Analysis of Pharmaceutical Manufacturing Data with a Logistic Regression Model

by Kimberly Schveder

Advisor: Richard Fan

December 12, 2017

Math 496H

Abstract

In the case study about addressing the precipitation in a particular generic pharmaceutical drug, a logistic regression was used to model the probability of consumer complaints and the relevant explanatory variables. We examined the significance of each of the explanatory variables. Interactions were discovered between some of the explanatory variables and an association was discovered between two of the explanatory variables.

Key words: Pharmaceutical, Manufacturing, Logistic Regression Modeling, CMH statistic, Cochran-Mantel-Haenszel Statistic, Chi-Square Statistic

Table of Contents

1. Introduction 4

2. Background 4-9

3. Statistical Methods 9-15

4. Conclusions and Further Research 15

5. References 15-17

6. Acknowledgments 17

7. Appendix 17-18

Analysis of Pharmaceutical Manufacturing Data with a Logistic Regression Model

Introduction

Statistical consulting is often used in business and industry, such as in manufactured product applications, service business applications, and process improvement. According to Hahn (2011), "the major goal of statistics during field support is to help develop an optimum servicing and problem avoidance strategy and leverage the information on field performance to build improved product in the future." Here, the focus will be on the manufactured product applications and specifically with field support and tracking. The data set that will be investigated is one from a pharmaceutical manufacturer. It tracks several characteristics of the approved drug that was placed on the market and the status of consumer complaint, and the date the consumer made the complaint. Here, the assessment of any association between the characteristics of the drugs from this batch and the number of consumer complaints will be completed with logistic regression and a logistic regression model. This will be completed in hope to make improvement in the drug quality in the future.

Background

Logistic regression is a popular statistical model that will be used to answer research questions about a variety of data on generic pharmaceutical drugs, taken from a pharmaceutical manufacturer, specifically for finding the probability of a consumer complaints given several potential explanatory variables. Such a model is so popular to use because the logistic model is based on the logistic function, which provides estimates that have to lie between zero and one and has a S-shape (Kleinbaum & Klein 2010), as shown in Figures 1 and 2. By definition, "logistic regression is a mathematical modeling approach that can be used to describe the relationship of several Xs to a dichotomous dependent variable" (Kleinbaum & Klein 2010). In this case of the pharmaceutical manufacturer's data set, our dichotomous dependent variables, Yes (Y, or 1) or No (N, or 0), with seventeen continuous potential in-process and product quality parameters as the explanatory inputs.

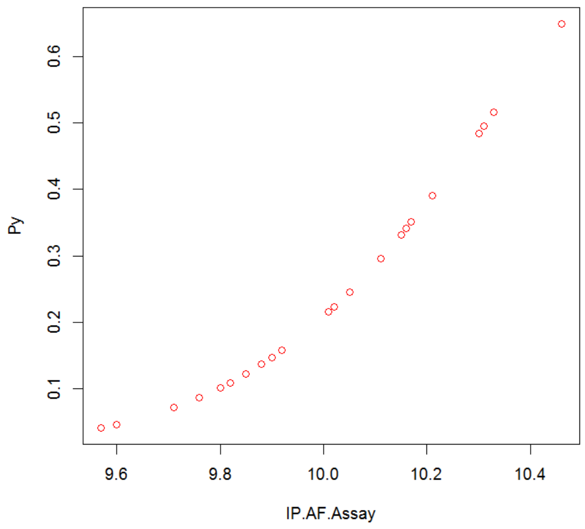


Figure 1

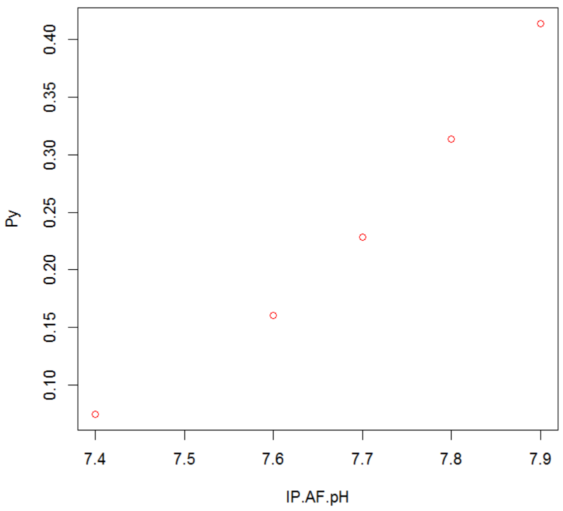


Figure 2

Figures 1 and 2: Examples of Logistic Regression Model Graphs

Here is a list of the seventeen in-process and finished product quality parameters that will be used to develop the model:

The response output is Complaints: 0 for "No consumer complaint reported" or 1 for "At least one consumer complaint reported."

