STAT X12 HW 8

Name(s)

# Rules:

Works in groups of 2 to 4 on this submission. **If you have impactful discussions with colleagues outside your group (or any other students if you are working alone), you must add that information in the “Citations” section.**

## Grass Density model

In Uselman et al. (2018), they conducted an experiment to study different factors that might impact restoration of agricultural fields. The study design is quite complex and we will need some additional discussions to fully understand “split-split plot randomized complete block designs” and how to model them correctly (it is coming though). For now, we will be using linear models (lm) to work with just the data from the “South” field in May 2016 (**note impacts on SOI of doing this** and **that this changes from the previous HW that worked with the 2017 version of the data set**), so you can read about their study design with that target in mind.

For this HW, we will focus on their “Grass Density” response and predictors of “Seeding Strategy” and “Irrigation” and their possible interaction. Figure 2 in their paper does a pretty good job of describing the study design and these predictor variables. These results directly relate to their Table 3 and the row for the interaction of Seeding Strategy and Irrigation in particular. Our results will differ for now because we are just focusing on that particular research question in a more complex model and because we were able to keep more levels of seeding strategy in our model by not considering Grass origin that was not defined for some of the seeding strategies. Focus on this particular interaction for your RQ, with interests in comparing levels of the treatment combinations, whether it ends up in an additive or further reduced model or in the interaction model, whichever you find most appropriate, starting with the interaction setting.

They note that they used log(y + 0.025) for the Grass density measurements, which started as counts of grass per meter squared (which came from counting grass in quadrats). The original information on the counts is not available and the size and number of quadrats searched changed across observations, so we can’t get back to the counts to use Poisson models here. But it is important to note that this was a count variable and would be better modeled that way if we had the original information. There might be some warranted discussion that they observed some 0s in the results…

* Uselman SM, Davison J, Baughman OW, Sullivan BW, Miller WW, Leger EA (2018) Restoring dryland old fields with native shrubs and grasses: Does facilitation and seed source matter? PLoS ONE 13(10): e0205760. <https://doi.org/10.1371/journal.pone.0205760>

Write a statistical report summarizing the results addressing these two predictors and your selected version of the 2016 Grass Density response based on the diagnostics. The researchers used the log(Grass Density + 0.025) version of the response but you are welcome to consider avoiding this transformation - results are provided for the raw responses and their version of the response - you should choose one to focus one for the report and explain why you chose it.

This should lead you to a final model to report and discuss in detail.

Your evidence and any model refinement should be using p-values, not some other model selection criterion. Remember to discuss size of any retained model components in the final model using the effects plots and Tukey’s pairwise comparisons.

For your final model, you should include and discuss R-squared, write out the estimated model with indicators clearly defined and effects plots. Use Tukey’s follow-up tests appropriately. And of course discuss and assess model assumptions.

Summary statistics of the response by combinations of the two predictors would be best generated using favstats to create your “Table 1”. Make sure to address whether this is a balanced design for the two predictors and report a table with counts of observations at each combination or have that information in your “Table 1” so you can discuss this.

**BONUS opportunity: make and discuss an alluvial diagram to support your discussion of the design of the experiment and your predictors of interest, possibly of balanced/not.**

Remember to carefully follow the report writing recommendations. I will be reading carefully for using the word “data” correctly and for writing clarity. Remember that the page limit is up to 4 pages of double-spaced text, but make sure you are focused with your discussions. See the demonstration report from our previous homework for formatting and your answer to the citations question for a start to your references.

