Untitled

Vatsal Jatakia

March 22, 2018

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(MASS)
barley = read.table("minnesota.txt",header = T)
summary(barley)
```

```
##      yield              gen           year              site
##  Min.   : 2.90   Manchuria: 58   Min.   :1927   Crookston  : 99
##  1st Qu.:26.80   Peatland : 57   1st Qu.:1929   Duluth     :107
##  Median :34.40   Trebi    : 57   Median :1932   GrandRapids:105
##  Mean   :35.63   Velvet   : 57   Mean   :1932   Morris     : 84
##  3rd Qu.:44.45   Glabron  : 56   3rd Qu.:1934   StPaul     :127
##  Max.   :75.50   ManxSA   : 51   Max.   :1936   Waseca     :125
##                  (Other)  :311
```

Q.1

```
ggplot(barley, aes(year, yield, color = site))+geom_line()+facet_grid(~site)+ggtitle("Variation of Yiel
```

For Crookston, the trend seems to be increasing until 1932 after which there is a steep decrease followed by another steep increase in 1934 again followed by a decrease. This trend is also observed in Duluth and St.Paul (except that the increase was in 1935). For Waseca and Grand Rapids, the trend remains constant followed by a decrease in 1933 and 1932 respectively and then it slightly increases before becoming constant again. However, for Morris the trend is quite different with a combination of increases and decreases. Hence, the trend is quite different for different sites.

Q.2

```
lm1 = rlm(yield~gen+year+site, barley, psi = psi.bisquare)
summary(lm1)
```

```
##
## Call: rlm(formula = yield ~ gen + year + site, data = barley, psi = psi.bisquare)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -35.2599  -7.1478  -0.5906   7.6154  41.5213
##
## Coefficients:
##                  Value     Std. Error t value
## (Intercept)     2234.8392  382.9779     5.8354
## genCompCross       1.8578    8.8990     0.2088
## genDryland        -3.4244    7.3629    -0.4651
## genGlabron         2.6368    2.8068     0.9394
## genHeinrichs       1.5482    4.1718     0.3711
## genJeans          -5.5720   12.3399    -0.4515
## genManchuria       1.1431    2.7979     0.4086
## genManxSA          5.4200    2.8364     1.9109
## genMechMixture     2.0578    8.8990     0.2312
## genMinsturdi       3.6424    4.1723     0.8730
## genNo474           3.4725    3.0358     1.1439
## genNo475           2.4655    3.8542     0.6397
## genOderbrucker    -1.4771    3.5897    -0.4115
## genOdessa          7.4590    3.5897     2.0779
## genPeatland        3.6242    2.7990     1.2948
## genSAxMan          6.0837    2.8486     2.1356
## genSD1340          5.8720    3.6911     1.5908
## genSpartan         3.8593    4.9513     0.7794
## genSvansota        2.0397    2.9574     0.6897
## genTrebi           7.9176    2.7990     2.8288
## genVelvet          2.2252    2.7990     0.7950
## genWisNo38         8.8736    3.1801     2.7904
## year              -1.1404    0.1985    -5.7452
## siteDuluth        -3.1490    1.6882    -1.8653
## siteGrandRapids   -8.6585    1.6938    -5.1119
## siteMorris        -3.4304    1.8015    -1.9042
## siteStPaul        -0.2594    1.6623    -0.1561
## siteWaseca        10.9905    1.6485     6.6668
##
## Residual standard error: 11 on 619 degrees of freedom
```

The scatterplot of this yield vs year suggests a high number of outliers which strengthens my choice of the robust model. The other models like linear and loess were giving a relatively low r-square values as well.

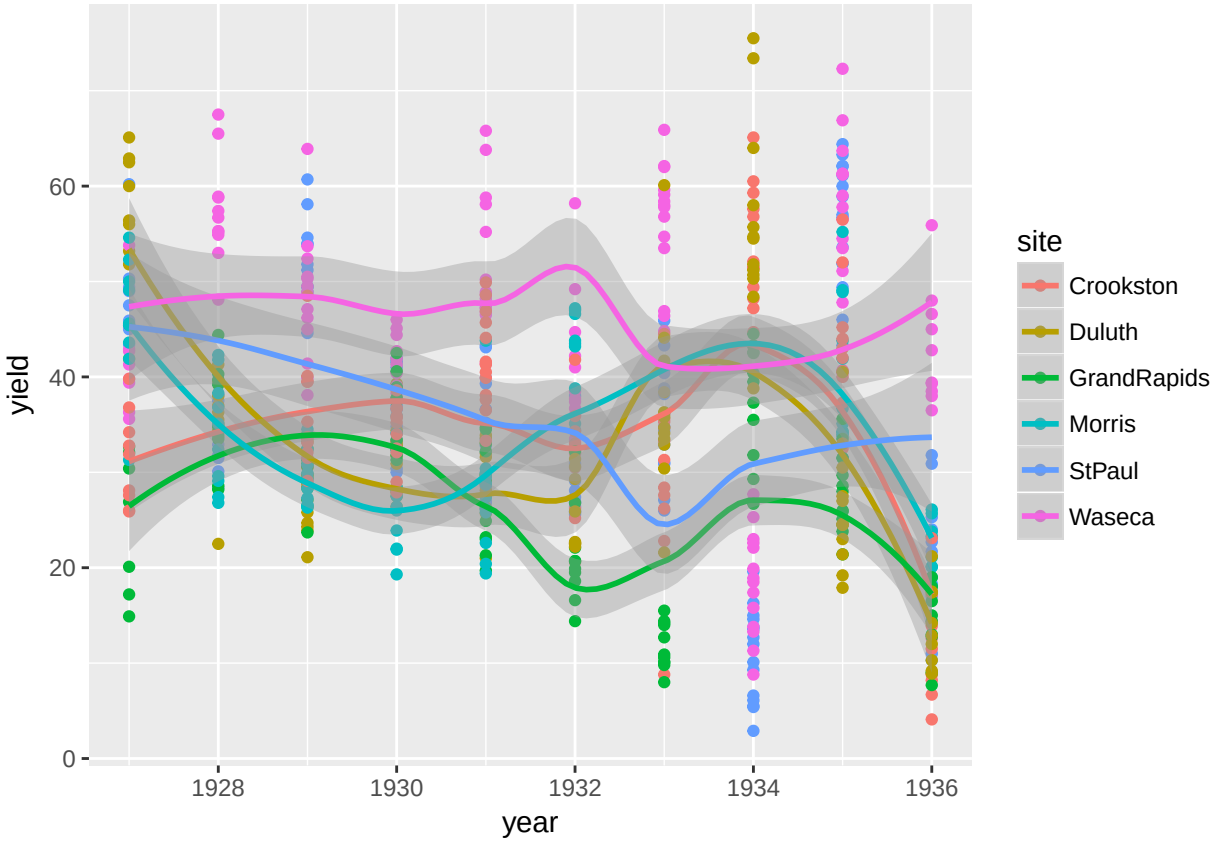
Q.3

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.4.3
```

