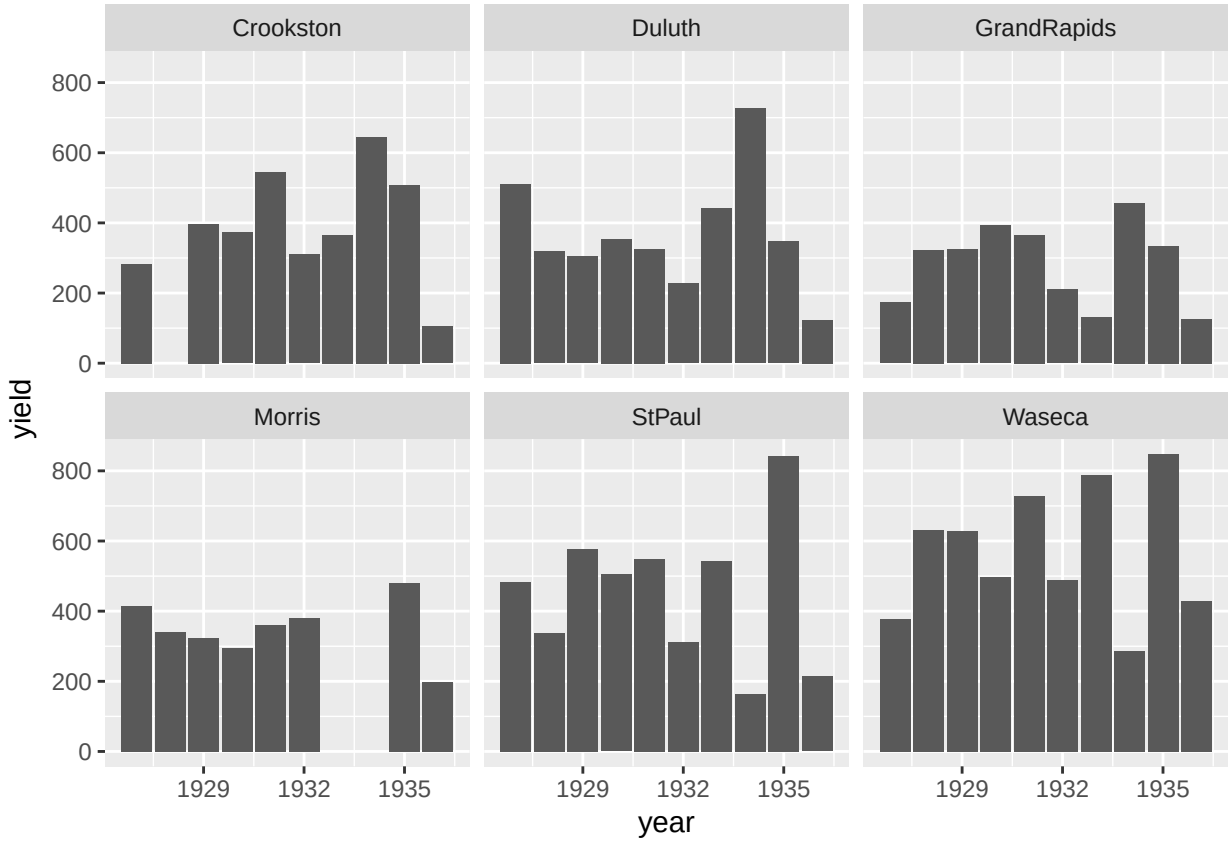
S670 Problem set 6

Erik Parker

March 20, 2017

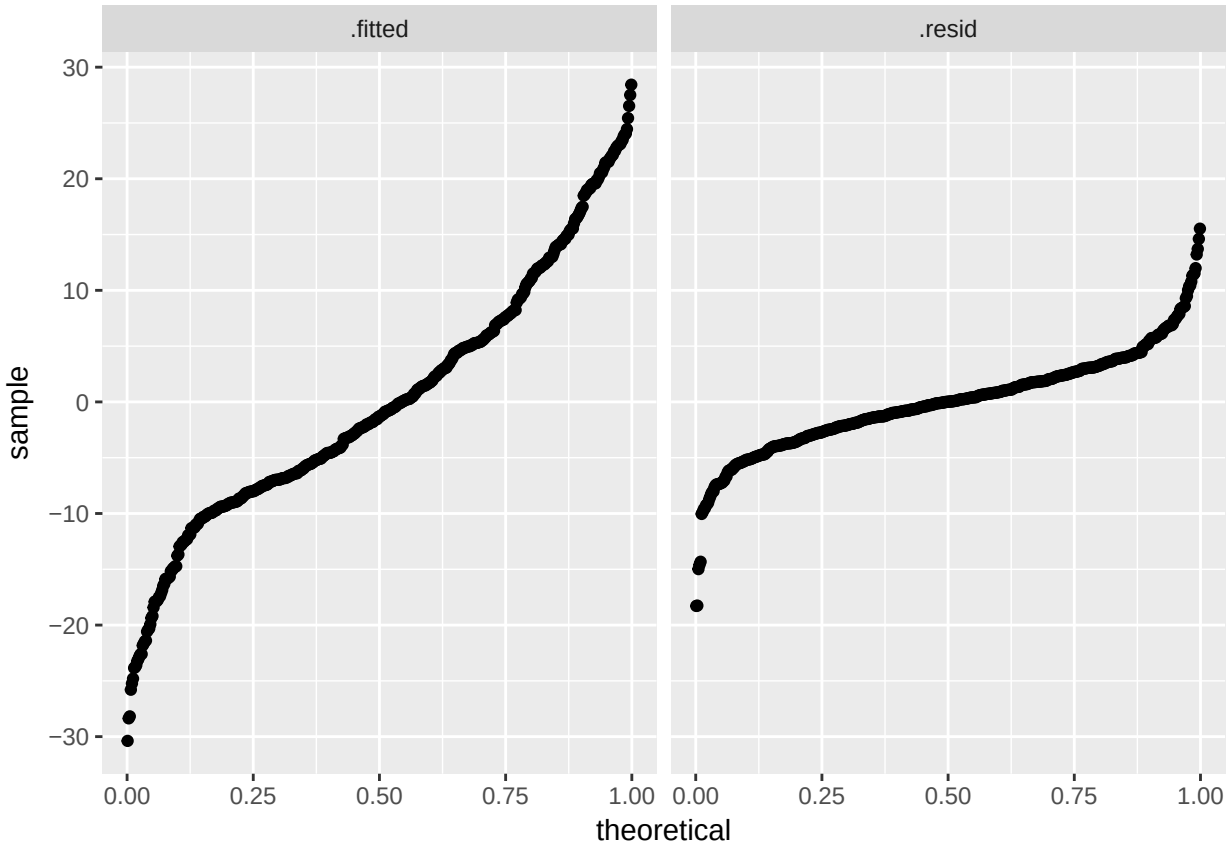
1. Draw an appropriate faceted graph showing how barley yield varied by year at each location, using color as necessary. When looking at successive years, was it more common for the yields to move in the same direction at all locations, or was it more common for the yields to increase at some locations and decrease at others?



The above plot shows that in general, the year to year yields of the sites seem to not move in the same direction at all the locations. Also in general, it seems that when there is disagreement in the directionality of yield changes between the sites (some sites increase year to year, while others decrease), Crookstown, Duluth, and Grand Rapids consistently move together in one direction, while St. Paul, Waseca, and to a lesser extent Morris move together in the other. The most clear example of this pattern can be seen from 1934 until 1936 where yields at the first three sites named decrease, while the others increase then decrease again. Also of note here is the disagreement between the yield change from 1931-32 between Morris and the other sites. It is clear here that the yields at Morris, very, slightly increase over that time span while the other sites all show some degree of reduction. The magnitude of increase at the Morris site is low enough though, and there are other sites with similiarly low decreases, that from just viewing the data alone it seems reasonable to attribute this contentious pattern to random variation in the data. In general it seems that the sites have their own shifting patterns that sometimes can be generalized across multiple sites, but more often are unique to that location.