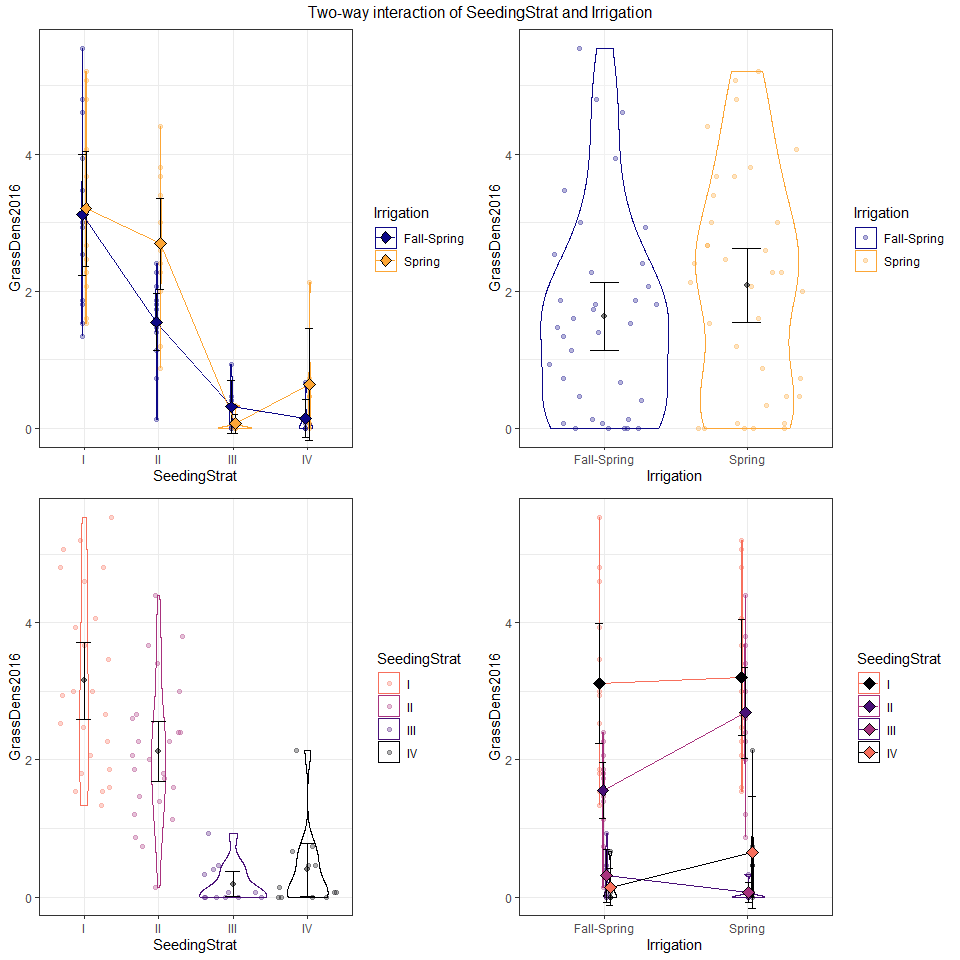
Uselman <- read\_csv("Uselman\_data.csv", na = "-")  
UselmanR <- Uselman %>% dplyr::filter(Site == "South") %>%   
 mutate(Date = factor(Date)) %>%  
 dplyr::filter(Date == "16-May") %>%   
 mutate(SeedingStrat = factor(SeedingStrat),  
 Irrigation = factor(Irrigation),  
 GrassOrig = factor(GrassOrig),  
 Blockf = factor(Block),  
 logShrubDens25 = log(ShrubDens2016 + 0.025),  
 logShrubDens1 = log(ShrubDens2016 + 1))  
  
dim(UselmanR)

## [1] 84 22

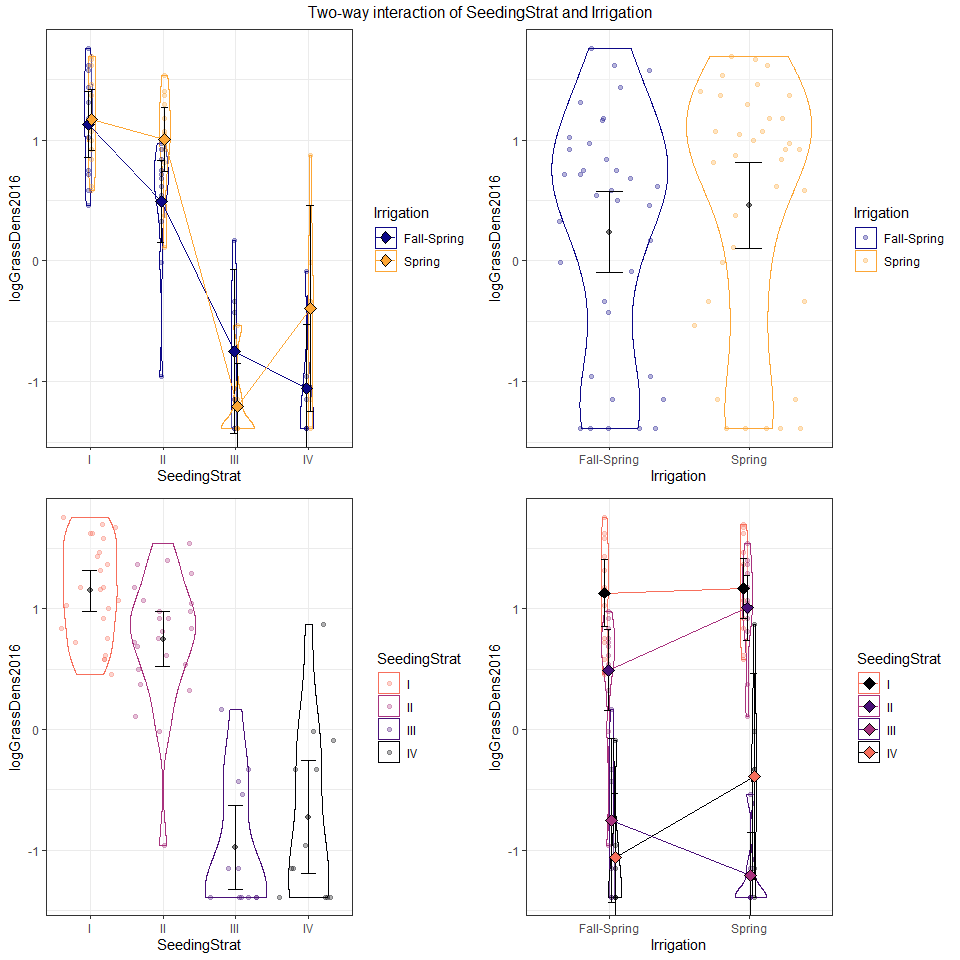
UselmanR <- UselmanR %>% drop\_na(ShrubOrig)   
dim(UselmanR)