1. before-filtration results for assay (IP.BF.Assay)
2. before-filtration results for pH (IP.BF.pH)
3. before-filtration results for density in g/mL (IP.BF.Density)
4. after-filtration results for assay in mg/mL (IP.AF.Assay)
5. after-filtration results for pH (IP.AF.pH)
6. after-filtration results for density in g/mL (IP.AF.Density)
7. finished product results for pH (FP.pH)
8. O_2 in headspace, in percentage form (O2.in.headspace.pcnt)
9. Assay, in percentage form (Assay.pcnt)
10. Total Impurities, in percentage form (Total.Impurities.pcnt)
11. APHA color, in APHA units (Color)
12. 10 µm HIAC particles (HIAC.gt.10.um.Particles)
13. 25 µm HIAC particles (HIAC.gt.25.um.particles)
14. Initial pH, before Log 1 (Initial.pH.Before.Log1)
15. Amount of NaOH added Ph Log 1, in mL (Amt.NaOH.added.Ph.Log1)
16. pH before Log 2 (pH.Before.Log.2)
17. Amount NaOH added Ph Log 2 in mL (Amt.NaOH.added.Ph.Log2)

The abbreviations of the variables as seen in the data set are shown in parentheses next to the actual name of the variable.

Here is a list of the definitions of some the chemical terms, so that the model can be interpreted in context:

1. pH is "a measure of acidity and alkalinity of a solution that is a number on a scale on which a value of 7 represents neutrality and lower numbers indicate increasing acidity and higher numbers increasing alkalinity..." (Merriam-Webster).
2. Assay is "to analyze (something, such as an ore) for one or more specific components" (Merriam-Webster).
3. Density is "the mass of a substance per unit volume" (Merriam-Webster).
4. Headspace is "the volume above a liquid or solid in a closed container " (Merriam-Webster).
5. "Impurity refers to the substances inside a confined amount of liquid, gas, or solid, which differ from the chemical composition of the material or compound. ... During production, impurities may be purposely, accidentally, inevitably, or incidentally added into the substance. ... Standards have been established by various organizations that attempt to define the permitted levels of various impurities in a manufactured product" (pharma-iq.com).
6. "The APHA color measurement method measures the yellow hue in liquids in comparison to a platinum cobalt reference solution" (hunterlab.com). Here, APHA color is referred to as color.
7. HIAC particles are counted. HIAC is one of Beckman-Coulter's trademark brands (BC is a company that makes particle counters). If you see a reference to "HIAC particles" (which is unlikely but possible) then what is actually meant would be a measurement of particles obtained via a HIAC counter.

To get an overview of the data being examined, Table 1 and Table 2 provide summary statistics and box plots of each explanatory variable, respectively.

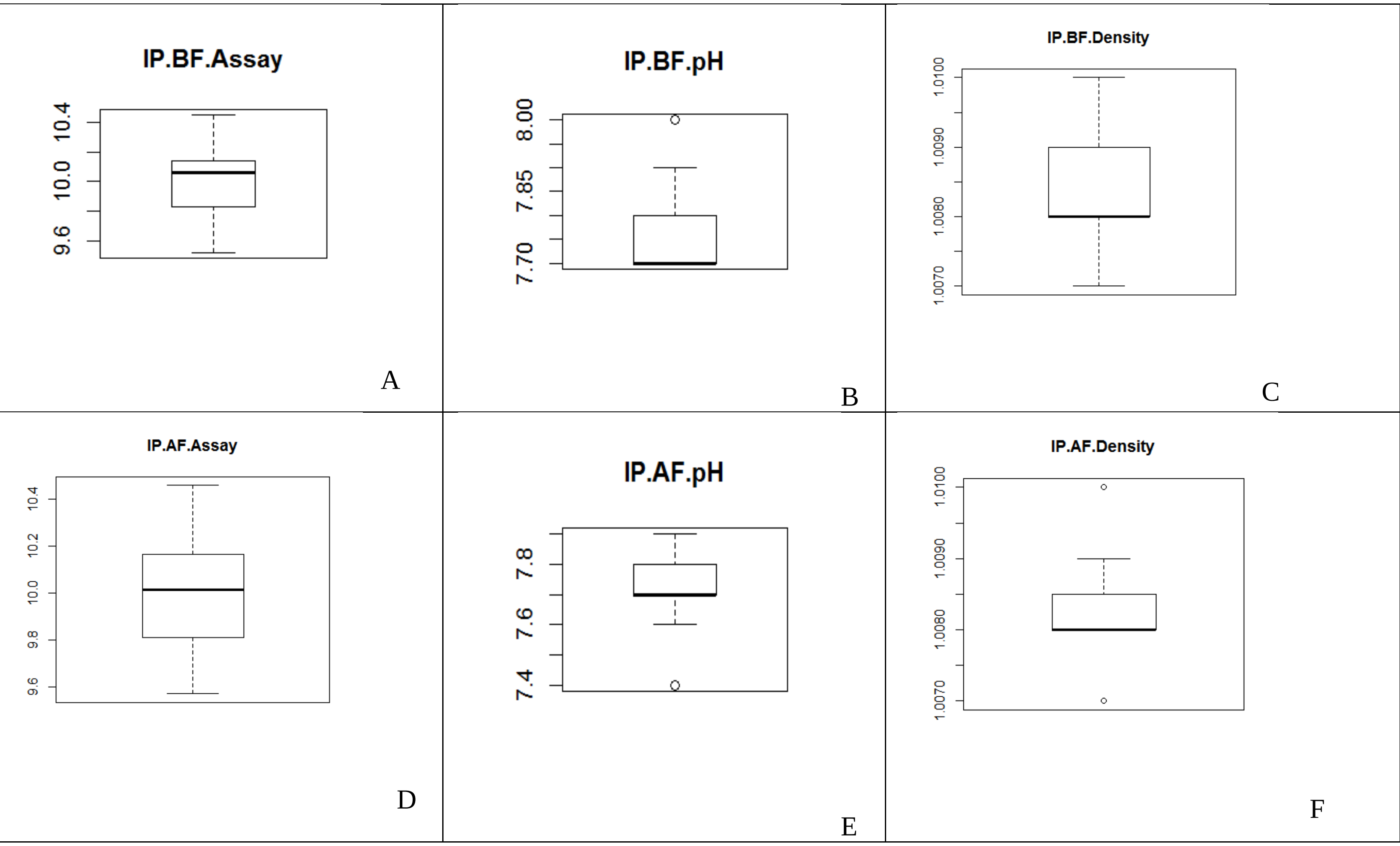
Table 1: Summary Statistics of each of the Explanatory Variables.

