Montana State University

Statistical Consulting and Research Services

Coliform Contamination in Private Well Water on the Crow Reservation

Lead Statistician:
Michaela Powell

Director:
Megan Higgs

Contributions from:
Laurie Rugemer

This material is provided to communicate advice from SCRS statisticians based on our best understanding of your research needs. We encourage you to use this report in discussions with colleagues. Please do not publish any portion of this material without permission.

When you make use of our work for publications or presentations, please be sure to acknowledge the funding we receive from NIGMS using the following: "Research reported in this publication was supported by Institutional Development Awards (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Awards P20GM103474, 5U54GM104944, U54GM115371, and 5P20GM104417. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health."

prepared for Emery Three Irons

October 20, 2018

1 Introduction

Emery Three Irons is a master's level graduate student in the Department of Land Resources and Environmental Sciences (LRES) at Montana State University (MSU). Advised by Scott Powell, Ph.D., who is an Associate Professor in the Department of LRES at MSU, Emery is working on his master's thesis which explores coliform contamination in private well water on the Crow Reservation. He has requested the assistance of the Statistical Consulting and Research Services in deciding on an appropriate analysis for the data he collected.

2 Background

Emery collected data from 100 private wells on the Crow Reservation. From each well, a water sample was taken to be tested for coliform contamination using the IDEXX Quanti-Tray 2000. After adding a reagent to 100mL of the water sample, the 48 large wells and 49 small wells on the tray are filled. The tray is placed in an incubator for 24 hours, after which the number of small wells and the number of large wells that have turned yellow are both recorded. Using the IDEXX Quanti-Tray 2000 Most Probable Number (MPN) Conversion Table, these two counts are used to estimate the coliform MPN per 100mL of well water. These MPN values range from 1 to 2419.6, where no yellow wells is given an MPN of "<1" and recorded as a "non-detect.". Using this method, two subsamples from the first 34 water samples were analyzed and three subsamples from the remaining 76 were analyzed.

For each well, Emery also recorded well stewardship characteristics (well cap type, well cap condition, well age, well depth, production formation) and physical characteristics about the location of the well (aquifer type, land cover type, distance to the nearest river). He has identified the following research questions of interest about the relationship between coliform contamination and these characteristics:

1. After accounting for well depth, what is the relationship between coliform contamination, the well being in alluvium, and the distance to the nearest river?
2. What is the relationship between coliform contamination and the condition and style of the well's cap?
3. What is the relationship between coliform contamination and the presence of agriculture cover in a 60 meter radius?
4. What is the relationship between coliform contamination and the presence of agriculture cover in a 90 meter radius?
5. What is the relationship between coliform contamination and the age of the well?

"Coliform contamination" can be described by a binary contaminated or not variable, or severity of contamination can be described on a continuous scale using the MPNs. The focus of this report is on the binary measure of contaminated or not which requires an explicit definition of what is classified as “contaminated.”

Using the MPNs for each well, a binary coliform contamination variable can be constructed; however, we first must define a criterion for whether or not the well is classified as contaminated. For the purpose of this report, a contaminated well is any well for which at least one of the subsamples had an MPN of 1 or greater; or equivalently, a non-contaminated well is any well for which all of the tests were non-detects (an MPN of <1).

After SCRS provided two sets of graphics exploring the five research questions of interest, it was agreed that it would be useful for Emery to start by implementing logistic regression for each of these questions using a binary coliform contamination variable as the response (defined as in this report or using a different contamination criterion). Emery requested an example logistic regression analysis for one of the research questions so that he can reference it when implementing logistic regression for the remaining questions.

3 Example Logistic Regression Analysis

The first research question will be used in this example.

Before beginning the analysis, the binary coliform contamination variable must be constructed. As mentioned previously, for this report, a contaminated well is any well for which at least one of the subsamples had an MPN of 1 or greater. When reading the following `R` code (R Core Team 2017), please note that the dataset has zeros entered in the place of the <1s.

```r
well <- read.csv("crowhomewelldata4.csv")
keep <- c("MPN_100mL_1_Coliform", "MPN_100mL_2_Coliform",
          "MPN_100mL_3_Coliform", "Well_Depth_ft",
          "Alluvium_or_Not", "Near_Distance_Stream")
well <- well[,keep]

colnames(well)[colnames(well)=="Near_Distance_Stream"] <- "Stream_Distance"

well$sum_MPN <- (well$MPN_100mL_1_Coliform + well$MPN_100mL_2_Coliform
                 + ifelse(is.na(well$MPN_100mL_3_Coliform)==TRUE, 0, well$MPN_100mL_3_Coliform))

well$contaminated <- ifelse(well$sum_MPN==0, 0, 1)
```

With the response variable constructed, we can now fit a model addressing the research question of interest. Rephrasing this question to relate to logistic regression, we have two specific questions:

1. *How are the odds of coliform contamination related to well depth, after accounting for distance to the nearest river and whether the well is in alluvium?*
2. *How are the odds of coliform contamination related to distance to the nearest river, after accounting for well depth and whether the well is in alluvium?*

In the plots exploring this question (included in previously mentioned exploratory graphics), the relationship between the distance to the river and odds of finding a contaminated well appears to depend on whether the well is in alluvium. In other words, there appears to be an interaction between alluvium and distance to the nearest river on the relationship with coliform contamination. We fit the following logistic model:

$$y_i \sim Binomial(1, \pi_i)$$

$$logit(\pi_i) = \beta_0 + \beta_1 WD_i + \beta_2 DR_i + \beta_3 I_{alluvium_i} + \beta_4 DR_i * I_{alluvium_i}$$

Where:

- y_i is the binary response for whether the i^{th} well is classified as contaminated.
- π_i is the probability that the i^{th} well is classified as contaminated.
- WD_i is the depth of the i^{th} well in feet.
- DR_i is the distance to the nearest river of the i^{th} well in feet.
- $I_{alluvium_i}$ is an indicator function of whether the i^{th} well is in alluvium ($I_{alluvium_i} = 1$ if in alluvium and $I_{alluvium_i} = 0$ if not in alluvium).

Please review Section 20.2.1 in The Statistical Sleuth for assessing assumptions of binary logistic regression (Ramsey and Schafer 2012, 606–7).

The following `R` code fits the model defined above and outputs the summary (R Core Team 2017).

```
well$Alluvium <- ifelse(well$Alluvium_or_Not=="Alluvium", "Yes", "No")

m1 <- glm(data=well, contaminated ~ Well_Depth_ft + Stream_Distance * Alluvium,
          family=binomial(link="logit"))

summary(m1)
```

```
##
## Call:
## glm(formula = contaminated ~ Well_Depth_ft + Stream_Distance *
##     Alluvium, family = binomial(link = "logit"), data = well)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6432  -1.3219   0.8047   0.9855   1.7724
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.506129   0.715721   0.707   0.4795
## Well_Depth_ft             0.004679   0.006233   0.751   0.4529
## Stream_Distance          -0.003352   0.001568  -2.138   0.0325
## AlluviumYes              -0.302539   0.784055  -0.386   0.6996
## Stream_Distance:AlluviumYes  0.004138   0.001919   2.156   0.0311
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.02  on 82  degrees of freedom
## Residual deviance: 102.91  on 78  degrees of freedom
##   (17 observations deleted due to missingness)
## AIC: 112.91
##
## Number of Fisher Scoring iterations: 5
```

Due to the inclusion of the interaction term between alluvium and distance to the nearest river, interpretation takes some care.

Consider that, for a well not in alluvium ($I_{alluvium_i} = 0$) the model simplifies to:

$$logit(\pi_i) = \beta_0 + \beta_1 WD_i + \beta_2 DR_i$$

And, for a well in alluvium ($I_{alluvium_i} = 1$) the model simplifies to:

$$logit(\pi_i) = (\beta_0 + \beta_3) + \beta_1 WD_i + (\beta_2 + \beta_4) DR_i$$

Using these two simplified models, we can exponentiate the coefficient estimates attached to WD_i to estimate how the odds of contamination multiplicatively change for a one foot increase in well depth, and we can exponentiate the coefficient estimates attached to DR_i to estimate how the odds of contamination multiplicatively change for a one foot increase in distance to the nearest river. However, first we will calculate the confidence intervals corresponding to each of the interpretations.

For wells not in alluvium, we can simply exponentiate the endpoints of the confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_2$, because `Not Alluvium` is the baseline group for the model fit. The following `R` code outputs the exponentiated confidence intervals (R Core Team 2017).

```r
exp(confint(m1))
```

```
##                                  2.5 %    97.5 %
## (Intercept)                  0.4287479 7.9316708
## Well_Depth_ft                0.9918587 1.0174838
## Stream_Distance              0.9930186 0.9992057
## AlluviumYes                  0.1477913 3.3659525
## Stream_Distance:AlluviumYes 1.0007366 1.0083928
```

