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

Leslie Gains-Germain, Kenny Flagg

Fall 2015

Contents

1 Introduction

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. In this report we describe the study and its potential issues, present our recommendations, and provide example R code.

2 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. The first three questions are related to model selection, and the last question is related to interpretation after a model is selected.

- 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?

3 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.

4 Recommendations

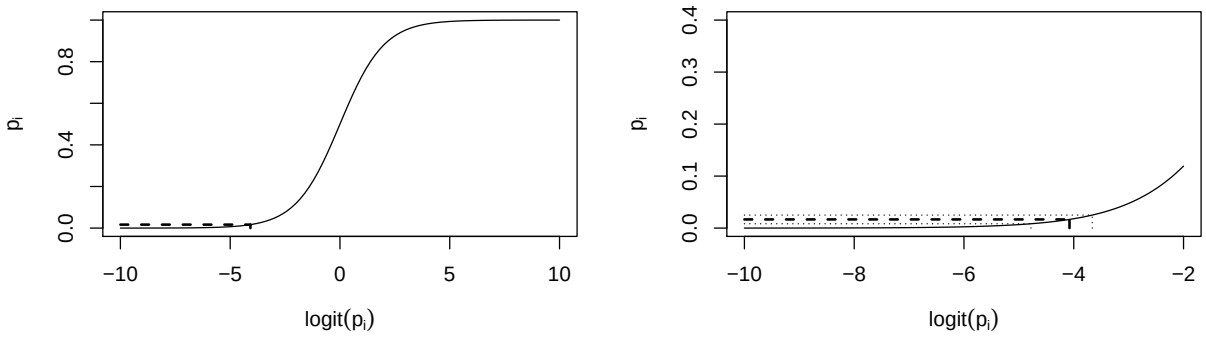
4.1 Potential Issues with Count Data

The response variable is the count of infected leaves out of the 30 leaves collected in each plot. The standard approach to analyzing count data is to use a Binomial distribution and model the probability p_i that a leaf in a given plot i will get infected. This modeling occurs on the scale of the log-odds, $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

One serious problem that can occur when working with counts is separation. This is when one variable or a combination of factors is perfectly associated with the response. For example, if a certain treatment results in an infection count of zero, then knowing that a leaf received that treatment determines that the leaf will not get infected, regardless of other variables. This causes difficulty in estimating the effect of that treatment on the infection rate, and results in a large standard error because the effect cannot be precisely determined.

Figure 1 provides a visual illustration of the separation problem. The left panel shows a small probability on the vertical axis and its corresponding logit value on the horizontal axis. The right panel zooms in on the lower left corner. The dotted lines show how a small amount of uncertainty on the probability scale translates to a relatively larger amount of uncertainty on the logit scale.

Figure 1: Illustration of separation.



The problem gets worse if the infected count is 0. In that case, the best estimate of the infection probability is 0, and $\text{logit}(0) = -\infty$. This will cause computational problems because the software will try to improve the model fit by decreasing the effect estimate toward $-\infty$. The software may fail to find a solution, and if it does report a solution, the standard error will be large because the solution cannot be found precisely.

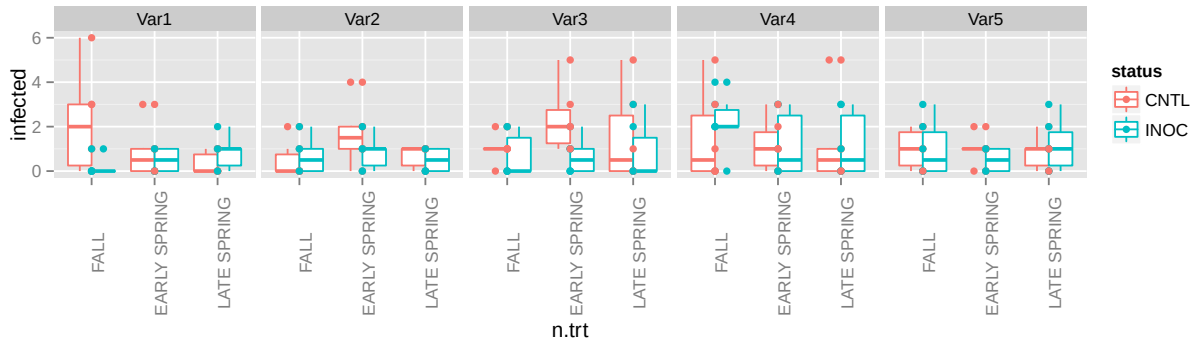
The following sections present examples of issues that could arise while analyzing Nar's data, and our recommendations on dealing with them. Except where stated otherwise, the examples use simulated datasets.