## [1] 72 22

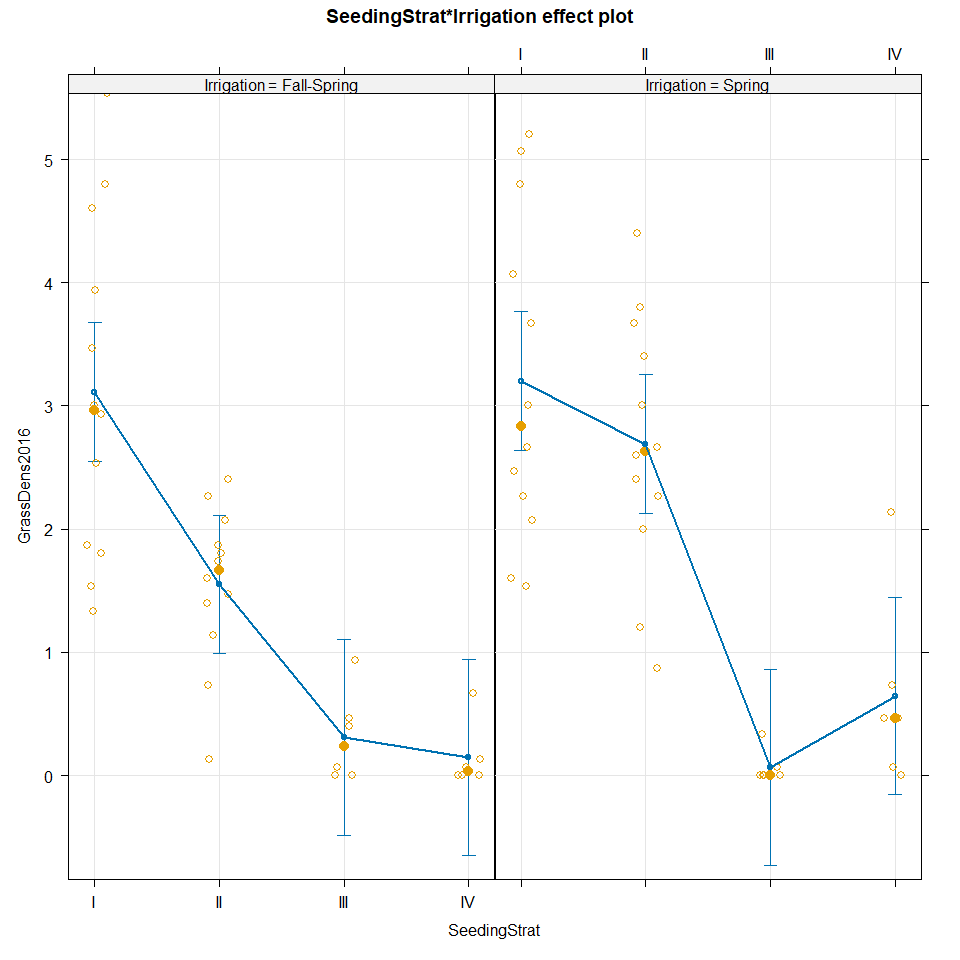
UselmanR <- UselmanR %>% mutate(SeedingStrat = factor(SeedingStrat),  
 Irrigation = factor(Irrigation),  
 GrassOrig = factor(GrassOrig),  
 logGrassDens2016 = log(GrassDens2016 + 0.25))  
  
  
ggintplot(response = "GrassDens2016", groupvars = c("SeedingStrat", "Irrigation"), data = UselmanR)



?ggintplot  
ggintplot(response = "logGrassDens2016", groupvars = c("SeedingStrat", "Irrigation"), data = UselmanR)