However, to calculate the confidence intervals for wells in alluvium, in addition to the confidence interval for $\hat{\beta}_1$, you may want the confidence interval for $(\hat{\beta}_2 + \hat{\beta}_4)$. The easiest way to attain this is to re-level the baseline group of the alluvium indicator variable to `Alluvium` and refit the model such that the parameter estimate for distance to the nearest river is for wells in alluvium. The following `R` code re-levels the alluvium indicator function, refits the model, and outputs the exponentiated confidence intervals (R Core Team 2017).

```r
well$Not_Alluvium <- ifelse(well$Alluvium_or_Not=="Not Alluvium", "Yes", "No")

m2 <- glm(data=well, contaminated ~ Well_Depth_ft + Stream_Distance * Not_Alluvium,
          family=binomial(link="logit"))
```

```r
exp(confint(m2))
```

```
##                                      2.5 %    97.5 %
## (Intercept)                      0.4376330 3.5245139
## Well_Depth_ft                    0.9918587 1.0174838
## Stream_Distance                  0.9986606 1.0032299
## Not_AlluviumYes                  0.2970927 6.7662965
## Stream_Distance:Not_AlluviumYes 0.9916771 0.9992639
```

With confidence intervals calculated, we can exponentiate the parameter estimates and make interpretations. Recall that the exponentiated coefficients correspond to an estimated multiplicative change in the odds.

- For both wells in alluvium and not in alluvium, each one foot increase in well depth is associated with an estimated 1.0047-fold change in the odds of coliform contamination, after adjusting for the distance to the nearest river (95% CI: 0.9919-fold to 1.0175-fold change).
- For wells not in alluvium, each one foot increase in the distance to the nearest river is associated with an estimated 0.9967-fold change in the odds of coliform contamination, after adjusting for well depth (95% CI: 0.9930-fold to 0.9992-fold change).
- For wells in alluvium, each one foot increase in the distance to the nearest river is associated with an estimated 1.0008-fold change in the odds of coliform contamination, after adjusting for well depth (95% CI: 0.9986-fold to 1.0032-fold change).

Interpretation	**Parameter Estimate**	**Exponentiated Parameter Estimate**	**Exponentiated Confidence Interval**
Well Depth	0.0047	1.0047	(0.9919, 1.0175)
Distance to the Nearest River (not in alluvium)	-0.0033	0.9967	(0.9930, 0.9992)
Distance to the Nearest River (in alluvium)	-0.0033+0.0041=0.0008	1.0008	(0.9986, 1.0032)

Table 1: Parameter estimates, exponentiated parameter estimates, and exponentiated confidence intervals used for each interpretation.

These “multiplicative change” interpretations can also be converted to “percent change”. If the exponentiated estimate is below one, we subtract it from one and then multiply by one hundred for the estimated percent decrease in the odds; if the exponentiated estimate is above one, we subtract one from it and multiply by one hundred for the estimated percent increase in the odds.

- For both wells in alluvium and not in alluvium, the odds of coliform contamination increase by an estimated 0.47% for each one foot increase in well depth, after adjusting for the distance to the nearest river (95% CI: 0.81% decrease to 1.75% increase).
- For wells not in alluvium, the odds of coliform contamination decrease by an estimated 0.33% for each one foot increase in in the distance to the nearest river, after

adjusting for well depth (95% CI: 0.70% decrease to 0.08% decrease).

- For wells in alluvium, the odds of coliform contamination increase by an estimated 0.08% for each one foot increase in the distance to the nearest river, after adjusting for well depth (95% CI: 0.14% decrease to 0.32% increase).

The interpretations would be more meaningful if instead of being for each one foot change (in well depth or in distance to the nearest river) they were for each ten foot change. To re-scale the interpretations, the coefficient estimates and bounds of the confidence intervals need to be multiplied by ten before exponentiating. The following R code does this re-scaling.

```
## FOR THE MODEL WITH "NOT ALLUVIUM" AS THE BASELINE GROUP

as.data.frame(exp(10*summary(m1)$coef[,1]))
```

```
##                              exp(10 * summary(m1)$coef[, 1])
## (Intercept)                                    157.79412012
## Well_Depth_ft                                    1.04789861
## Stream_Distance                                  0.96703822
## AlluviumYes                                      0.04853885
## Stream_Distance:AlluviumYes                      1.04224487
```

```
## FOR THE MODEL WITH "NOT ALLUVIUM" AS THE BASELINE GROUP

exp(10*confint(m1))
```

```
##                                     2.5 %       97.5 %
## (Intercept)                 2.099035e-04 9.854777e+08
## Well_Depth_ft               9.215059e-01 1.189255e+00
## Stream_Distance             9.323391e-01 9.920858e-01
## AlluviumYes                 4.971527e-09 1.866723e+05
## Stream_Distance:AlluviumYes 1.007391e+00 1.087170e+00
```

```
## FOR THE MODEL WITH "ALLUVIUM" AS THE BASELINE GROUP

as.data.frame(exp(10*summary(m2)$coef[,1]))
```

```
##                                  exp(10 * summary(m2)$coef[, 1])
## (Intercept)                                            7.6591448
## Well_Depth_ft                                          1.0478986
## Stream_Distance                                        1.0078906
## Not_AlluviumYes                                       20.6020546
## Stream_Distance:Not_AlluviumYes                        0.9594674
```

```
## FOR THE MODEL WITH "ALLUVIUM" AS THE BASELINE GROUP

exp(10*confint(m2))
```

```
##                                         2.5 %       97.5 %
## (Intercept)                     2.576916e-04 2.957959e+05
## Well_Depth_ft                   9.215059e-01 1.189255e+00
## Stream_Distance                 9.866864e-01 1.032772e+00
## Not_AlluviumYes                 5.356982e-06 2.011455e+08
## Stream_Distance:Not_AlluviumYes 9.198197e-01 9.926634e-01
```

