The effect of nitrogen application time on infection and development of the Wheat Streak Mosaic Virus (WSMV)

Leslie Gains-Germain, Kenny Flagg

Spring 2015

Contents

1 Introduction **2**

2 Statistical Experience and Assistance Needed **2**

3 Objectives and Questions **2**

4 Study Design/Data collection **2**

5 Recommendations **4**
5.1 Binomial Generalized Linear Model 4
5.2 Graphical Data Exploration 4
5.3 Model Fitting and Possible Issues 6
5.4 Model Refinement 7
5.5 Interpretation 9

6 Scope of inference **12**

7 Additional Comments **12**

8 Appendix: R Code **12**
8.1 Data Preprocessing 12
8.2 `qplot` 13
8.3 Heatmap 13
8.4 Summary Effects Plot 14

9 References **14**

1 Introduction

2 Statistical Experience and Assistance Needed

Nar B. Ranabhat is a PhD candidate in the Department of Land Resources and Environmental Science. He is currently working on several projects regarding the spread and development of the Wheat Streak Mosaic Virus (WSMV) in Winter Wheat. Nar has taken Statistics 511 and 512 here at Montana State University, and he is currently sitting in on Mixed Models. Nar has asked for our help in building a model, fitting it in R, and interpreting the results.

3 Objectives and Questions

Nar is interested in the effect of variety, nitrogen application time, and inoculation on the probability of virus infection in Winter Wheat. He is primarily interested in the effect of nitrogen application time, and he wants to assess evidence for all possible two and three way interactions. Nar's goal is to select the simplest possible model, and then use this model to describe the effects of the above variables.

We address the following questions in this report:

- Is there evidence that the interaction between nitrogen application and variety differs between inoculated and control plots?
- Is there evidence that the nitrogen application effect varies across varieties? Is there evidence that the nitrogen application effect differs between inoculated and control?
- Is there evidence that the variety effect differs between inoculated and control?
- Conditional on the results to the above questions, how do we appropriately describe the estimated effects for variety, nitrogen application time, and inoculation on the probability of virus infection?

4 Study Design/Data collection

Nar conducted this experiment in a MSU field at the base of the Bridger Mountains. He divided the field into six 31.5 by 27.5 meter blocks. The blocks were then divided into five rows, with 5 meters of space between each row. Each row was randomly assigned to a variety of Winter Wheat with separate randomizations in each block. Three of the chosen varieties are known to be resistant to WSMV (SNMS, TAM 112, and MACE). The pronghorn variety (PRHG) is known to be susceptible to WSMV. The last variety, Yellowstone (YSTN), is a common variety in Montana, and its susceptibility to WSMV is unknown.

After assigning rows to varieties, the rows were then divided into six 1.5 by 5 meter plots. Each plot was randomly assigned to one of six combinations of nitrogen application time

and inoculation status, with separate randomizations in each row. The combinations were fall/inoculated, fall/control, early spring/inoculated, early spring/control, late spring/inoculated, and late spring/control.

Plots assigned to the inoculated treatment had five infected plants transplanted to the middle of the plot. These plants were infected in the greenhouse by clipping an infected leaf to the healthy plant. The plants were infected in the tillering stage of their life cycle at an age of about one month. The Wheat Streak Mosaic Virus is transmitted by the wheat curl mite, tiny organisms that are nearly invisible to the naked eye. They can easily move from leaf to leaf on a plant, and they are transported from plant to plant by the wind (Sloderbeck 2008). It is conceivable that the mites on the infected plants introduced to the inoculated plots could move to healthy plants within that plot. But, Nar says that it is nearly impossible for the mites to move from an inoculated plot to a non-inoculated plot with such a low population of mites. Plants with the virus in non-inoculated plots are assumed to have been infected by mites in the surrounding environment.

All plots were planted in mid-September and allowed to grow over the winter. The following spring and summer, 30 leaves were picked from different plants within each plot and taken to the lab to be tested for presence of the virus. The selection of the leaves from each plot was not random nor technically systematic. The six technicians were instructed to collect a sample of plants spread evenly throughout the plot. There is no possibility of visual bias because the symptoms are not heavy enough in this area of Montana for the plants to show clear visual evidence of the virus.