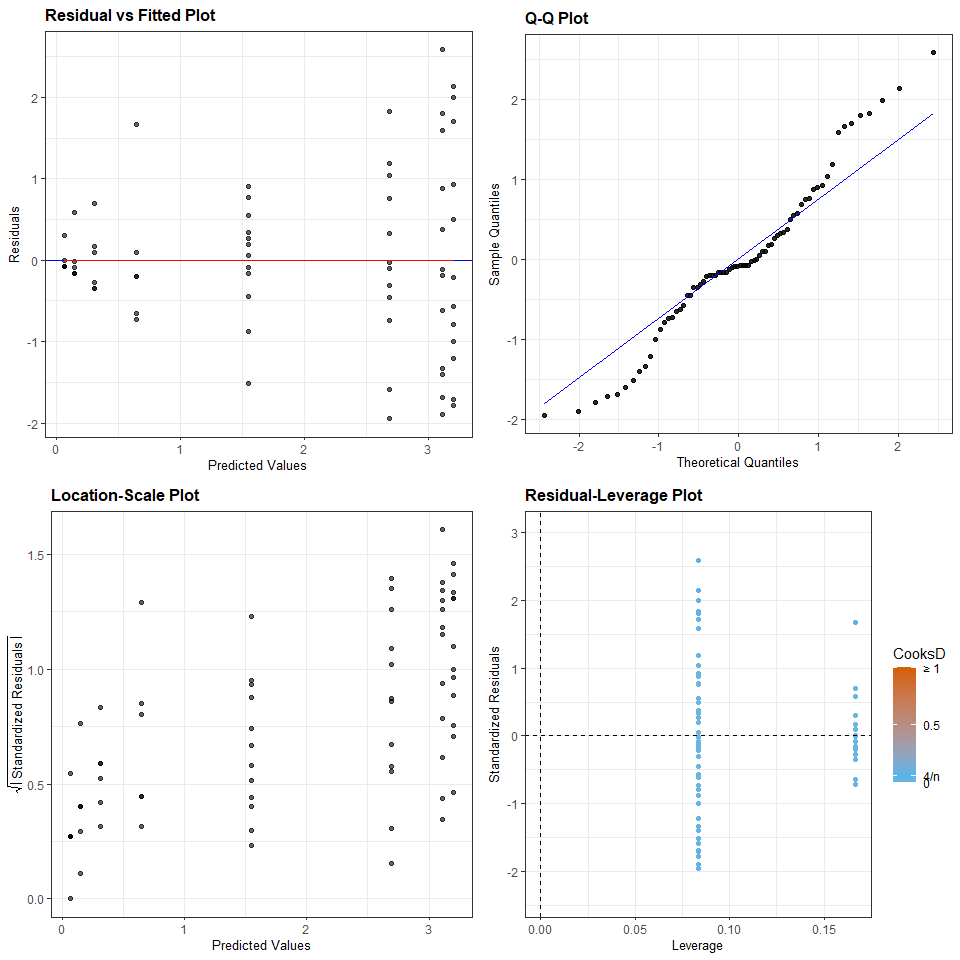
m0 <- lm(GrassDens2016 ~ SeedingStrat\*Irrigation, data = UselmanR)  
plot(allEffects(m0, residuals = T), grid = T)



summary(m0)

##   
## Call:  
## lm(formula = GrassDens2016 ~ SeedingStrat \* Irrigation, data = UselmanR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.82222 -0.45000 -0.07222 0.47917 2.42222   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 3.11111 0.28212 11.028 < 2e-16  
## SeedingStratII -1.56111 0.39897 -3.913 0.000224  
## SeedingStratIII -2.80000 0.48864 -5.730 2.92e-07  
## SeedingStratIV -2.96667 0.48864 -6.071 7.65e-08  
## IrrigationSpring 0.08889 0.39897 0.223 0.824405  
## SeedingStratII:IrrigationSpring 1.05000 0.56423 1.861 0.067348  
## SeedingStratIII:IrrigationSpring -0.33333 0.69104 -0.482 0.631193  
## SeedingStratIV:IrrigationSpring 0.41111 0.69104 0.595 0.553997  
##   
## Residual standard error: 0.9773 on 64 degrees of freedom  
## Multiple R-squared: 0.6427, Adjusted R-squared: 0.6036   
## F-statistic: 16.44 on 7 and 64 DF, p-value: 3.427e-12

