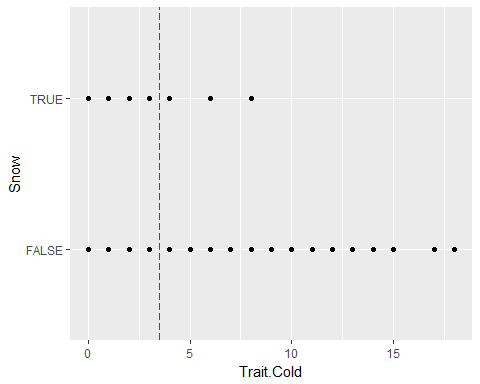
Rain’s confidence interval is still significant within the negative bounds, so it doesn’t appear that the more extreme outliers are influencing the Rain-Cold Drink Sales relationship.

#### Snow

Let’s try and do the same for Snow. Maybe there are outliers that produce this weird positive effect of Snow on cold drink sales that’s way stronger than temperature?



It seems like there are three outliers for Snow==TRUE that have about 4 and more on Trait.Cold. Could it be them producing the effect?

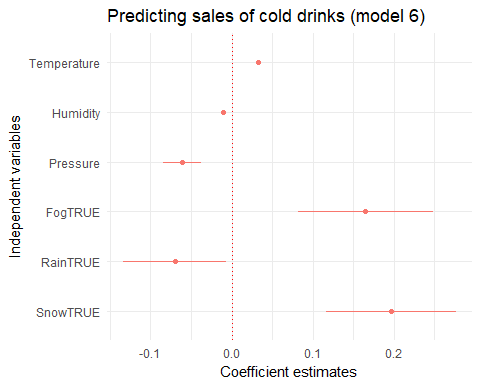


The point plot confirms the outliers are at 3.5 and over. Let’s try to get rid of them and see what happens:

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

## Joining, by = c("Day", "Hour", "Season", "ExamPeriod", "LowSalesPeriod", "DrinksSold", "Size.Mean", "Content.Espresso", "Content.Drip", "Content.Water", "Content.Tea", "Content.Milk", "Content.SpecialtyMilk", "Content.Chocolate", "Content.Seasonal", "Content.Juice", "Trait.Cold", "Trait.HighInSugar", "Trait.HighInCaffeine", "Trait.Froth", "Temperature", "DewPoint", "Humidity", "WindSpeed", "Pressure", "Clear", "Fog", "Rain", "Snow")

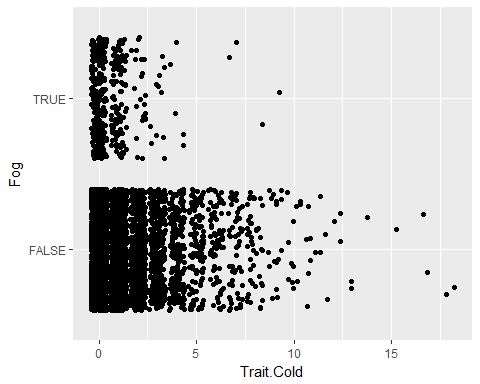
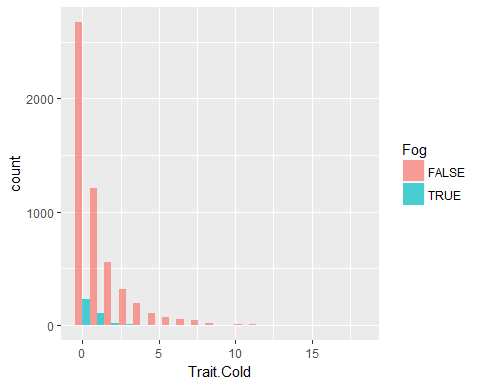
##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature + Humidity + Pressure +   
## Fog + Rain + Snow, data = lessOutliers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5494 -0.5221 -0.1213 0.5386 2.8792   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.2148088 1.2490957 5.776 8.06e-09 \*\*\*  
## Temperature 0.0325898 0.0009679 33.670 < 2e-16 \*\*\*  
## Humidity -0.0102488 0.0007051 -14.535 < 2e-16 \*\*\*  
## Pressure -0.0611171 0.0121849 -5.016 5.44e-07 \*\*\*  
## FogTRUE 0.1649844 0.0425765 3.875 0.000108 \*\*\*  
## RainTRUE -0.0699480 0.0323314 -2.163 0.030547 \*   
## SnowTRUE 0.1964327 0.0406259 4.835 1.37e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7048 on 5657 degrees of freedom  
## Multiple R-squared: 0.2526, Adjusted R-squared: 0.2518   
## F-statistic: 318.6 on 6 and 5657 DF, p-value: < 2.2e-16