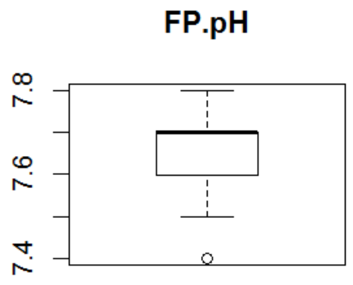
Variable Name	Min	Q_1	Median	Q_3	Max	Mean	Stdev
IP.BF.Assay	9.520	9.830	10.06	10.14	10.45	9.996	0.247
IP.BF.pH	7.700	7.700	7.700	7.800	8.000	7.776	0.0926
IP.BF.Density	1.007	1.008	1.008	1.009	1.010	1.008	0.0007
IP.AF.Assay	9.570	9.815	10.015	10.162	10.460	10.002	0.240
IP.AF.pH	7.400	7.700	7.700	7.800	7.900	7.704	0.102
IP.AF.Density	1.007	1.008	1.008	1.008	1.010	1.008	0.0007
FP.pH	7.400	7.600	7.700	7.700	7.800	7.656	0.123
O2.in.headspace.pcnt	1.300	1.800	2.000	2.500	3.700	2.120	0.542
Assay.pcnt	96.20	97.90	100.00	101.60	105.90	99.990	2.470
Total.Impurities.	0.200	0.300	0.300	0.500	0.700	0.384	0.118

pcnt							
Color	26.00	36.00	40.00	50.00	63.00	43.440	9.824
HIAC.gt.10.um. Particles	230.0	540.0	660.0	930.0	1530.0	716.4	284.501
HIAC.gt.25.um. particles	0.0	0.0	10.0	10.0	90.0	10.800	18.240
Initial.pH.Before .Log1	6.670	6.94	7	7.01	7.2	6.973	0.1009
Amt.NaOH.adde d.Ph.Log1	200.0	300.0	320.0	350.0	400.0	325.80	39.043
pH.Before.Log.2	7.700	7.780	7.800	7.800	7.900	7.787	0.053
Amt.NaOH.adde d.Ph.Log2	0.0	0.0	0.0	20.0	55.0	10.600	17.520

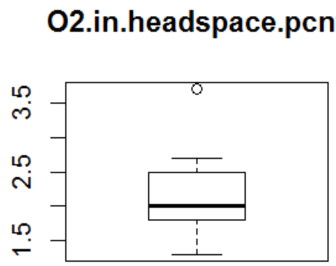
Below are the boxplots for each of the variables. Only a few of the variables have skewed data: IP.BF.pH, HIAC.gt.25.um.particles, and Amt.NaOH.added.Ph.Log2.

Table 2: Boxplots of each of the Explanatory Variables.