The sampling was all done in one day on two occasions. The first sampling date occurred at the end of May when the plants were in the tillering stage, and the second sampling date occurred in early July when plants were in the flowering stage. In the first year of the study, $2013 - 2014$, all plots were sampled on both dates. In the second year of the study, only plots with susceptible varieties were sampled on the first date, and all plots were sampled on the second date The advice given in this report is only for analysis of the data collected on the second date in July. After the leaves were picked in the field, they were sent to the lab and screened for the virus via the ELISA procedure. The data are collected and organized in a spreadsheet for the first year. The second year leaves are currently stored in Ziploc bags in the freezer and are waiting to be analyzed in the lab. The researchers do not expect this waiting period to affect their ability to detect the virus.

This is a three year study, and planting for the third year of the study is currently taking place. Each year, the assignment of varieties to rows and treatments to plots is re-randomized. Furthermore, planting takes place in the gaps between the plots from the previous year.

5 Recommendations

5.1 Binomial Generalized Linear Model

Because the response variable is a Binomial count, and because the design includes random assignment of variety to rows within blocks, a Binomial generalized linear model is appropriate. We recommend using a three-way interaction to estimate the probability of infection for each variety/treatment/inoculation status combination and a block/variety interaction to control for differences between blocks and between rows in the same block.
The model is

$$y_i \sim \text{Binomial}(30, p_i),$$

$$\text{logit}(p_i) = \mu + \alpha_{b[i]} + \beta_{v[i]} + \gamma_{t[i]} + \delta_{j[i]} + (\alpha\beta)_{b[i],v[i]} + (\beta\gamma)_{v[i],t[i]} + (\beta\delta)_{v[i],j[i]} + (\gamma\delta)_{t[i],j[i]} + (\beta\gamma\delta)_{v[i],t[i],j[i]}$$

where

- p_i is the probability of infection in the ith plot,
- $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$ is the natural logarithm of the odds of infection, and
- $b[i]$, $v[i]$, $t[i]$, and $j[i]$ index the block, variety, nitrogen application timing, and inoculation status of observation i.

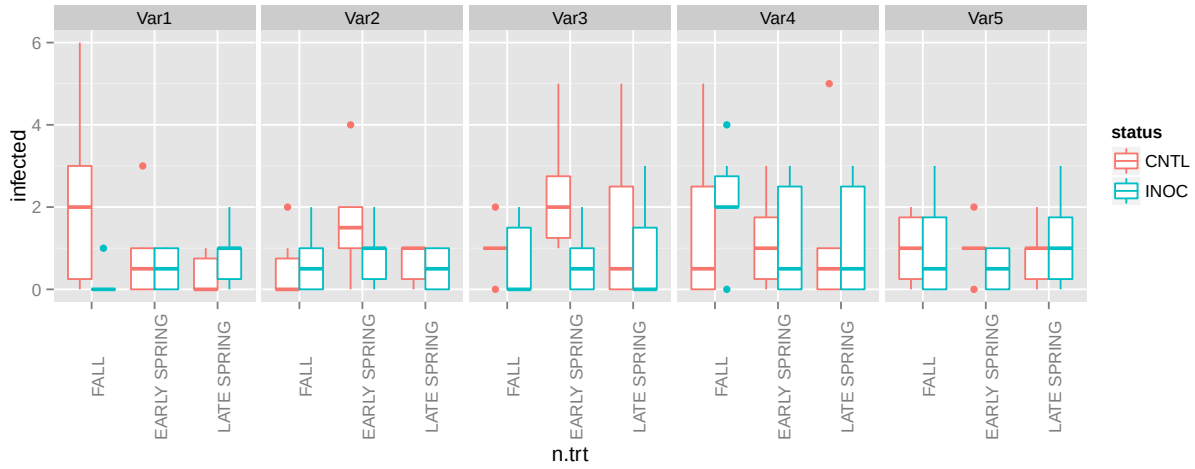
The assumptions for this model are that plants within a plot are independent and have the same probability of being infected, and that plots are independent given the other variables in the model.