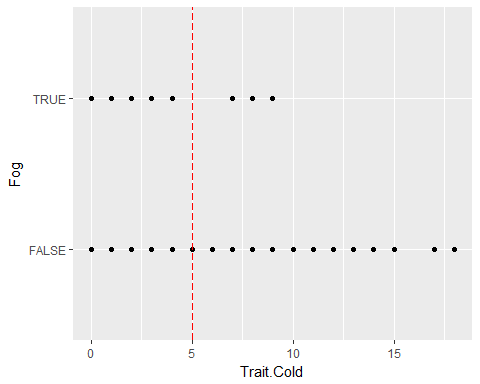
The effect is a little bit smaller, but still more or less in the same spot.

#### Fog

For good measure, let’s try and do the same for Fog.



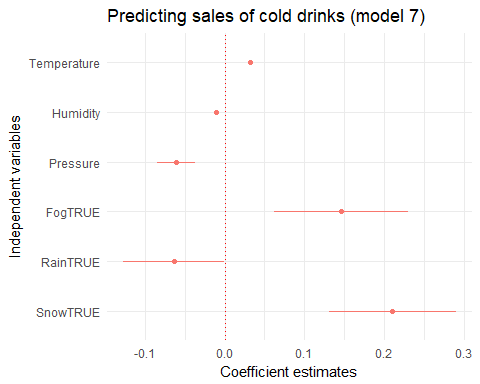
If anything, *no fog* should be a predictor for cold drinks. Still, let’s remove the outliers for FogTRUE (I’m eyeballing the cutoff at 5):



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

## Joining, by = c("Day", "Hour", "Season", "ExamPeriod", "LowSalesPeriod", "DrinksSold", "Size.Mean", "Content.Espresso", "Content.Drip", "Content.Water", "Content.Tea", "Content.Milk", "Content.SpecialtyMilk", "Content.Chocolate", "Content.Seasonal", "Content.Juice", "Trait.Cold", "Trait.HighInSugar", "Trait.HighInCaffeine", "Trait.Froth", "Temperature", "DewPoint", "Humidity", "WindSpeed", "Pressure", "Clear", "Fog", "Rain", "Snow")

##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature + Humidity + Pressure +   
## Fog + Rain + Snow, data = lessOutliers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5467 -0.5220 -0.1227 0.5405 2.8824   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.2057616 1.2490627 5.769 8.40e-09 \*\*\*  
## Temperature 0.0324767 0.0009689 33.518 < 2e-16 \*\*\*  
## Humidity -0.0102504 0.0007052 -14.535 < 2e-16 \*\*\*  
## Pressure -0.0610246 0.0121846 -5.008 5.66e-07 \*\*\*  
## FogTRUE 0.1463350 0.0427867 3.420 0.00063 \*\*\*  
## RainTRUE -0.0635290 0.0323710 -1.963 0.04975 \*   
## SnowTRUE 0.2103983 0.0405082 5.194 2.13e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.705 on 5656 degrees of freedom  
## Multiple R-squared: 0.2507, Adjusted R-squared: 0.25   
## F-statistic: 315.5 on 6 and 5656 DF, p-value: < 2.2e-16



Same result for fog: the effect is a little bit smaller, but still significant.