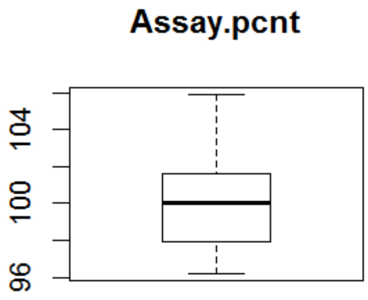



FP.pH
7.4
7.6
7.8

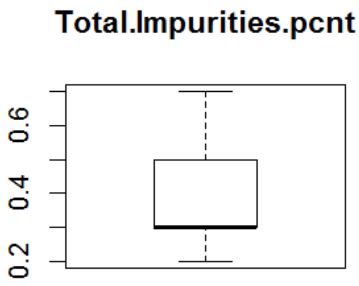
G


O2.in.headspace.pcnt
1.5
2.5
3.5

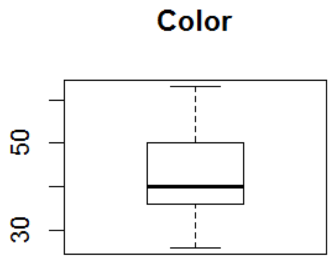
H


Assay.pcnt
96
100
104

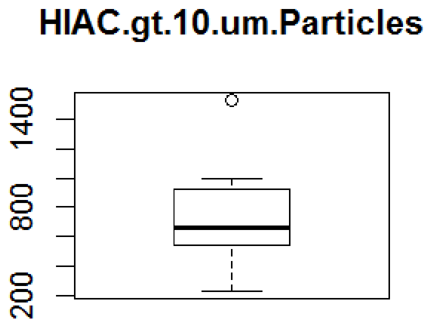
I


Total.Impurities.pcnt
0.2
0.4
0.6

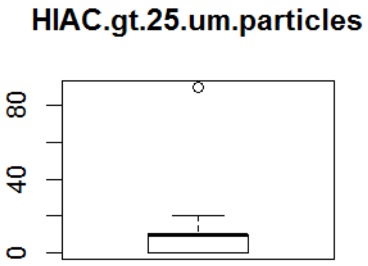
J


Color
30
50

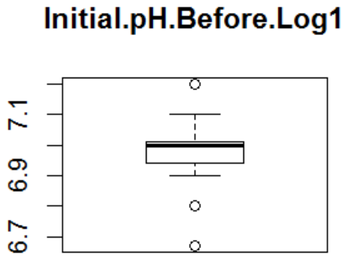
K


HIAC.gt.10.um.Particles
200
800
1400

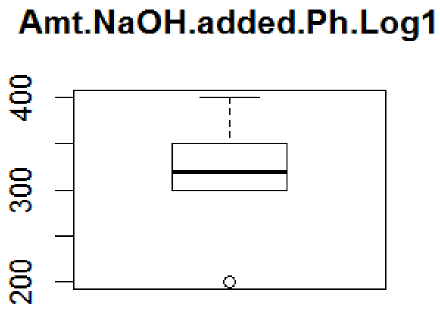
L


HIAC.gt.25.um.particles
0
40
80

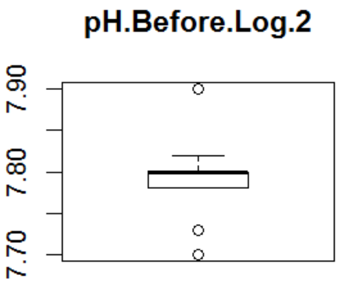
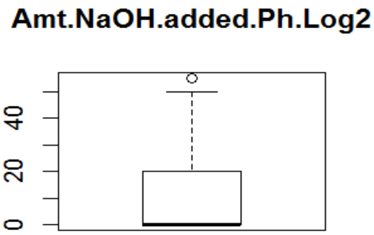
M


Initial.pH.Before.Log1
6.7
6.9
7.1

N


Amt.NaOH.added.Ph.Log1
200
300
400

O

 P	 Q	

With all variables and chemical words defined, the process of modeling can begin.

Statistical Methods

How was the logistic model derived? Why does it work and why is it valid? To derive the logistic regression model, "we assume that the case of interest (or 'true') is coded to 1, and the alternative case (or 'false') is coded to 0" (Zumel, 2011). This model also has the assumption that the log-odds of an observation output y is "expressed as a linear function of the K input variables x:

$$log\frac{P(x)}{1-P(x)} = \sum_{j=0}^{K}\beta_j x_j\text{" (equation 1)}$$

To get a total of K+1 parameters, we set the $x_0 = 1$ to the get the first term of the series $b_{0,}$ a constant term.

The left hand side of the above equation is called the logit of probability P. This is where the "logistic" in logistic regression model comes from. To get rid of the log, we can raise both side of the equation to the base of the log to get the following ratio:

$$\text{Odds Ratio} = \frac{probability\ of\ success}{probability\ of\ failure} = \frac{P(x)}{1-P(x)} = e^{\left(\sum_{j=0}^{K}\beta_j x_j\right)} = \prod_{j=0}^{K} e^{\beta_j x_j}$$

Hence, logistic regression models are multiplicative in their inputs. Thus, using this knowledge, the coefficients can be interpreted better. “The value exp(β_j) tells us how the odds of the response being “true” increase (or decrease) as x_j increases by one unit, all other things being equal. For example, suppose b_j = 0.693. Then exp(β_j) = 2.” The logit equation can be inverted to get a simplified expression for P:

Letting $z = \sum_{j=0}^{K}\beta_j X_j$ we get

$$\frac{P(x)}{1-P(x)} = e^{z}$$

Multiply both sides by $(1 - P(x))$ to get

$$P(x) = e^{z}(1 - P(x))$$

Then distribution exp(z)

$$P(x) = e^{z} - e^{z}P(x)$$

Then, add e^{z}P(x) to the left side,

$$P(x) + e^{z}P(x) = e^{z}$$

Then factor out a P(x)

$$P(x)(1 + e^{z}) = e^{z}$$

Divide by (1+e^{z}) to get

$$P(x) = \frac{e^{z}}{1 + e^{z}}$$

This ends the derivation of the logistic regression model. ■

If the base of the log, from the logit of P, is the number e = 2.718281828459, then the logit equation can be written as

$$P(x) = \frac{e^{z}}{1 + e^{z}} = \frac{e^{\sum_{j=0}^{K} \beta_j x_j}}{1 + e^{\sum_{j=0}^{K} \beta_j x_j}}$$

In particular, we will be using this model, with z = $a + Bx$, to examine the complaints of precipitation. Here, complaints received from the consumer base is a binary outcome: either Yes (Y) or No (N) are the possibilities:

$$P(Complaint) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

For each of the in-process and finished quality parameters, a single-parameter logistic regression model was completed so that each parameter can be assessed of statistical significance with respect to the likelihood of a consumer complaint.