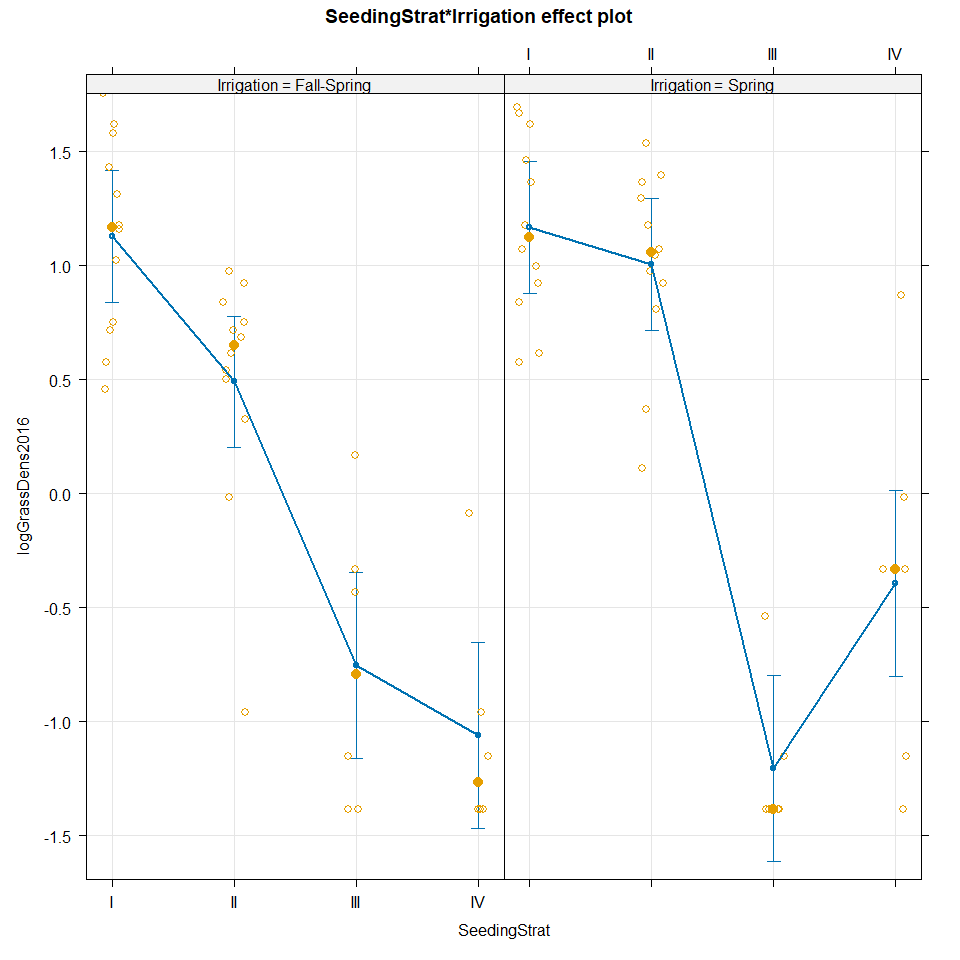
resid\_panel(m0, "R")



Anova(m0)

## Anova Table (Type II tests)  
##   
## Response: GrassDens2016  
## Sum Sq Df F value Pr(>F)  
## SeedingStrat 101.183 3 35.3141 1.379e-13  
## Irrigation 3.675 1 3.8479 0.05416  
## SeedingStrat:Irrigation 5.084 3 1.7744 0.16095  
## Residuals 61.125 64

m1 <- lm(logGrassDens2016 ~ SeedingStrat\*Irrigation, data = UselmanR)  
plot(allEffects(m1, residuals = T), grid = T)