There are several reasons why we recommend using a fixed-effects generalized linear model instead of a generalized linear mixed model. We observed overdispersion in Nar's data, and there are no simple tools for including overdispersion in mixed models. Therefore, the standard errors in the mixed model output would be based on an incorrect estimate of the variance and might not be trustworthy. Additionally, including many coefficients and nested effects in a mixed model can cause computational problems where the software will fail to find unique solutions for the parameters. These problems are made worse in a Binomial GLMM when the counts are small, as in these data. The Binomial GLM can adjust for overdispersion and avoid many of the computational difficulties associated with the GLMM. The GLM can correctly model the nested design through a block by variety interaction. The only advantage of the GLMM is that it would estimate block- and row-level variances, but it is our opinion that this is outweighed by the lack of an overdispersion estimate and the computational difficulties.

5.2 Graphical Data Exploration

Before fitting a model, we recommend starting by creating plots for exploratory data analysis. This is useful for seeing which interactions may be present, and for identifying patterns that may represent violations of assumptions or potential computational problems.

Figure 1: Example interaction plot using `qplot`.



A good way to visually compare varieties, treatments, and inoculation status is with boxplots of infection counts for each combination of factor levels. The `qplot` function in the `ggplot2` package is a simple way to do this.

We noticed that infection counts of 0 were recorded in 67 of the 180 plots in Nar's dataset. This is a large fraction of the data; if the distribution of these zeros is related to the variables in the model then it may be impossible to estimate coefficients for factor levels where infection rates are nearly zero.

We recommend sorting the data and creating a heatmap to look for patterns of zeros. Figure 2 shows an example using simulated data. The most serious problems will come from the zeros, so we chose a contrasting color (black) for plots with infection counts of 0.

Any treatment or variety with many black squares is a reason for concern. In the example, all but one of the inoculated fall-application plots for Var1 had counts of 0. The software will have difficulty estimating the the three-way interaction involving fall nitrogen application and variety 1.

If a treatment has many zeros across all varieties or a variety has many zeros, Nar should consider whether he has reason to believe that treatment or variety has qualitative differences from the others that would cause it to have a much lower infection rate. These data would be considered as coming from a different population than the data with fewer zeros. **If and only if it is justified**, Nar could omit all of the data for this treatment or variety when fitting the model. The scope of inference for this model must then be limited to varieties and treatments that have a nonzero infection rate. A separate analysis would be needed for more resistant population.

If information about the physical layout of the field is available, it would be possible to construct a similar heatmap to look for spatial patterns. If correlation between nearby plots

Figure 2: Example heatmap visualization for identifying patterns of zeros.