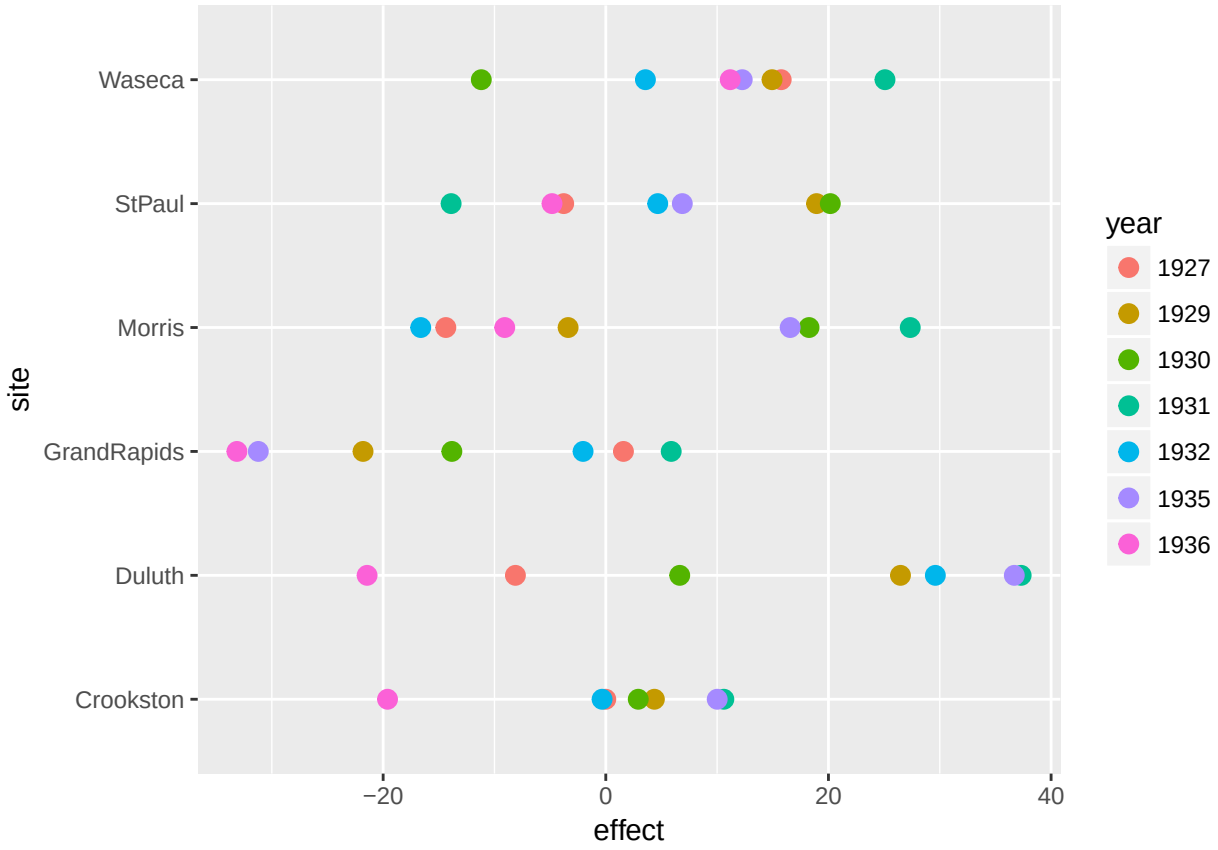
2. Fit a model with yield as the response and gen (variety), year, and site as explanatory variables, with the goal of determining whether Morris 1931-1932 is an anomaly. Justify why you chose this model and not some other one. Because of outliers, you should use a robust fitting method.