Additionally, possible interactions, after discussing them with the chemical and process engineers, were investigated by looking at multivariable models. The following logistic regression model was used to test for two-way interactions between several variables:

$$P(Complaint) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + B_3 X_1 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + B_3 X_1 X_2}}$$

The two-way interactions that were investigated used the in-process after-filtration results for assay, pH and density, and several other interaction terms using the finished product results for assay, density, 10 µm HIAC particles, and 25 µm HIAC particles, and the before process results for assay.

The analyses were completed in R console 64 bit, version 3.4.1. The code that was used can be found in the appendix.

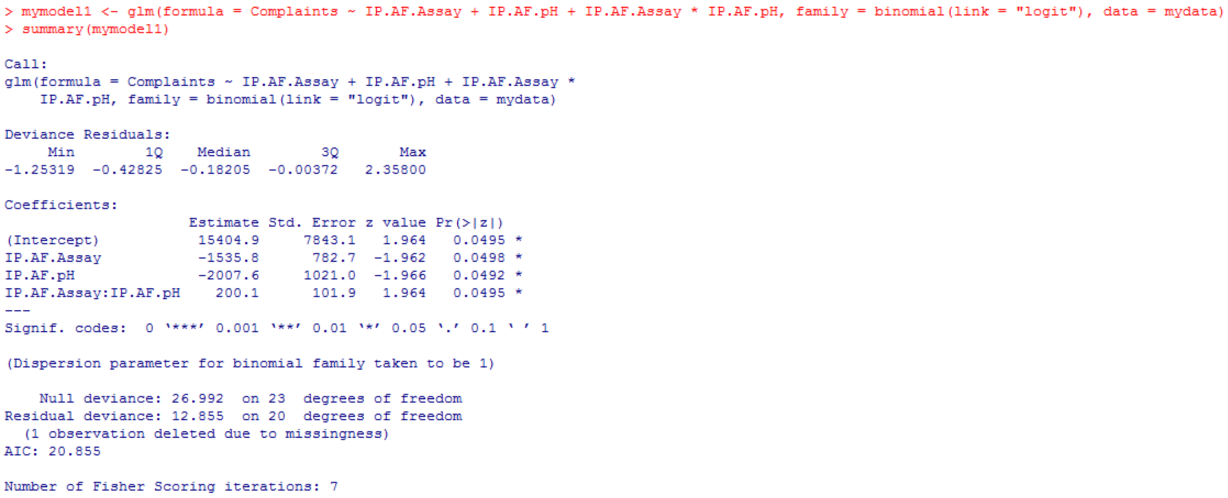
In Table 3 below, from each of the 25 given lots, the resulting statistics are listed for each of the 17 single-parameter logistic model. (Here, only the coefficients were estimated in each of the models).

Table 3. Single parameter logistic regression summary table

Description	**Estimate**	**z-value**	**p-value**	**Standard Error**
IP AF pH	4.357	0.814	0.416	5.353
IP BF pH	-1.544	-0.289	0.773	5.350
FP pH	0.992	0.249	0.803	3.986
IP BF Assay (mg/mL)	2.790	1.282	0.200	2.176
IP AF Assay (mg/mL)	4.240	1.695	**0.090**	2.501
IP AF Density (g/mL)	976.100	1.385	0.166	704.9
HIAC >25 um particles	0.033	1.172	0.241	0.028
O2 in headspace (%)	-0.341	-0.369	0.712	0.924
Total Impurities (%)	-3.864	-0.811	0.417	4.763
HIAC >10 um Particles	-0.003	-1.261	0.207	0.002
Amt NaOH added Ph Log 1 (mL)	0.0023	0.186	0.853	0.013
Assay (% label)	-0.001	-0.005	0.996	0.194
Color (APHA units)	-0.039	-0.080	0.936	0.049
Amt NaOH added Ph Log 2 (mL)	-1.106	-0.005	0.996	217.678
Initial pH (Before Log 1)	0.01794	0.004	0.997	4.737
IP BF Density (g/mL)	547.000	0.822	0.411	665.100
pH (Before Log 2)	6.700	0.711	0.477	9.424

Notice that there are no p-values below the commonplace 0.05 significance level, for each of the parameters. However, one of these parameters, IP AF Assay, has a p-value below 0.1 and thus is marginally significant. HIAC > 25, HIAC > 10, IP AF Assay, IP AF Density, and IP BF Assay all have p-values less than 0.25. Thus, a second model was run to test to see if there is evidence of interactions between HIAC > 25 and HIAC > 10, IP AF Assay and IP AF Density, IP AF Assay and IP BF Assay, HIAC > 25 and IP AF Assay, and HIAC > 10 and IP AF Assay. Additionally, the process and chemical engineers suspect that the After Process Assay and After Process pH have an interaction, so we also test this interaction as well.