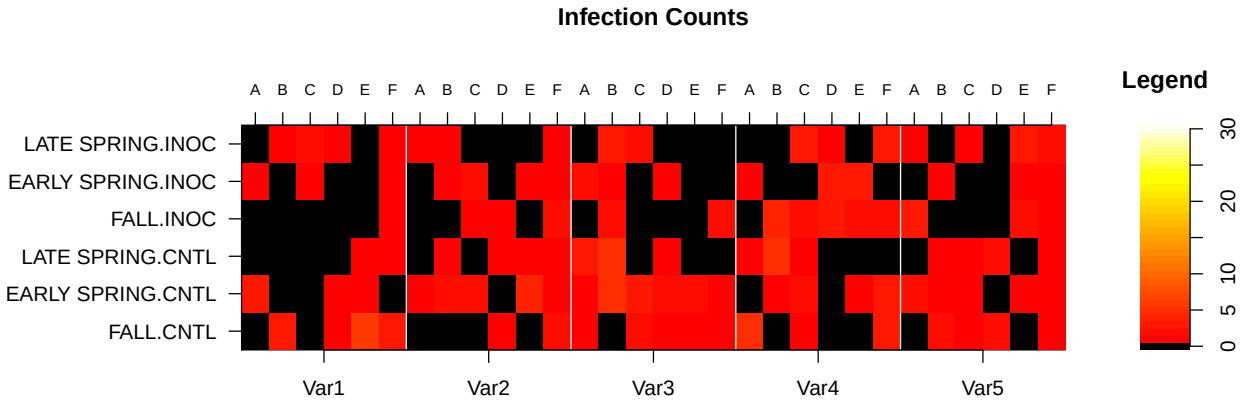
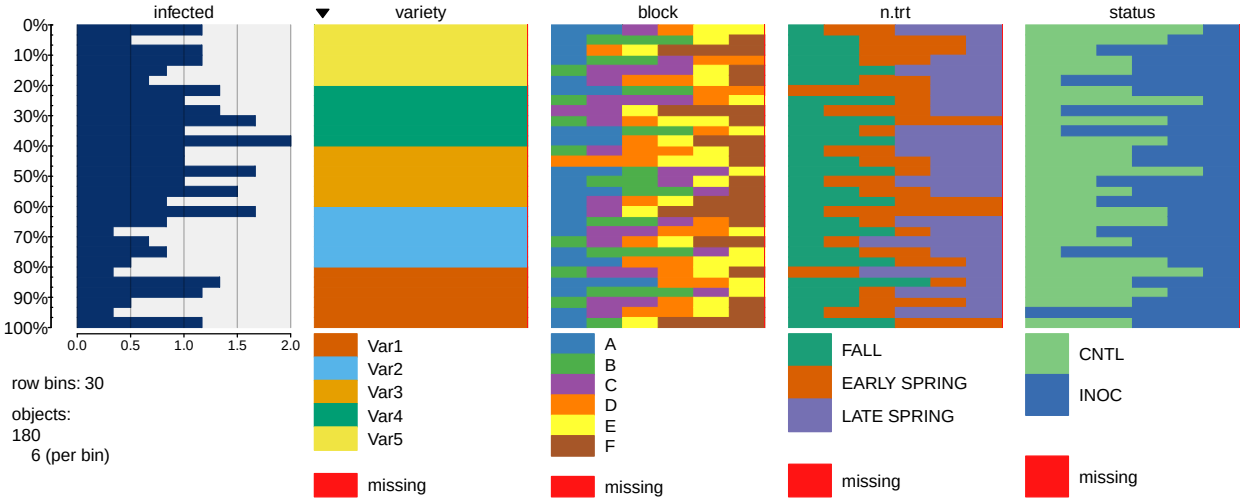


Figure 3: Example `itableplot` output.



is seen, the model should be modified to account for this.

Another option for examining the data graphically is the `itableplot` function in the `tabplot` package. This is a useful interactive tool for visualizing large datasets and identifying patterns.

5.3 Model Fitting and Possible Issues

The following code will fit the three-way interaction model:

```
#this forces the constraint that the block effects sum to 0
contrasts(fert.sim1$block) <- contr.sum(6, contrast=TRUE)
glm3way <- glm(cbind(infected, total) ~ block/variety + variety*n.trt*status2,
               family = quasibinomial, data = fert.sim1)
```

If it is decided that certain observations come from an extremely resistant population and can be omitted from the current analysis:

```
noVar1 <- glm(cbind(infected, total) ~ block*variety + variety*n.trt*status2,
              family = quasibinomial, data = fert.sim1, subset = variety!='Var1')
```

If there is no interest in comparing between inoculated plots and control plots, separate models could be fit:

```
onlyINOC <- glm(cbind(infected, total) ~ block*variety + variety*n.trt,
                family = quasibinomial, data = fert.sim1, subset = status=='INOC')

onlyCNTL <- glm(cbind(infected, total) ~ block*variety + variety*n.trt,
                family = quasibinomial, data = fert.sim1, subset = status=='CNTL')
```

5.4 Model Refinement

In this section, we describe how to go through the process of model selection. As discussed in the previous section, we recommend starting with the most complicated three-way interaction model and then assessing which terms can be removed. In general, start by considering removal of the highest order terms.