4.2 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.

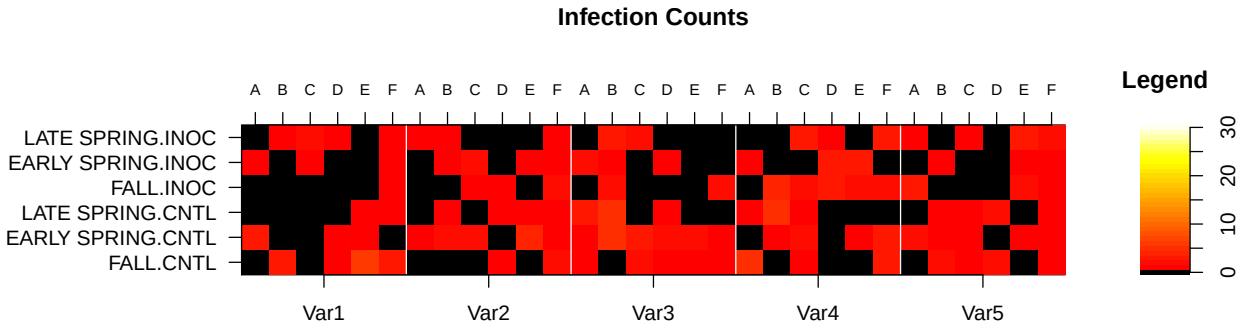
A good way to visually compare varieties, treatments, and inoculation status is with boxplots of infection counts for each combination of factor levels. The `qplot` function in the `ggplot2` package is a simple way to do this.

Figure 2: Example interaction plot using `qplot`.



A graphical analysis is also a good way to see if separation is present in the data. In Nar's dataset, 67 of the 180 counts are 0. This is a large fraction of the data, so it is very important to identify possible separation. We recommend sorting the data and creating a heatmap to look for patterns of zeros. Figure 3 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. The R code to produce this graphic appears in the appendix.

Figure 3: Example heatmap visualization for identifying patterns of zeros.



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.

The graphic could be presented as evidence that the omitted data does not fit the same model as the data with higher infection counts. A separate analysis would be needed if additional information is desired about the 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 is seen, the model should be modified to account for this.

4.3 Quasibinomial Generalized Linear Model

We recommend starting with a quasibinomial generalized linear model. This model includes a three-way interaction to estimate the infection probability 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

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

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

where

- p_i is the probability of infection in the ith plot,
- $\text{logit}(p_i) = \log\left(\dfrac{p_i}{1-p_i}\right)$ 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,
- ω is an overdispersion parameter, so that the variance of the response in plot i is $30\omega p_i(1-p_i)$.

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.

The overdispersion parameter accounts for slight violations of these assumptions. We expect the plants within a single plot to have different infection probabilities based on their proximity to plants where mites are present. This issue manifests itself in the form of excess variation because the probability that a particular plant gets infected will not be the same as the estimated probability for plants in its plot. The overdispersion is an estimate of this extra variability; without it the standard errors will be underestimated.

4.4 Binomial Generalized Linear Mixed Model

Another option is to use random effects to model the nested structure of row within block. This model includes a three-way interaction of fixed effects and does not include an overdispersion parameter. In mixed models, the random effects are meant to explain all the variation in the data, so overdispersed models are not used in this setting.

The model is

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

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

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,
- $a_b \sim \text{N}(0, \sigma_a^2)$ and $(ab)_{b,v} \sim \text{N}(0, \sigma_{ab}^2)$ are random adjustments for each block and each variety within each block.

The assumptions for this model are that plants within a plot are independent and have the same probability of being infected, that blocks are independent, that rows in different blocks are independent, that plots in different rows are independent given the other variables in the model, and that there are no additional sources of variation besides block, row, and the Binomial distribution for the infection counts.

This model accounts for block-to-block and row-to-row variation by including variance terms for the block and row random effects. The benefit of this is that the amount of variation described by the blocks and rows will be estimated. The drawback is that, if there is additional variation not explained by these random effects, the model will fit poorly.

4.5 Advantage of the Quasibinomial Generalized Linear Model over the Binomial Generalized Linear Mixed Model

There are several reasons why we recommend using a fixed-effects generalized linear model instead of a generalized linear mixed model in this report. These reasons are related to overdispersion and computational issues.