4 Choosing the Contamination Criterion

In this report, a well was classified as contaminated if at least one of the tests had an MPN of 1 or greater; however, a different criterion could be chosen. To help with this decision, the MPNs of the wells with mixed test results (both non-detects and MPNs of 1 or greater) are shown in Figure 1. The wells with non-detects and small MPNs could perhaps motivate a different contamination criterion. For example, a well could be classified as contaminated if at least one of the tests had an MPN of 5 or greater. When considering the following figure, please note, again, that the dataset has zeros entered in the place of the non-detects, so there are wells that appear to have an MPN of zero when in actuality they are non-detects.

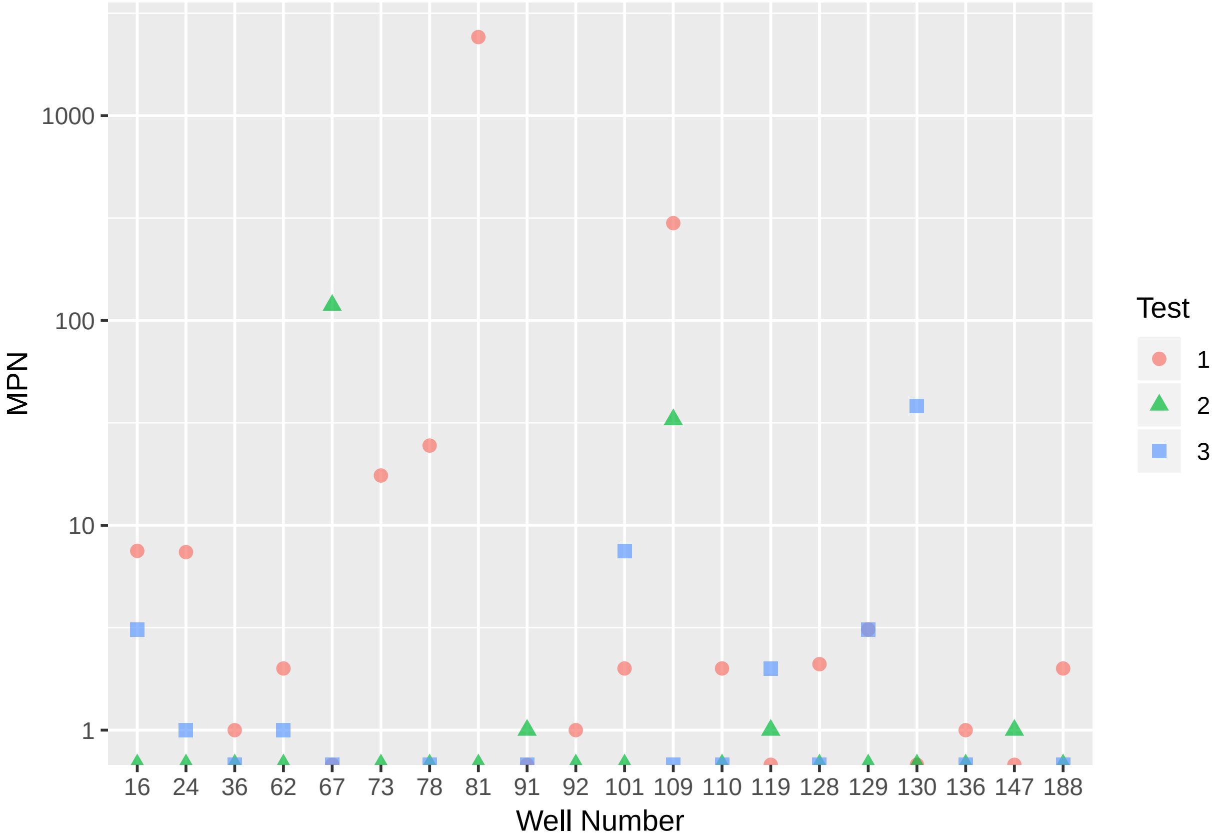

Figure 1: Plot including MPNs of wells with mixed test results.

Works Cited

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ramsey, F., and D. Schafer. 2012. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Cengage Learning. https://books.google.com/books?id=eSlLjA9TwkUC.