We'll start by assessing evidence for the three-way interaction term. One way to consider it's removal would be to fit a model without the three-way interaction, and then conduct a drop-in-deviance F-test comparing the full and reduced models, as you learned to do in Stat 512. This method can be time consuming, so instead we recommend using the `Anova` function in the `car` package with the option `test="F"`. This provides a **Type II** sums of squares table, and the tests provided are similar to extra sums of squares F-tests based on the Pearson residuals (provide reference here to info about pearson resids vs deviance resids). The tests in this table are conditional upon terms of the same and lower order being in the model. For example, the `n.trt:status2` row in the table tests for removal of the $n.trt * status2$ two-way interaction, conditional upon all other two way interactions and main effects being in the model. As a result, this table can be used to assess evidence for removal of each term.

```
Anova(glm3way, test="F")

## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(infected, total)
## Error estimate based on Pearson residuals
##
##                          SS  Df    F Pr(>F)
## block                   7.2   5 1.14  0.344
## variety                 8.7   4 1.71  0.152
## n.trt                   1.2   2 0.49  0.613
## status2                 4.4   1 3.46  0.065
## block:variety          18.5  20 0.73  0.788
```

```
## variety:n.trt           9.9   8 0.98  0.453
## variety:status2         6.4   4 1.26  0.288
## n.trt:status2           2.5   2 1.01  0.368
## variety:n.trt:status2  10.8   8 1.07  0.388
## Residuals             158.2 125
```

Here is how we would interpret the results in each row of the table. Note, it is not appropriate to consider removal of any of the terms involving block. Block to block variability is expected because it was included as a design variable, and so it should be included in the model. The *block* $*$ *variety* interaction also needs to be included to account for the nesting structure in the design.

1. `variety:n.trt:status2` "There is no evidence that the interaction between variety and nitrogen application time differs between inoculated and control plots (p-value= 0.388 from F-stat= 1.07 on 8 and 125 df)."

2. `n.trt:status2` "There is no evidence that the relationship between nitrogen application time and the odds of virus infection differs between inoculated and control plots after controlling for all other two way interactions in the model (p-value= 0.368 from F-stat= 1.01 on 2 and 133 df).

3. `variety:status2` "There is no evidence that the relationship between variety and the odds of virus infection differs between inoculated and control plots after controlling for all other two way interactions in the model (p-value= 0.288 from F-stat= 1.26 on 4 and 133 df).

4. `variety:n.trt` "There is no evidence that the relationship between nitrogen application time and the odds of virus infection depends on variety after controlling for all other two way interactions in the model (p-value= 0.453 from F-stat= 0.98 on 8 and 133 df).

If no evidence is found for any of the interactions, the simplest model we recommend fitting should contain the *block* $*$ *variety* interaction and all main effects (shown below). If this model is selected, each effect is then interpreted after controlling for the effects of the other variables in the model. For example, to assess evidence for the `n.trt` effect in this model, we would say, "There is no evidence that the odds of virus infection depends on nitrogen application time after controlling for block, variety, inoculation status, and the *block*$*$*variety* interaction."

```
glmnointeractions <- glm(cbind(infected, total) ~ block/variety + variety+n.trt
                         + status2,
                family = quasibinomial, data = fert.sim1)
Anova(glmnointeractions, test="F")

## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(infected, total)
## Error estimate based on Pearson residuals
##
```

```
##                   SS  Df    F Pr(>F)
## block            7.3   5 1.15  0.339
## variety          8.6   4 1.70  0.153
## n.trt            1.3   2 0.49  0.612
## status2          4.4   1 3.50  0.063
## block:variety   18.8  20 0.74  0.779
## Residuals      186.7 147
```

The next section describes how to interpret effects once a model is selected.

5.5 Interpretation

In this section, we describe how to interpret effects if evidence of a three-way interaction is found. The effects plots in the `effects` packages provide nice visuals for displaying estimated probabilities of infection across levels of mulitiple variables.