Since the inoculated plants were planted in the center of the plots, we expect to see spatial dependence where plants near the centers are more likely to become infected than plants further from the center. Then individual plants could have infection probabilities that differ from the infection rate estimated for the plot they are in. This would result in the data having more variation than what the Binomial model accounts for, a situation known as overdispersion. We observed overdispersion in Nar's data, and mixed models have limited ability to account for overdispersion beyond what can be explained by the variables in the model. Therefore, the standard errors in the mixed model output would be based on an incorrect estimate of the variance and might not be trustworthy. The standard errors in the GLMM model would be smaller than they should be, causing the p-values for effects of interest to be artificially small. This problem is easily remedied by using a quasibinomial fixed-effects GLM model.

We illustrate this problem in the output below. We fit a quasibinomial logistic regression model, and a binomial generalized linear mixed effects model to the real data sent to us by Nar (details on model fitting in section 4.6 and 4.7). The coefficients, standard errors, and tests are shown for the three way interaction coefficients in each model. The estimates are similar between the two models, but the standard errors are much larger in the quasibinomial GLM model. Without accounting for overdispersion, the standard errors for the estimates in the mixed effects model are too small. For this reason, we advise Nar not to trust the results from the mixed effects model.

Table 1: Estimates, standard errors, and tests for three way interaction coefficients in the quasibinomial model.

	Estimate	Std. Error	t value	Pr(>|t|)
varietyPRHN:n.trtEARLY SPRING:status2CNTL	1.34	1.32	1.02	0.31
varietyTAM:n.trtEARLY SPRING:status2CNTL	1.15	1.50	0.76	0.45
varietyYSTN:n.trtEARLY SPRING:status2CNTL	1.92	1.38	1.39	0.17
varietyPRHN:n.trtLATE SPRING:status2CNTL	1.56	1.28	1.22	0.23
varietyTAM:n.trtLATE SPRING:status2CNTL	1.60	1.48	1.08	0.28
varietyYSTN:n.trtLATE SPRING:status2CNTL	1.40	1.54	0.91	0.37

Table 2: Estimates, standard errors, and tests for three way interaction coefficients in the GLMER model

	Estimate	Std. Error	z value	Pr(>|z|)
varietyPRHN:n.trtEARLY SPRING:status2CNTL	1.37	0.66	2.08	0.04
varietyTAM:n.trtEARLY SPRING:status2CNTL	1.16	0.75	1.56	0.12
varietyYSTN:n.trtEARLY SPRING:status2CNTL	1.94	0.69	2.83	0.00
varietyPRHN:n.trtLATE SPRING:status2CNTL	1.57	0.64	2.45	0.01
varietyTAM:n.trtLATE SPRING:status2CNTL	1.61	0.74	2.18	0.03
varietyYSTN:n.trtLATE SPRING:status2CNTL	1.41	0.77	1.84	0.07

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 avoid many of the computational difficulties associated with the GLMM.

Rows are nested within blocks in the study design, and each row is randomly assigned to a variety. The nesting structure of the design can be accounted for in a GLMM model by including random effects for row and block. Originally, however, random effects were developed to reflect a feature of the study design. For example, if 10 operators were chosen at random from a larger population of operators, *operator* would be modeled as a random effect. In this case, the sample of operators truly is random in the design, and the operators selected are intended to represent a larger population of operators. In Nar's study design, row and block are not truly random effects because they were not selected from a larger population of rows and blocks. It is important to recognize this and understand that the only reason row and block are included as random effects in the GLMM model is to account for the nesting structure of row within block. As a result, we think it is more appropriate to model the nested design using fixed effects in the binomial generalized linear model, and this strategy is more consistent with the study design.

The binomial GLM can correctly model the nested design through a block by variety interaction. In the GLM model, it is possible to include a `block/variety` term, and this is equivalent to including the `block*variety` interaction. Including this interaction accounts for the nesting because it allows the differences in varieties to vary by block.

The main advantages of the GLMM is that it would estimate block- and row-level variances, and it would require fewer degrees of freedom to account for the nesting structure in the design. It is our opinion that this is outweighed by the lack of an overdispersion estimate and the computational difficulties.