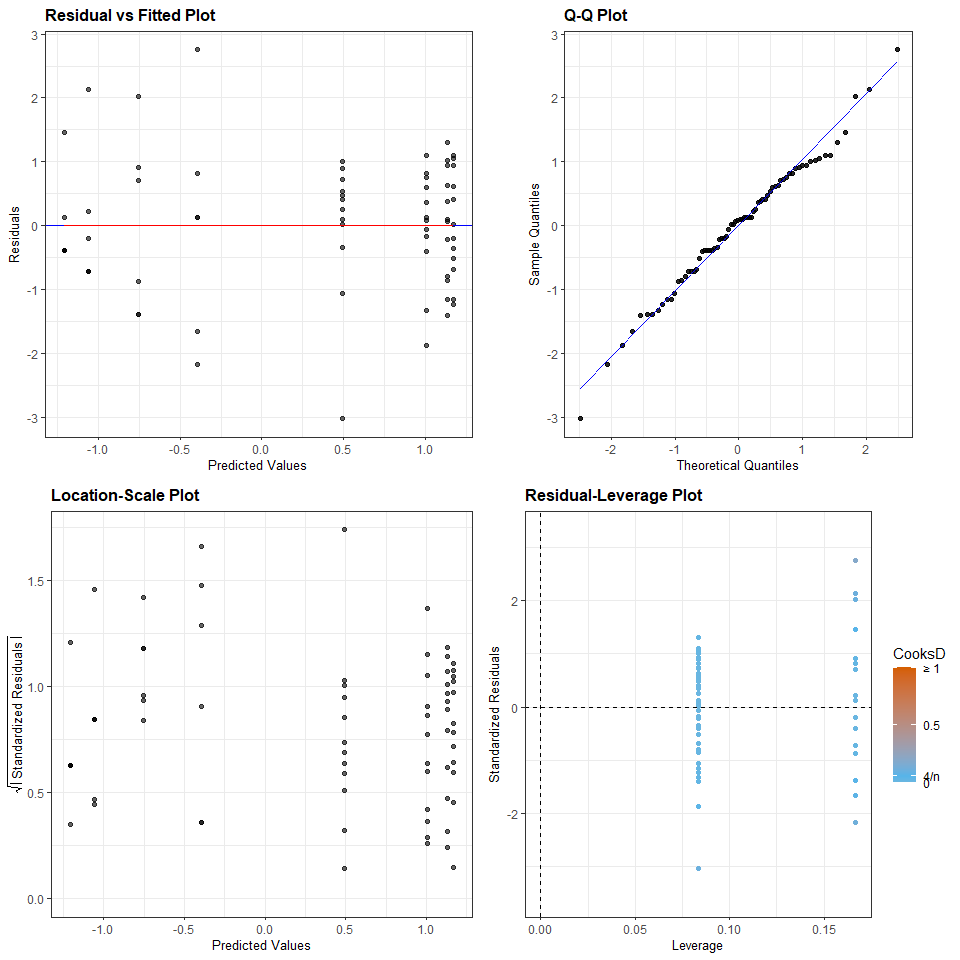
summary(m1)

##   
## Call:  
## lm(formula = logGrassDens2016 ~ SeedingStrat \* Irrigation, data = UselmanR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.45012 -0.32719 0.04442 0.32889 1.26030   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.13027 0.14455 7.819 6.72e-11  
## SeedingStratII -0.63900 0.20442 -3.126 0.00267  
## SeedingStratIII -1.88328 0.25037 -7.522 2.24e-10  
## SeedingStratIV -2.18938 0.25037 -8.745 1.58e-12  
## IrrigationSpring 0.03797 0.20442 0.186 0.85325  
## SeedingStratII:IrrigationSpring 0.47762 0.28910 1.652 0.10341  
## SeedingStratIII:IrrigationSpring -0.49063 0.35407 -1.386 0.17065  
## SeedingStratIV:IrrigationSpring 0.62934 0.35407 1.777 0.08025  
##   
## Residual standard error: 0.5007 on 64 degrees of freedom  
## Multiple R-squared: 0.7827, Adjusted R-squared: 0.7589   
## F-statistic: 32.93 on 7 and 64 DF, p-value: < 2.2e-16

m1 %>% tbl\_regression(intercept = T)

| **Characteristic** | **Beta** | **95% CI***1* | **p-value** |
| --- | --- | --- | --- |
| (Intercept) | 1.1 | 0.84, 1.4 | <0.001 |
| SeedingStrat |  |  |  |
| I | — | — |  |
| II | -0.64 | -1.0, -0.23 | 0.003 |
| III | -1.9 | -2.4, -1.4 | <0.001 |
| IV | -2.2 | -2.7, -1.7 | <0.001 |
| Irrigation |  |  |  |
| Fall-Spring | — | — |  |
| Spring | 0.04 | -0.37, 0.45 | 0.9 |
| SeedingStrat \* Irrigation |  |  |  |
| II \* Spring | 0.48 | -0.10, 1.1 | 0.10 |
| III \* Spring | -0.49 | -1.2, 0.22 | 0.2 |
| IV \* Spring | 0.63 | -0.08, 1.3 | 0.080 |
| *1*CI = Confidence Interval | | | |

resid\_panel(m1, "R")



?emmeans  
Anova(m1)

## Anova Table (Type II tests)  
##   
## Response: logGrassDens2016  
## Sum Sq Df F value Pr(>F)  
## SeedingStrat 54.238 3 72.1058 < 2e-16  
## Irrigation 0.874 1 3.4838 0.06656  
## SeedingStrat:Irrigation 2.681 3 3.5638 0.01889  
## Residuals 16.047 64

m1\_em <- emmeans(m1, pairwise ~ SeedingStrat\*Irrigation, adjust = "tukey")  
m1\_em