### Transforming IVs

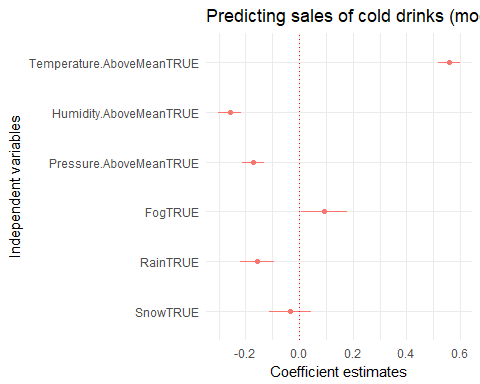
Since the fact that snow and fog are positive predictors for cold drinks, and rain is a negative predictor is *not* caused by outliers, I see one other possible cause:

**the lack of precision in measurement (is it snowing right now? Yes or No) amplifies the effect of Snow, Rain, and Fog compared to the continuous variables (Temperature, Humidity, Pressure).**

If this is in fact the case, I could restructure the continuous variables according to their means (use binary variables coded TRUE if above mean), and run the model using binary predictors only (for now, using the same variables that made it so far into the model).

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)

##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature.AboveMean + Humidity.AboveMean +   
## Pressure.AboveMean + Fog + Rain + Snow, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1946 -0.4662 -0.1946 0.5374 3.2185   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.63669 0.02416 26.358 < 2e-16 \*\*\*  
## Temperature.AboveMeanTRUE 0.55792 0.02130 26.199 < 2e-16 \*\*\*  
## Humidity.AboveMeanTRUE -0.25899 0.02247 -11.526 < 2e-16 \*\*\*  
## Pressure.AboveMeanTRUE -0.17052 0.02104 -8.104 6.45e-16 \*\*\*  
## FogTRUE 0.09246 0.04360 2.121 0.034 \*   
## RainTRUE -0.15767 0.03272 -4.819 1.48e-06 \*\*\*  
## SnowTRUE -0.03373 0.03991 -0.845 0.398   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7394 on 5660 degrees of freedom  
## Multiple R-squared: 0.179, Adjusted R-squared: 0.1781   
## F-statistic: 205.6 on 6 and 5660 DF, p-value: < 2.2e-16



So, now that continuous variables have been transformed into binaries, their effect is much more strongly felt compared to Fog, Rain, and Snow.

Intuitively, high Temperature has a very high coefficient compared to other variables, and the confidence interval is rather small.

Snow didn’t survive as predictor: it’s now insignificant, and most of the confidence interval is in the negative.

On the other hand, both Fog and Rain are still significant predictors.

Intuitively, we can be more sure about Rain as a negative predictor compared to being sure about Fog as a positive predictor (Rain’s confidence interval is farther from 0).

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

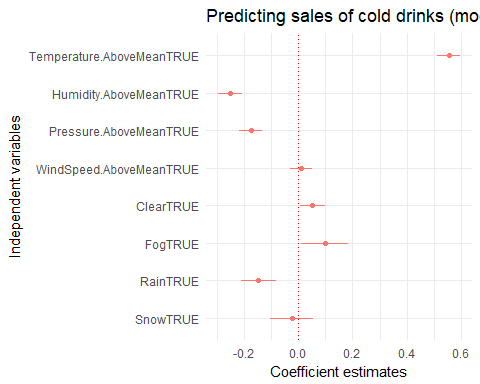
* Snow isn’t a predictor for sales of cold drinks (intuitively),
* Rain has a negative effect on sales of cold drinks (intuitively),
* Foggy weather definitely has a positive effect on sales of cold drinks (counter-intuitively), but this effect may be very small to the point where it’s negligible.

This makes much more sense than what we were seeing previously when modeling using both binary and continuous variables.