4.6 Fitting the GLM

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

```
glm3way <- glm(cbind(infected, total) ~ block*variety + variety*n.trt*status2,
               family = quasibinomial, data = fert.sim1)
```

The summary output should be examined for unexpectedly large estimates or standard errors. A warning message may or may not appear in this situation. 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")
```

4.7 Fitting the GLMM

To fit the three-way interaction model:

```
glmm3way <- glmer(cbind(infected, total) ~ variety*n.trt*status2 + (1|block/variety),
                  control = glmerControl(optimizer = "bobyqa"),
                  family = binomial, data = fert.sim1)
```

The `control = glmerControl(optimizer = "bobyqa")` argument tells the `glmer` function which numerical optimizer to use to find the estimates. Several optimizers are included with R, but `bobyqa` works well on problems that are mathematically difficult. We recommend using it for all `glmer` fits.

The following warning message will probably appear:

```
## Warning in commonArgs(par, fn, control, environment()):  maxfun < 10 * length(par)^2 is
not recommended.
```

This message alone does not indicate any problems, but it appears because the model has a large number of parameters, which could cause the numerical optimizer to take many iterations to converge. By default, the algorithm will run for a maximum of 10,000 iterations. That number can be increased by specifying a larger `maxfun` option:

```
glmm3way <- glmer(cbind(infected, total) ~ variety*n.trt*status2 + (1|block/variety),
                  control = glmerControl(optimizer = "bobyqa",
                                         optCtrl = list("maxfun" = 20000)),
                  family = binomial, data = fert.sim1)
```

If a warning message says something like "maximum number of function evaluations exceeded" or mentions "max|grad|" and a tolerance level, then the optimizer ran for the full number of iterations without finding a solution. Increasing `maxfun` as above may solve the problem.

If messages about eigenvalues or "unable to evaluate scaled gradient" occur:

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :  Model is
nearly unidentifiable:  large eigenvalue ratio
##  - Rescale variables?
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :  unable to
evaluate scaled gradient
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :  Model failed
to converge:  degenerate  Hessian with 2 negative eigenvalues
```

These indicate that unique coefficient estimates do not exist. This is most likely caused by separation, so the output should be examined for large standard errors. The best solution would be to fit a simpler model (such as one with fewer interaction terms) or to remove problematic variables or data. The decision to remove any terms or data must be justified by prior knowledge and not based on the R output. If this cannot be justified, then the generalized linear mixed model should not be used.

4.8 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 model and then assessing which terms can be removed. In general, start by considering removal of the highest order terms.

Warning: the results reported in this section are from simulated data. This section is meant to be used as an example that can be followed when working with the real dataset.

We'll start by assessing evidence for the three-way interaction term. One way to consider its 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. 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. These tables can be used to assess all the effects at the same level together. If you choose to remove a term, however, it is important to fit a new model and reconsider the Anova table every time a simpler model is selected. This is necessary because the estimate of overdispersion changes when the model changes, causing the tests in the table to change as well.

The Anova table for the three-way interaction model is shown below (Table 3). This table should only be used to assess evidence for the three-way interaction term. We recommend the following wording,

> 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).

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

Table 3: Anova table for the 3 way interaction model.

	SS	Df	F	Pr(>F)
block	7.19	5	1.14	0.3445
variety	8.66	4	1.71	0.1519
n.trt	1.24	2	0.49	0.6135
status2	4.38	1	3.46	0.0652
block:variety	18.50	20	0.73	0.7876
variety:n.trt	9.95	8	0.98	0.4529
variety:status2	6.40	4	1.26	0.2876
n.trt:status2	2.55	2	1.01	0.3684
variety:n.trt:status2	10.83	8	1.07	0.3884
Residuals	158.18	125		

If you drop the 3 way interaction term, the model with all the two-way interactions should be fit and the Anova table analysis should be repeated.

```
glm2way <- glm(cbind(infected, total) ~ block*variety + variety*n.trt
               + variety*status2+n.trt*status2,
               family = quasibinomial, data = fert.sim1)
Anova(glm2way, test="F")
```

Table 4: Anova table for the model with all 2 way interactions.

	SS	Df	F	Pr(>F)
block	7.25	5	1.11	0.3594
variety	8.66	4	1.65	0.1646
n.trt	1.24	2	0.47	0.6234
status2	4.38	1	3.35	0.0696
block:variety	18.52	20	0.71	0.8129
variety:n.trt	9.95	8	0.95	0.4781
variety:status2	6.40	4	1.22	0.3043
n.trt:status2	2.55	2	0.97	0.3805
Residuals	174.05	133		

Here is how we would interpret the results in the applicable rows in the above the two-way interaction Anova table (Table 4).

1. `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 and main effects in the model (p-value= 0.380 from F-stat= 0.973 on 2 and 133 df).

2. `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 and main effects in the model (p-value= 0.304 from F-stat= 1.22 on 4 and 133 df).

3. `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 and main effects in the model (p-value= 0.478 from F-stat= 0.950 on 8 and 133 df).