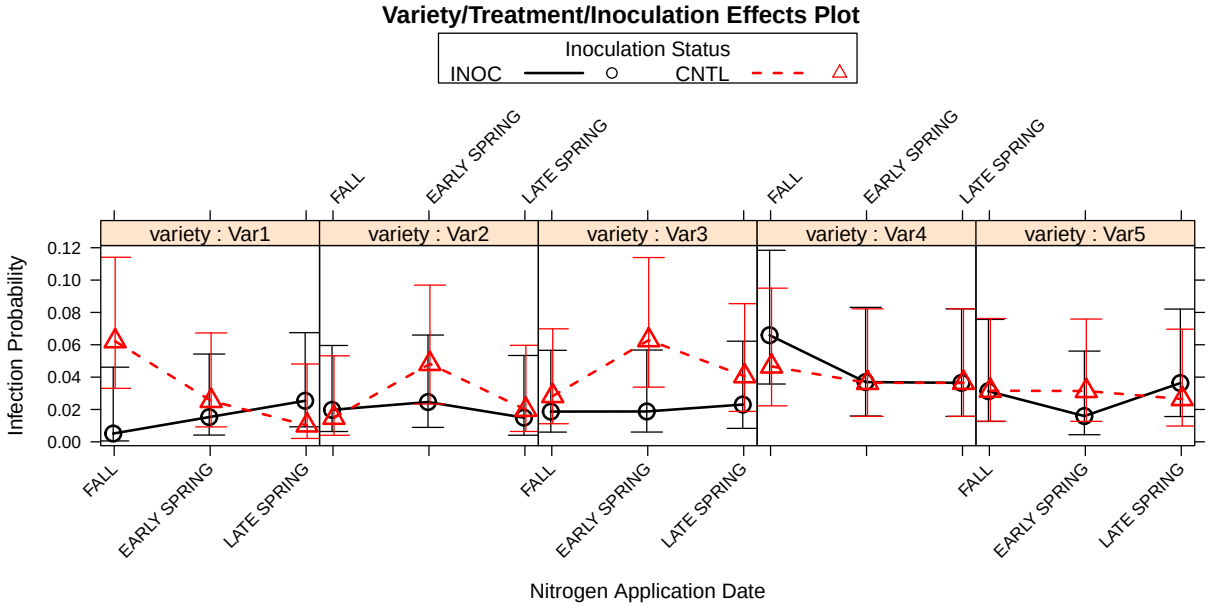


Figure 4: The fitted model can be nicely summarized with plots from the `effects` package.

The `Effect` function is convenient because it provides the estimated probabilities at each treatment combination, averaged over all blocks. The results are provided in a organized table (see below).

```
Effect(c('variety','n.trt','status2'), glm3way, se=TRUE)

##
##  variety*n.trt*status2 effect
## , , status2 = INOC
##
##         n.trt
## variety    FALL EARLY SPRING LATE SPRING
```

```
##    Var1 0.00513       0.0153      0.0254
##    Var2 0.01976       0.0246      0.0149
##    Var3 0.01869       0.0188      0.0230
##    Var4 0.06592       0.0369      0.0365
##    Var5 0.03130       0.0159      0.0363
##
## , , status2 = CNTL
##
##        n.trt
## variety   FALL EARLY SPRING LATE SPRING
##    Var1 0.0622       0.0253      0.0102
##    Var2 0.0149       0.0482      0.0198
##    Var3 0.0283       0.0629      0.0406
##    Var4 0.0466       0.0365      0.0365
##    Var5 0.0316       0.0314      0.0264
```

We recommend displaying the effects plot and output and then describing the effects of interest in words. When describing the effects in words, we recommend describing the effects in terms of multiplicative changes in the odds of virus infection. It is more straightforward to calculate confidence intervals when effects are described are this scale rather than the probability scale. We recommend changing the reference level of the model to find the estimates and confidence intervals for the effects of interest. Notice that when we fit the model in Section 5.3, we specified the constraint that the sum of the block effects would add to 0. As a result, the estimated effects given in the model summary are the effects averaged over all blocks.