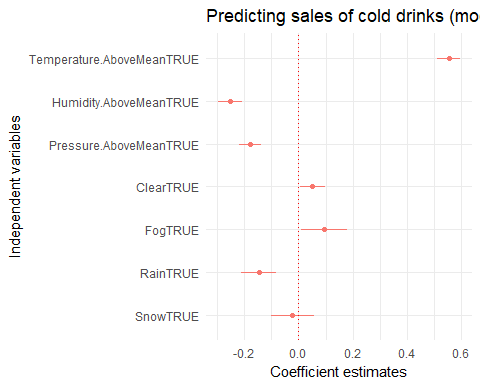
## Final model

Now, let’s create the most parsimonious model again, this time running it with all continuous variables converted to binaries:

##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature.AboveMean + Humidity.AboveMean +   
## Pressure.AboveMean + WindSpeed.AboveMean + Clear + Fog +   
## Rain + Snow, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2295 -0.4907 -0.1879 0.5609 3.1977   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.614271 0.029313 20.955 < 2e-16 \*\*\*  
## Temperature.AboveMeanTRUE 0.554282 0.021546 25.725 < 2e-16 \*\*\*  
## Humidity.AboveMeanTRUE -0.251130 0.022822 -11.004 < 2e-16 \*\*\*  
## Pressure.AboveMeanTRUE -0.175204 0.022045 -7.948 2.28e-15 \*\*\*  
## WindSpeed.AboveMeanTRUE 0.009366 0.020892 0.448 0.6539   
## ClearTRUE 0.051588 0.023498 2.195 0.0282 \*   
## FogTRUE 0.097249 0.043848 2.218 0.0266 \*   
## RainTRUE -0.146784 0.033247 -4.415 1.03e-05 \*\*\*  
## SnowTRUE -0.023411 0.040367 -0.580 0.5620   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7392 on 5658 degrees of freedom  
## Multiple R-squared: 0.1797, Adjusted R-squared: 0.1785   
## F-statistic: 154.9 on 8 and 5658 DF, p-value: < 2.2e-16

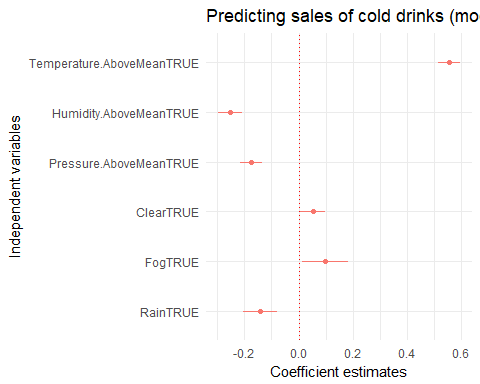


WindSpeed isn’t surviving as predictor, neither does Snow. Removing WindSpeed first:



Snow still doesn’t survive as predictor, removing:

##   
## Call:  
## lm(formula = sqrt(Trait.Cold) ~ Temperature.AboveMean + Humidity.AboveMean +   
## Pressure.AboveMean + Clear + Fog + Rain, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2259 -0.4932 -0.1860 0.5596 3.1929   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.61652 0.02385 25.845 < 2e-16 \*\*\*  
## Temperature.AboveMeanTRUE 0.55653 0.02050 27.145 < 2e-16 \*\*\*  
## Humidity.AboveMeanTRUE -0.25442 0.02234 -11.390 < 2e-16 \*\*\*  
## Pressure.AboveMeanTRUE -0.17611 0.02108 -8.354 < 2e-16 \*\*\*  
## ClearTRUE 0.05281 0.02326 2.270 0.0232 \*   
## FogTRUE 0.09637 0.04355 2.213 0.0270 \*   
## RainTRUE -0.14203 0.03246 -4.375 1.23e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7391 on 5660 degrees of freedom  
## Multiple R-squared: 0.1796, Adjusted R-squared: 0.1787   
## F-statistic: 206.5 on 6 and 5660 DF, p-value: < 2.2e-16



In this final model, we see that

* *above-mean temperature* is a strong positive predictor for the sales of cold drinks,
* *above-mean humidity* follows 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 close to zero.

It appears that we can confidently report the effects of temperature, humidity, pressure, and rain on sales of cold drinks.

Visibility doesn’t appear to make a difference for sales of cold drinks, judging by the effects of clear skies and fog—the two are essentially opposites, and can’t both be positive predictors.