If you remove another term from the model, refit the simpler model and reassess evidence for each effect. For example, suppose we remove the variety by inoculation status interaction term (Table 5). Our assessment of the variety by nitrogen application interaction effect in the following simpler model would be,

> There is no evidence that the relationship between nitrogen application time and the odds of virus infection depends on variety after controlling for all main effects and the two way interaction between variety and inoculation status (p-value= 0.4536 from F-stat= 0.98 on 8 and 135 df).

Table 5: Anova table for the model with two way interactions between variety and nitrogen and variety and inoculation status.

	SS	Df	F	Pr(>F)
block	7.26	5	1.11	0.3583
variety	8.65	4	1.65	0.1648
n.trt	1.24	2	0.47	0.6233
status2	4.38	1	3.35	0.0696
block:variety	18.53	20	0.71	0.8125
variety:n.trt	10.27	8	0.98	0.4536
variety:status2	6.71	4	1.28	0.2801
Residuals	176.62	135		

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 (Table 6). 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.

To assess evidence for the `n.trt` effect in the simplest recommended 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 (p-value= 0.612 from F-stat= 0.49 on 2 and 147 df).

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

Table 6: Anova table for the model with no interactions.

	SS	Df	F	Pr(>F)
block	7.27	5	1.15	0.3392
variety	8.65	4	1.70	0.1525
n.trt	1.25	2	0.49	0.6119
status2	4.44	1	3.50	0.0635
block:variety	18.80	20	0.74	0.7793
Residuals	186.65	147		

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

4.9 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 multiple variables.

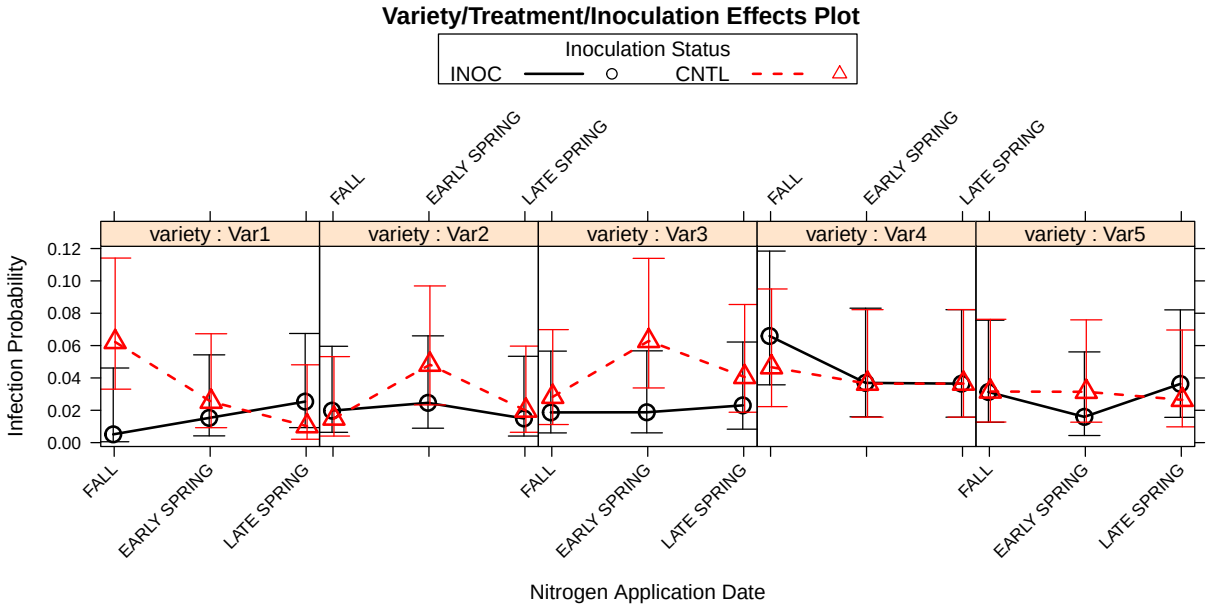


Figure 4: The fitted model can be nicely summarized with plots from the `effects` package.

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 4.6, 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.

5 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.

6 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, the sample size should be large enough that the researchers expect at least one infected leaf to be found in each plot. Additionally we recommend placing the inoculated plants at random locations within the plots, and collecting leaves from a simple random sample of locations in each plot.

7 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.

7.1 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.

```
# 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")

# This forces the constraint that the block effects sum to 0
contrasts(fert.sim1$block) <- contr.sum(6, contrast=TRUE)
```

7.2 qplot

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

7.3 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.

```
## 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(3, 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(3, 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)
```

7.4 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"))
```

8 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.html`.