## $emmeans  
## SeedingStrat Irrigation emmean SE df lower.CL upper.CL  
## I Fall-Spring 1.130 0.145 64 0.841 1.4190  
## II Fall-Spring 0.491 0.145 64 0.203 0.7800  
## III Fall-Spring -0.753 0.204 64 -1.161 -0.3446  
## IV Fall-Spring -1.059 0.204 64 -1.467 -0.6507  
## I Spring 1.168 0.145 64 0.879 1.4570  
## II Spring 1.007 0.145 64 0.718 1.2956  
## III Spring -1.206 0.204 64 -1.614 -0.7973  
## IV Spring -0.392 0.204 64 -0.800 0.0166  
##   
## Confidence level used: 0.95   
##   
## $contrasts  
## contrast estimate SE df t.ratio p.value  
## (I Fall-Spring) - (II Fall-Spring) 0.639 0.204 64 3.126 0.0510  
## (I Fall-Spring) - (III Fall-Spring) 1.883 0.250 64 7.522 <.0001  
## (I Fall-Spring) - (IV Fall-Spring) 2.189 0.250 64 8.745 <.0001  
## (I Fall-Spring) - I Spring -0.038 0.204 64 -0.186 1.0000  
## (I Fall-Spring) - II Spring 0.123 0.204 64 0.604 0.9987  
## (I Fall-Spring) - III Spring 2.336 0.250 64 9.330 <.0001  
## (I Fall-Spring) - IV Spring 1.522 0.250 64 6.079 <.0001  
## (II Fall-Spring) - (III Fall-Spring) 1.244 0.250 64 4.970 0.0001  
## (II Fall-Spring) - (IV Fall-Spring) 1.550 0.250 64 6.192 <.0001  
## (II Fall-Spring) - I Spring -0.677 0.204 64 -3.312 0.0310  
## (II Fall-Spring) - II Spring -0.516 0.204 64 -2.522 0.2051  
## (II Fall-Spring) - III Spring 1.697 0.250 64 6.778 <.0001  
## (II Fall-Spring) - IV Spring 0.883 0.250 64 3.527 0.0168  
## (III Fall-Spring) - (IV Fall-Spring) 0.306 0.289 64 1.059 0.9629  
## (III Fall-Spring) - I Spring -1.921 0.250 64 -7.674 <.0001  
## (III Fall-Spring) - II Spring -1.760 0.250 64 -7.029 <.0001  
## (III Fall-Spring) - III Spring 0.453 0.289 64 1.566 0.7684  
## (III Fall-Spring) - IV Spring -0.361 0.289 64 -1.249 0.9135  
## (IV Fall-Spring) - I Spring -2.227 0.250 64 -8.896 <.0001  
## (IV Fall-Spring) - II Spring -2.066 0.250 64 -8.252 <.0001  
## (IV Fall-Spring) - III Spring 0.147 0.289 64 0.507 0.9996  
## (IV Fall-Spring) - IV Spring -0.667 0.289 64 -2.308 0.3054  
## I Spring - II Spring 0.161 0.204 64 0.789 0.9931  
## I Spring - III Spring 2.374 0.250 64 9.482 <.0001  
## I Spring - IV Spring 1.560 0.250 64 6.231 <.0001  
## II Spring - III Spring 2.213 0.250 64 8.837 <.0001  
## II Spring - IV Spring 1.399 0.250 64 5.586 <.0001  
## III Spring - IV Spring -0.814 0.289 64 -2.815 0.1093  
##   
## P value adjustment: tukey method for comparing a family of 8 estimates

pwpm(m1\_em)

## I Fall-Spring II Fall-Spring III Fall-Spring IV Fall-Spring  
## I Fall-Spring [ 1.130] 0.0510 <.0001 <.0001  
## II Fall-Spring 0.639 [ 0.491] 0.0001 <.0001  
## III Fall-Spring 1.883 1.244 [-0.753] 0.9629  
## IV Fall-Spring 2.189 1.550 0.306 [-1.059]  
## I Spring -0.038 -0.677 -1.921 -2.227  
## II Spring 0.123 -0.516 -1.760 -2.066  
## III Spring 2.336 1.697 0.453 0.147  
## IV Spring 1.522 0.883 -0.361 -0.667  
## I Spring II Spring III Spring IV Spring  
## I Fall-Spring 1.0000 0.9987 <.0001 <.0001  
## II Fall-Spring 0.0310 0.2051 <.0001 0.0168  
## III Fall-Spring <.0001 <.0001 0.7684 0.9135  
## IV Fall-Spring <.0001 <.0001 0.9996 0.3054  
## I Spring [ 1.168] 0.9931 <.0001 <.0001  
## II Spring 0.161 [ 1.007] <.0001 <.0001  
## III Spring 2.374 2.213 [-1.206] 0.1093  
## IV Spring 1.560 1.399 -0.814 [-0.392]  
##   
## Row and column labels: SeedingStrat:Irrigation  
## Upper triangle: P values adjust = "tukey"  
## Diagonal: [Estimates] (emmean)   
## Lower triangle: Comparisons (estimate) earlier vs. later

multcomp::cld(m1\_em, alpha = 0.05, Letters = LETTERS)