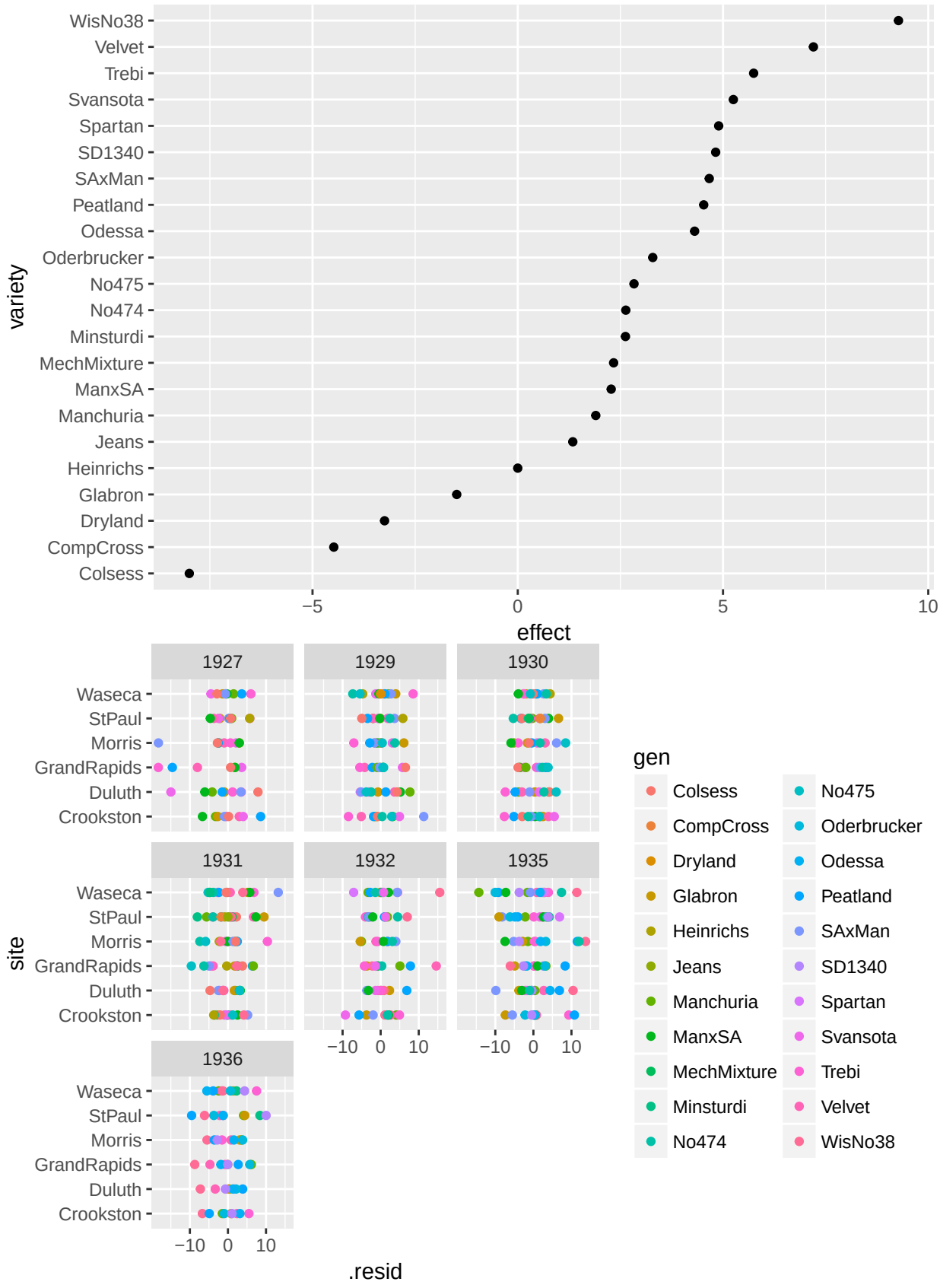
```
model.rlm <- rlm(yield ~ site * year + gen, data = barley.factor, psi = psi.bisquare)
```