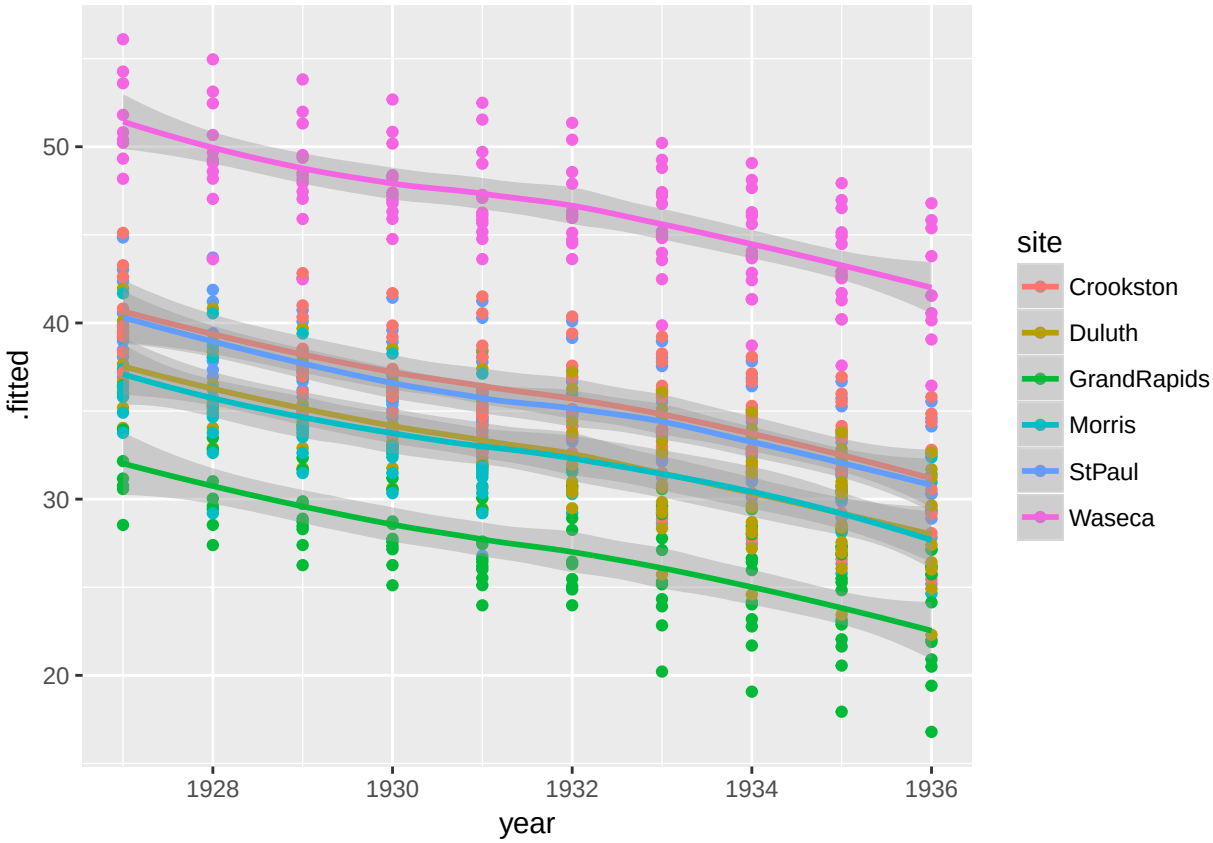
```
df1 = augment(lm1)

ggplot(barley, aes(year, yield, color = site))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```




```
ggplot(df1, aes(x = year, y = .fitted, color = site))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```



Considering the Morris trend line from 1931 to 1932 in the above 2 plots, the plot for fitted values suggests a decrease in the trend, while the actual trend suggests an increase in the yield from 1931 to 1932. Due to these differences in observations, the opinion shifts towards the possibility of a mistake being committed while entering the data and that the yields in 1931 and 1932 might have been reversed.