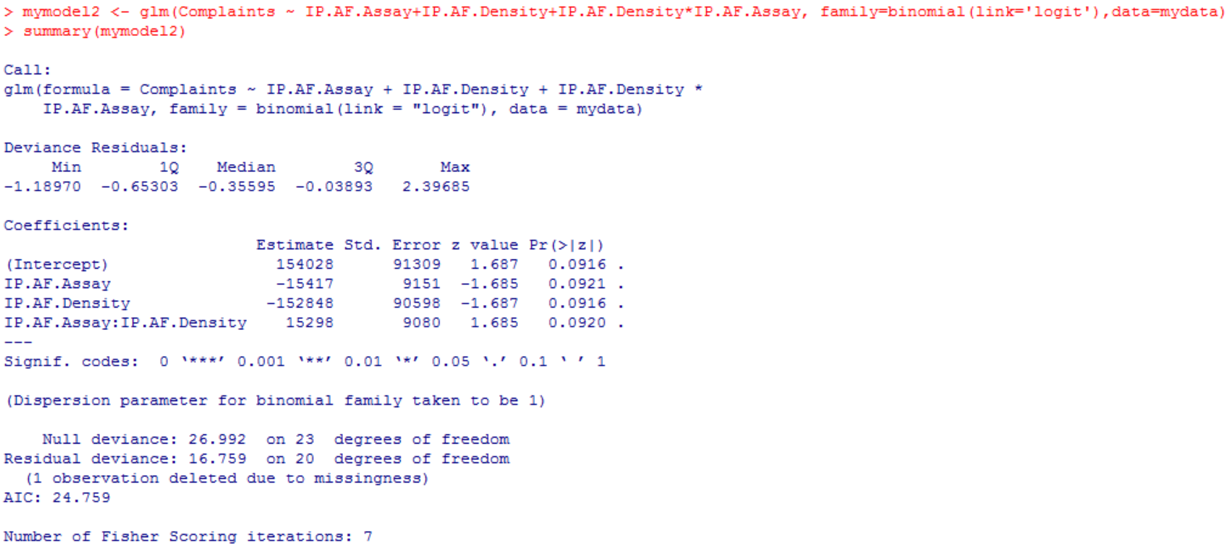
The interaction term between after-process assay and after-process pH came out to be statistically significant, at a 5% significance level, with a p-value for the interaction term equaling 0.0495. See Figure 3 for these results.



```
> mymodel1 <- glm(formula = Complaints ~ IP.AF.Assay + IP.AF.pH + IP.AF.Assay * IP.AF.pH, family = binomial(link = "logit"), data = mydata)
> summary(mymodel1)

Call:
glm(formula = Complaints ~ IP.AF.Assay + IP.AF.pH + IP.AF.Assay *
    IP.AF.pH, family = binomial(link = "logit"), data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.25319  -0.42825  -0.18205  -0.00372   2.35800

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           15404.9     7843.1   1.964   0.0495 *
IP.AF.Assay           -1535.8      782.7  -1.962   0.0498 *
IP.AF.pH              -2007.6     1021.0  -1.966   0.0492 *
IP.AF.Assay:IP.AF.pH    200.1      101.9   1.964   0.0495 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.992  on 23  degrees of freedom
Residual deviance: 12.855  on 20  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 20.855

Number of Fisher Scoring iterations: 7
```

Figure 3: Test of Interaction Between After-Process Assay and After-Process pH

Interaction between In Process After Filtration Assay and In Process After Filtration Density came out to be marginally significant with a p-value for the interaction term equaling 0.0920. See Figure 4 for the results.



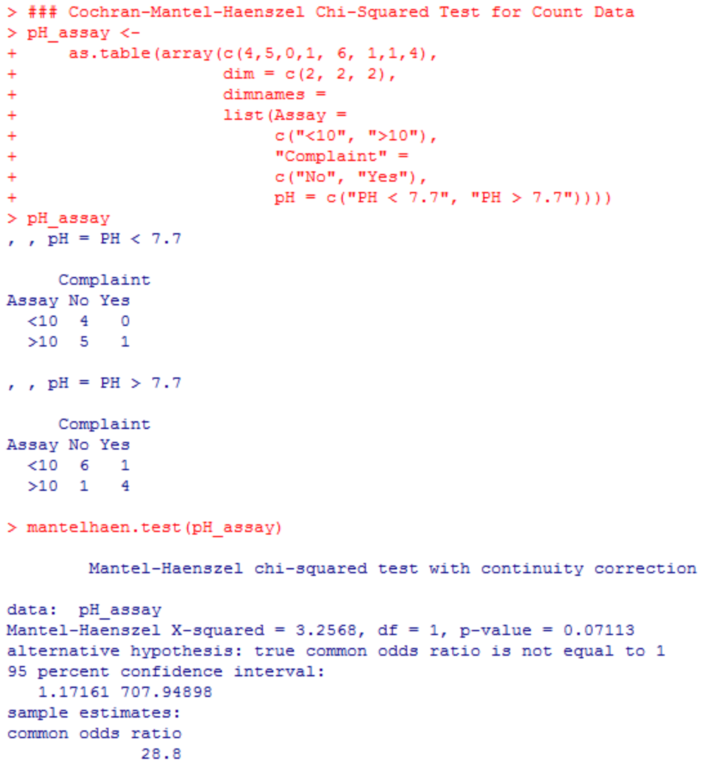
```
> mymodel2 <- glm(Complaints ~ IP.AF.Assay+IP.AF.Density+IP.AF.Density*IP.AF.Assay, family=binomial(link='logit'),data=mydata)
> summary(mymodel2)

Call:
glm(formula = Complaints ~ IP.AF.Assay + IP.AF.Density + IP.AF.Density * 
    IP.AF.Assay, family = binomial(link = "logit"), data = mydata)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.18970  -0.65303  -0.35595  -0.03893   2.39685  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 154028      91309   1.687   0.0916 .
IP.AF.Assay                 -15417       9151  -1.685   0.0921 .
IP.AF.Density              -152848      90598  -1.687   0.0916 .
IP.AF.Assay:IP.AF.Density    15298       9080   1.685   0.0920 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.992  on 23  degrees of freedom
Residual deviance: 16.759  on 20  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 24.759

Number of Fisher Scoring iterations: 7
```