## SeedingStrat Irrigation emmean SE df lower.CL upper.CL .group  
## III Spring -1.206 0.204 64 -1.614 -0.7973 A   
## IV Fall-Spring -1.059 0.204 64 -1.467 -0.6507 A   
## III Fall-Spring -0.753 0.204 64 -1.161 -0.3446 A   
## IV Spring -0.392 0.204 64 -0.800 0.0166 A   
## II Fall-Spring 0.491 0.145 64 0.203 0.7800 B   
## II Spring 1.007 0.145 64 0.718 1.2956 BC   
## I Fall-Spring 1.130 0.145 64 0.841 1.4190 BC   
## I Spring 1.168 0.145 64 0.879 1.4570 C   
##   
## Confidence level used: 0.95   
## P value adjustment: tukey method for comparing a family of 8 estimates   
## significance level used: alpha = 0.05   
## NOTE: If two or more means share the same grouping symbol,  
## then we cannot show them to be different.  
## But we also did not show them to be the same.

library(modelsummary)  
datasummary\_balance(GrassDens2016 ~ SeedingStrat, data = UselmanR, output = 'markdown',  
 title = 'Grass ~ Seeding Strat')

Grass ~ Seeding Strat

|  | I (N=24) | | II (N=24) | | III (N=12) | | IV (N=12) | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. |
| GrassDens2016 | 3.2 | 1.3 | 2.1 | 1.0 | 0.2 | 0.3 | 0.4 | 0.6 |

datasummary\_balance(GrassDens2016 ~ Irrigation, data = UselmanR, output = 'markdown',  
 title = 'Grass ~ Irrigation')

Grass ~ Irrigation

|  | Fall-Spring (N=36) | | Spring (N=36) | |
| --- | --- | --- | --- | --- |
|  | Mean | Std. Dev. | Mean | Std. Dev. |
| GrassDens2016 | 1.6 | 1.5 | 2.1 | 1.6 |

datasummary\_balance(SeedingStrat + GrassDens2016 ~ Irrigation, data = UselmanR, output = 'markdown',  
 title = 'Grass ~ Irrigation & Seeding:IDK if this helps at all')

Grass ~ Irrigation & Seeding:IDK if this helps at all

|  | | Fall-Spring (N=36) | | Spring (N=36) | |
| --- | --- | --- | --- | --- | --- |
|  |  | Mean | Std. Dev. | Mean | Std. Dev. |
| GrassDens2016 |  | 1.6 | 1.5 | 2.1 | 1.6 |
|  |  | N | Pct. | N | Pct. |
| SeedingStrat | I | 12 | 33.3 | 12 | 33.3 |
|  | II | 12 | 33.3 | 12 | 33.3 |
|  | III | 6 | 16.7 | 6 | 16.7 |
|  | IV | 6 | 16.7 | 6 | 16.7 |

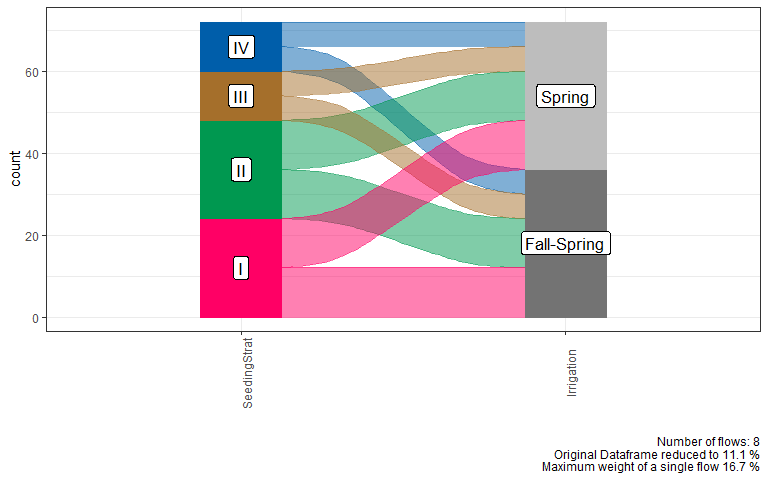
?ggintplot

Bonus stuff

tally(SeedingStrat~Irrigation, data = UselmanR)

## Irrigation  
## SeedingStrat Fall-Spring Spring  
## I 12 12  
## II 12 12  
## III 6 6  
## IV 6 6

?tally  
alluvial\_wide(data = UselmanR %>% dplyr::select(SeedingStrat,Irrigation))

 This is a balanced design. Using the tally output, we can see that observations are balanced and there are enough data to conduct and interaction model. Additionally, the alluvial plot visually depicts that the data frame is balanced in observations between the two categorical predictors, seeding strategy and irrigation.