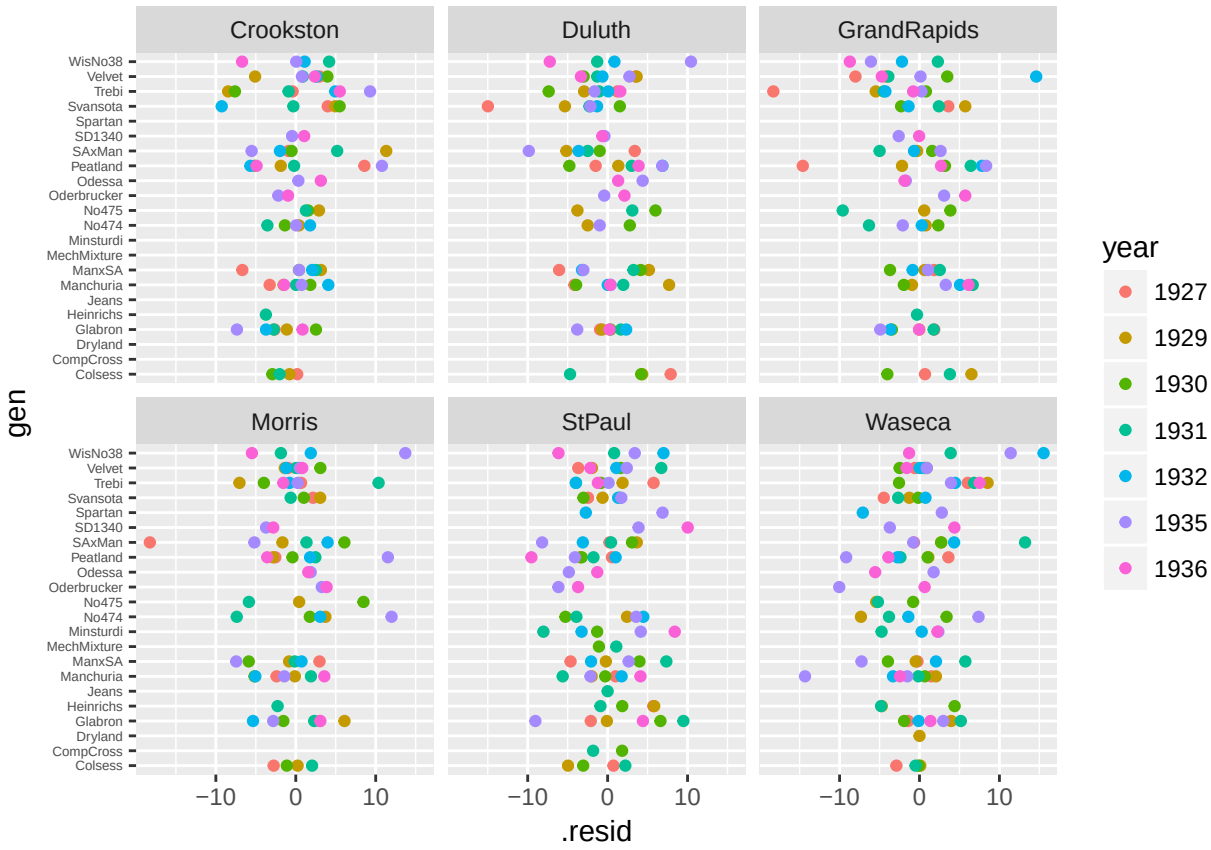


This model was fit as it generated to the best residual-fit plot out of all the alternative models attempted. I chose to drop the years 1928, 33, and 34 from the data used to construct my model as they had missing values which prevented me from performing an interaction between site and year with them included, an interaction that proved to be necessary when the residauls were examined. Dropping these years removed a lot of the largest outliers from the data, and allowed for the best performing model to be constructed while still retaining the majority of the data, so I believe it was an appropriate step to take.

3. Draw plots of the fit and/or residuals with the goal of determining whether Morris 1931-1932 is a mistake, or whether it can be explained as natural variation. As best as you can tell, was there a mistake?




WisNo38
Velvet
Trebi
Svansota
Spartan
SD1340
SAxMan
Peatland
Odessa
Oderbrucker
No475
No474
Minsturdi
MechMixture
ManxSA
Manchuria
Jeans
Heinrichs
Glabron
Dryland
CompCross
Colsess
variety
−5
0
5
10
effect
1927
1929
1930
1931
1932
1935
1936
Waseca
StPaul
Morris
GrandRapids
Duluth
Crookston
site
−10
0
10
.resid
gen
Colsess
CompCross
Dryland
Glabron
Heinrichs
Jeans
Manchuria
ManxSA
MechMixture
Minsturdi
No474
No475
Oderbrucker
Odessa
Peatland
SAxMan
SD1340
Spartan
Svansota
Trebi
Velvet
WisNo38



Based on these new plots of the fit of my model, and the residuals plots of said model (two not shown here due to size issues, but included in the code), it seems as though there is a large amount of natural variation in the barley yields between these sites, for the different years, and for the different varieties. This leads me to believe that the data from Morris between 1931 and 1932 was not a reversal mistake, and was instead just a result of the extensive barley yield variation seen across years within, and between sites in these data. So, as best as I can tell, the reversal in the Morris data was not a mistake, and was instead just a natural result of the large amount of variationpresent in these data.