Figure 4: Test of Interaction Between After-Process Assay and After-Process Density

Table 4 and Figure 5 were then created to verify the result of Interaction between After-process Assay and After-process pH. Because the respective medians for pH and Assay are approximately 7.7 and 10 mg/ml (see Tables 1 and 2), this contingency table classifies the mean pH as either above or below 7.7, and mean assay as above or below 10 mg/ml. (It should be noted that three of the values were excluded because they had a mean pH equal to 7.7. They wouldn't have changed the model because they had the value of the median mean pH in the dataset.) The three-way contingency table gives us a non-parametric test for the pH-assay interaction.

Table 4. Three-way classification table for pH-assay association

	pH < 7.7				**pH > 7.7**		
	Complaint				**Complaint**		
Assay	N	Y	Total		N	Y	Total
< 10.0 mg/mL	4	0	4		6	1	7
> 10.0 mg/mL	5	1	6		1	4	5
Total	9	1	10		7	5	12
Cochran-Mantel-Haenszel Chi-Square:						3.2568	
p-value, Cochran-Mantel-Haenszel Chi-Square:						0.07113	



```
> ### Cochran-Mantel-Haenszel Chi-Squared Test for Count Data
> pH_assay <-
+     as.table(array(c(4,5,0,1, 6, 1,1,4),
+                    dim = c(2, 2, 2),
+                    dimnames =
+                    list(Assay =
+                         c("<10", ">10"),
+                         "Complaint" =
+                         c("No", "Yes"),
+                         pH = c("PH < 7.7", "PH > 7.7"))))
> pH_assay
, , pH = PH < 7.7

     Complaint
Assay No Yes
  <10  4   0
  >10  5   1

, , pH = PH > 7.7

     Complaint
Assay No Yes
  <10  6   1
  >10  1   4

> mantelhaen.test(pH_assay)

        Mantel-Haenszel chi-squared test with continuity correction

data:  pH_assay
Mantel-Haenszel X-squared = 3.2568, df = 1, p-value = 0.07113
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
   1.17161 707.94898
sample estimates:
common odds ratio
             28.8
```

Figure 5: Results from R for Test of Association of pH-assay

A Cochran-Mantel-Haenszel (CMH) statistic is a test statistic that tests the null hypothesis of conditional independence in 2 × 2 × k contingency tables. This was proposed by Mantel and Haenszel in 1959 and is an alternative method to "the test of conditional independence as a logistic model analysis for a 2 × 2 × k contingency table" (Agresti 2013b). The CMH test relates to, but is not based on the same logistic model as this test.

Let's derive and then explain this statistic. According to Agresti (2013b), in the contingency table, Mantel and Haentzel

> treated the response (column) marginal totals as fixed. Thus, in each partial table k of cell counts {n_{ijk}}, their analysis conditioned on both the treatment (e.g. group) totals {n_{1+k}, n_{2+k}} and the response outcome totals {n_{+1k}, n_{+2k}}. The usual sampling schemes then yield a hypergeometric distribution for the first cell count n_{11k} in each partial table. That count determines {n_{12k}, $n_{21k,}$ n_{22k}}, given the marginal totals.
>
> Under H_0, the hypergeometric mean and variance of n_{11k} are
>
> $$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}} \; and \; var(n_{11k}) = \frac{n_{1+k}n_{+1k}n_{2+k}n_{+2k}}{[n_{++k}^2(n_{++k}-1)]}$$

Cell counts from different partial tables are independent. The test statistic combines information from the K tables by comparing $\sum_k n_{11k}$ to its null expected value. It equals

$$CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k var(n_{11k})}$$

This statistic has a large sample chi-squared null distribution with df = 1. Notice that the CMH statistic has the form of and is an example of a chi-square statistic, a statistic that is a measure of the overall closeness of the observed frequencies to the expected frequencies.

If the chi-square statistic is large due to the expected and observed values being far apart, the null hypothesis of independence is rejected ("The Chi-Square Statistic," 2003). The chi-square curve is used to judge whether the calculated statistic is large enough. It should be stated, however, that “a nonsignificant CMH statistic suggests either that there is no association or that no pattern of association has enough strength or consistency to dominate any other pattern” (Cochran-Mantel-Haenszel Statistics, 2009). The CMH statistic for the contingency table above has a p-value of 0.07113, making it marginally significant. Thus, this verifies that there is likely to be some association between the pH and assay.