You can start by looking at the summary of the model, `summary(glm3way)`. The coefficients of interest will be those found in the rows `n.trtEARLY SPRING` and `n.trtLATE SPRING`. Below we show how to interpret these coefficients in words, and how the meaning changes when the reference level is manipulated.

```
#use the default reference level to describe the effects across
#nitrogen application for inoculated plants of variety 1 (averaged over blocks)
exp(coef(glm3way)[11])

## n.trtEARLY SPRING
##              3.01

exp(confint(glm3way, 11))

##   2.5 %  97.5 %
##   0.295 104.899

exp(coef(glm3way)[12])

## n.trtLATE SPRING
##             5.05

exp(confint(glm3way, 12))

##   2.5 %  97.5 %
##   0.652 166.418
```

For these data, we would describe the nitrogen application effect for inoculated plants of variety 1 as follows:

> For inoculated plants of variety 1, the odds of virus infection when nitrogen is applied in the early spring is estimated to be 3.01 times the odds of virus infection when nitrogen is applied in the fall, with a 95% likelihood-based confidence interval from 0.295 to 104.9 times. For inoculated plants of variety 1, the odds of virus infection when nitrogen is applied in the late spring is estimated to be 5.05 times the odds of virus infection when nitrogen is applied in the fall, with a 95% likelihood-based confidence interval from 0.652 to 166.4 times.

```
#suppose we want to the effects of introgen application for
#inoculated plants of variety 3
#first relevel
var3 <- relevel(fert.sim1$variety, "Var3")
#then fit model with var3
glm.refvar3 <- glm(cbind(infected, total) ~ block/var3 + var3*n.trt*status2,
               family = quasibinomial, data = fert.sim1)
#interpret results
exp(coef(glm.refvar3)[11])
exp(confint(glm.refvar3, 11))
exp(coef(glm.refvar3)[12])
exp(confint(glm.refvar3, 12))

#now suppose we want to the effects of introgen application for
#non-inoculated plants of variety 3
#first relevel the status variable
statuscntl <- relevel(fert.sim1$status2, "CNTL")
#then fit model with var3
glm.refvar3cntl <- glm(cbind(infected, total) ~ block/var3 + var3*n.trt*statuscntl,
               family = quasibinomial, data = fert.sim1)
#interpret results
exp(coef(glm.refvar3cntl)[11])
exp(confint(glm.refvar3cntl, 11))
exp(coef(glm.refvar3cntl)[12])
exp(confint(glm.refvar3cntl, 12))
```

We would describe the nitrogen application effect for inoculated plants of variety 3 as follows:

> For inoculated plants of variety 3, the odds of virus infection when nitrogen is applied in the early spring is estimated to be about equal to the odds of virus infection when nitrogen is applied in the fall, with a 95% likelihood-based confidence interval from 0.191 to 5.28 times. For inoculated plants of variety 3, the odds of virus infection when nitrogen is applied in the late spring is estimated to be 1.24 times the odds of virus infection when nitrogen is applied in the fall, with a 95% likelihood-based confidence interval from 0.269 to 6.201 times.

We would describe the nitrogen application effect for non-inoculated plants of variety 3 as follows:

For non-inoculated plants of variety 3, the odds of virus infection when nitrogen is applied in the early spring is estimated to be 2.3 times the odds of virus infection when nitrogen is applied in the fall, with a 95% likelihood-based confidence interval from 0.797 to 7.72 times. For non-inoculated plants of variety 3, the odds of virus infection when nitrogen is applied in the late spring is estimated to be 1.45 times the odds of virus infection when nitrogen is applied in the fall, with a 95% likelihood-based confidence interval from 0.446 to 5.16 times.

6 Scope of inference

The field where the study took place was not randomly selected from a larger population of fields, so inference from this study does not extend beyond the field at the base of the Bridger mountains where the experiment was conducted. There are potential confounding factors in this field that may not be present in other fields around Montana. For example, the soil in this specific field may have high starting levels of nitrogen from previous studies that could affect the observed relationship between nitrogen application time and probability of virus infection. If inference beyond the study field is desired, it would be necessary to justify that the study field is similar to other wheat fields in the population of interest in terms of soil composition and all other variables.

Plants were randomly assigned to treatments, so it is justified to infer that the observed infection rates in this field are caused by the timing of the nitrogen application and the inoculation status of the plots. A causal relationship *cannot* be inferred between variety and virus infection, however, because plants were not randomly assigned to varieties.

7 Additional Comments

While very low infection rates are, in practice, a desirable result, they present technical and computational problems for analysis. In similar future studies, random sampling should be used within each plot and the sample size should be large enough that the researchers expect at least one infected leaf to be found in each plot.