Conclusion

Even though the above model is meant to model the probability that a complaint will be filed from a customer that is associated with the precipitation in the drugs, it might be beneficial to interpret the filing of at least one complaint as a combination of the variables that likely affect the presence of precipitation in the batch of drugs. The association that was found between pH and assay (a.k.a., the concentration of a chemical in a drug), and the interactions found between after process pH and after process assay and after process assay and after process density, will likely provide some insight into potential physical or chemical mechanisms that, under particular conditions of the process, may result in the precipitation of the drug. Thus, the identification of these physical and chemical mechanisms is worth the investigation of the engineers and management of the manufacturing facilities. Another potential investigation is to determine what the complaints are about.

References

Agresti, A. (2013a). Logistic Regression. In D. Balding & N. Cressie & G. Fitzmaurice (Eds.), *Categorical Data Analysis* (pp. 163-206), (Vol 1, 3rd ed.). Hoboken, NJ; John Wiley & Sons, Inc.

Agresti, A. (2013b). Building, Checking, and Applying Logistic Regression Models. In D. Balding & N. Cressie & G. Fitzmaurice (Eds.), *Categorical Data Analysis* (pp. 207-250), (Vol 1, 3rd ed.). Hoboken, NJ; John Wiley & Sons, Inc.

Assay. 2017. In Merriam-Webster.com

Retrieved September 17, 2017, from https://www.merriam-webster.com.

C, K. (2014, November 26). APHA Color: A Measurement of Liquid Purity. Retrieved from https://www.hunterlab.com/blog/color-chemical-industry/apha-color-measurement-liquid-purity/

Cochran-Mantel-Haenszel Statistics. (2009). Retrieved from http://www.okstate.edu/sas/v7/sashtml/books/stat/chap26/sect27.htm

Density. 2017. In Merriam-Webster.com

Retrieved September 17, 2017, from https://www.merriam-webster.com.

Hahn, G. J., Doganaksoy, N., & Blumberg, C. J. (2011). A Career in Statistics: Beyond the Numbers [electronic resource]. Hoboken, N.J. : Wiley, c2011.

Headspace. 2017. In Merriam-Webster.com

Retrieved September 17, 2017, from https://www.merriam-webster.com.

Impurity. 2017. In Pharma-IQ.com

Retrieved September 17, 2017, from https://www.pharma-iq.com/

Kleinbaum, D. & Klein, M. (2010). Introduction to Logistic Regression. In M. Gail & K. Krickeberg & J.M. Samet & A. Tsiatis & W. Wong (Eds.), *Logistic Regression: A Self Learning Text.* (pp. 1-40). Retrieved https://mregresion.files.wordpress.com/2011/04/logistic-regression-a-self-learning-text.pdf

pH. 2017. In Merriam-Webster.com

Retrieved Sept 17, 2017, from https://www.merriam-webster.com.

The Chi-Square Test Statistic. (2003, September 8). Retrieved from http://www.stat.wmich.edu/s216/book/node114.html

Zumel, N. (2011, September 14). The Simpler Derivation of Logistic Regression. Retrieved from http://www.statsblogs.com/2011/09/14/the-simpler-derivation-of-logistic-regression/

Acknowledgments

Thank you Jeff Gardner from DataPharm for letting me use the data set to do this project and for giving me guidance in the research project. Thank you Dr. Fan for also giving me research guidance and support, and helping me to learn the R code for this project.

Appendix

The R code from this project can be found here.

```
## To import and read the data and data names

mydata=read.csv("spdata01.csv")

names(mydata)

## summary statistics

attach(mydata)

summary(Initial.pH.Before.Log1); sd(Initial.pH.Before.Log1)

boxplot(Initial.pH.Before.Log1); title("Initial.pH.Before.Log1")

summary(IP.BF.Density); sd(IP.BF.Density,na.rm=T)

boxplot(IP.BF.Density, na.rm=T); title("IP.BF.Density")

##Logit Model and Graph

mymodel <- glm(Complaints ~ IP.AF.Assay,family=binomial(link='logit'),data=mydata)

summary(mymodel)

       ### Logit Model Graph

beta0=mymodel$coefficients[1]; beta1=mymodel$coefficients[2]

Py=exp(beta0+beta1*IP.AF.pH)/(1+exp(beta0+beta1*IP.AF.pH))

plot(IP.AF.pH, Py, type="p",col="red")

## interaction

mymodel1 <- glm(Complaints ~ IP.AF.Assay+IP.AF.pH+IP.AF.Assay*IP.AF.pH,
family=binomial(link='logit'),data=mydata)

summary(mymodel1)

### Cochran-Mantel-Haenszel Chi-Squared Test for Count Data

pH_assay <-

    as.table(array(c(4,5,0,1, 6, 1,1,4),

                   dim = c(2, 2, 2),

                   dimnames =
                   list(Assay =

                        c("<10", ">10"),

                        "Complaint" =

                        c("No", "Yes"),

                        pH = c("PH < 7.7", "PH > 7.7"))))

pH_assay

mantelhaen.test(pH_assay)
```