8 Appendix: R Code

Since Nar expressed a desire for assistance with R, we have included the code we used to create the output for this report. Note that we used a simulated dataset.

8.1 Data Preprocessing

We changed the baseline levels of `n.application` and `status`. If the YLST variety is the primary variety of interest, the `relevel` function could be used on the `variety` variable the same way.

```r
# Correctly order nitrogen application levels, the default is alphabetical
fert.sim1$n.trt <- relevel(fert.sim1$n.application, 'FALL')

# If status=INOC is the most interesting case, make INOC the baseline
fert.sim1$status2 <- relevel(fert.sim1$status, 'INOC')
```

8.2 qplot

```r
require(ggplot2)
qplot(x = n.trt, y = infected, geom = 'boxplot', color = status,
      facets = .~variety, data = fert.sim1) +
  theme(axis.text.x = element_text(angle = 90))
```

8.3 Heatmap

Nar only needs to change `fert.sim1` inside the `arrange` call to the name of his data frame. The rest of this code is self-contained and will create the heatmap with appropriate labels for varieties, treatments, and blocks.

```r
## Heatmap with meaningful ordering

# Order the dataset, using the arrange function from the dplyr package
require(dplyr)
fert.ordered <- arrange(fert.sim1, variety, block, status, n.trt)

# Create a matrix of the arranged responses
infected.arranged <- matrix(fert.ordered$infected, ncol = 6, byrow = TRUE)

# Set up two panels, right one for a legend
layout(t(1:2), widths = c(9, 1))

# Plot the heatmap, with zeros in black and segments separating the blocks
par(mar = c(5, 10, 6, 2)) # Set big margins
image(z = infected.arranged, y = 1:6,

      # Variety blocks are 1 unit wide, centered at 0.5, 1.5, etc
      x = seq(0.5, 5.5, 1/6),

      # Use black for 0, and use heatmap colors (red-orange-yellow-white) for 1 to 30
      col = c('black', heat.colors(30)), zlim = c(0, 30),

      # Don't automatically create axes or labels
      xlab = '', ylab = '', yaxt = 'n', xaxt = 'n')

# Use white line segments to visually separate the varieties
segments(x0 = 1.5:4.5, y0 = 0.5, y1 = 6.5, col = 'white')

# Place a title at the top, and label Varieties, treatment:status levels, and blocks around the image
title('Infection Counts', line = 4)
```

```
axis(3, labels = rep(levels(fert.ordered$block), 5), cex.axis = 0.75,

     # Each block is plotted in a column with width 1/6, so put the labels in the middle
     at = seq(7/12, 5 + 5/12, 1/6))
axis(2, labels = levels(with(fert.ordered, interaction(n.trt, status))), at = 1:6, las = 2)
axis(1, labels = levels(fert.ordered$variety), at = 1:5)

# Legend
par(mar = c(5, 1, 6, 2))
image(y = seq(-0.5, 30.5, 1), z = matrix(0:30, nrow = 1), axes = FALSE, ylab = '',
      col = c('#000000', heat.colors(30)), zlim = c(0, 30))
title('Legend', line = 1.5)
axis(4)
```

8.4 Summary Effects Plot

```
require(effects)
plot(Effect(c('variety','n.trt','status2'), glm3way),
     multiline = TRUE, type = 'response', se = TRUE, ci.style='bars',
     x.var = 'n.trt', rotx = 45, layout = c(5, 1),
     main = 'Variety/Treatment/Inoculation Effects Plot',
     xlab = 'Nitrogen Application Date', ylab = 'Infection Probability',
     key.args = list(title = 'Inoculation Status'))
```

9 References

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Ramsey, F.L., Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis, Third Edition*. Boston, MA: Brooks/Cole, Cengage Learning.

Sloderbeck, J., Michaud, P., Whitworth, Robert. "Wheat Pests." CurlMite. Kansas State University, 1 May 2008. Web. 18 Sept. 2015.
http://entomology.k-state.edu/extension/insect-information/crop-pests/wheat/curlmite.htm