PFI: Yields

Evan “Pete” Walsh

October 27, 2014

Yields for SB and Corn in acre buschel per pound.

Read data and load libraries.

```r
pfi <- read.csv("/Users/marianwaitwalsh/GitHub/PFI/data/PFI_clean.csv")
weather <- read.csv("/Users/marianwaitwalsh/GitHub/PFI/data/IA_annual_rainfall_raw.csv")
library(dplyr)
library(tidyr)
library(reshape2)
library(ggplot2)
library(leaps)
```

Subset the `PFI` data to get yields for just corn and SB.

```r
yields <- pfi %>%
  filter(item_type == "Unit Quantity", crop %in% c("Corn","SB")) %>%
  select(-c(item, item_type)) %>%
  group_by(field_id, year)
```

Clean the weather data and join with `yields`.

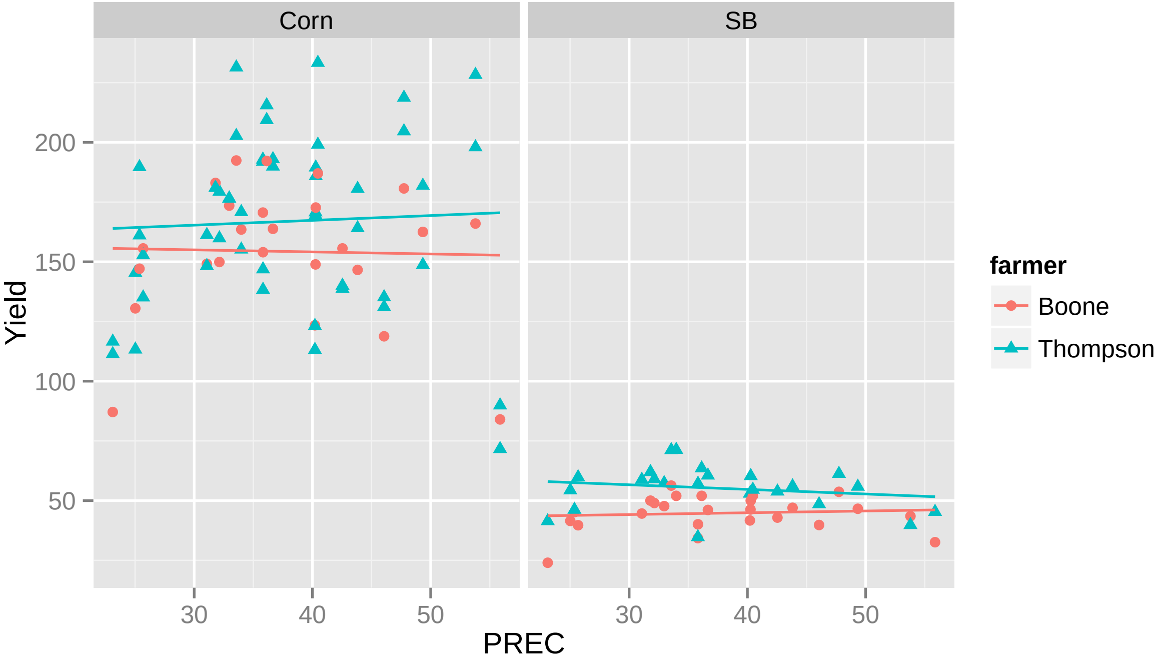
```r
wBoone <- weather %>%
  filter(stationName == "Boone") %>%
  gather(key, value, 5:373) %>%
  separate(key, into = c("year", "key"), sep = "\\_") %>%
  spread(key, value) %>%
  select(year, MAXT, MINT, PREC)

wBoone$year <- sapply(wBoone$year, FUN = function(x) extract_numeric(x))
yields2 <- inner_join(yields, wBoone, by = "year")
yields2$MAXT <- as.numeric(yields2$MAXT)
yields2$MINT <- as.numeric(yields2$MINT)
yields2$PREC <- as.numeric(yields2$PREC)
```

Examine the relationship between annual precipitation `PREC`, mean maximum temperature `MAXT` and mean minimum temperature `MINT`.

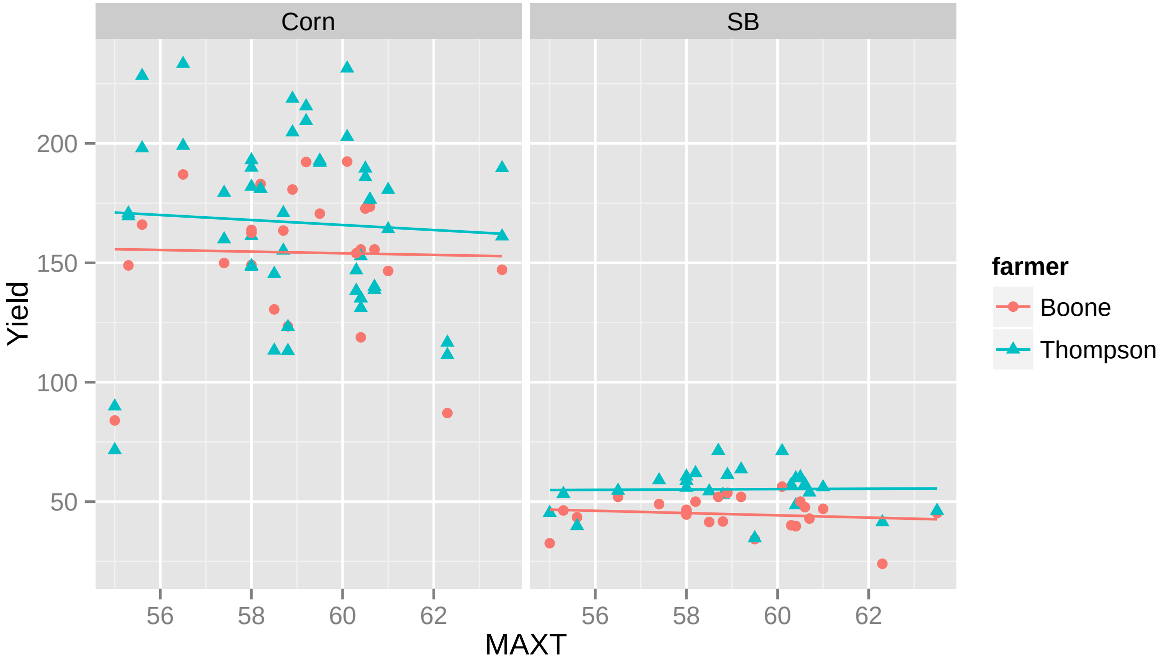
```r
head(yields2)
```

```
## Source: local data frame [6 x 8]
## Groups: field_id, year
##
##   year   farmer field_id crop  value MAXT MINT  PREC
## 1 1988    Boone        1 Corn  87.10 62.3 34.9 23.11
## 2 1988    Boone        2   SB  24.00 62.3 34.9 23.11
## 3 1988 Thompson        1 Corn 116.63 62.3 34.9 23.11
## 4 1988 Thompson        2   SB  41.45 62.3 34.9 23.11
## 5 1988 Thompson        4 Corn 111.39 62.3 34.9 23.11
## 6 1989     Boone        1 Corn 130.50 58.5 34.5 25.02
```

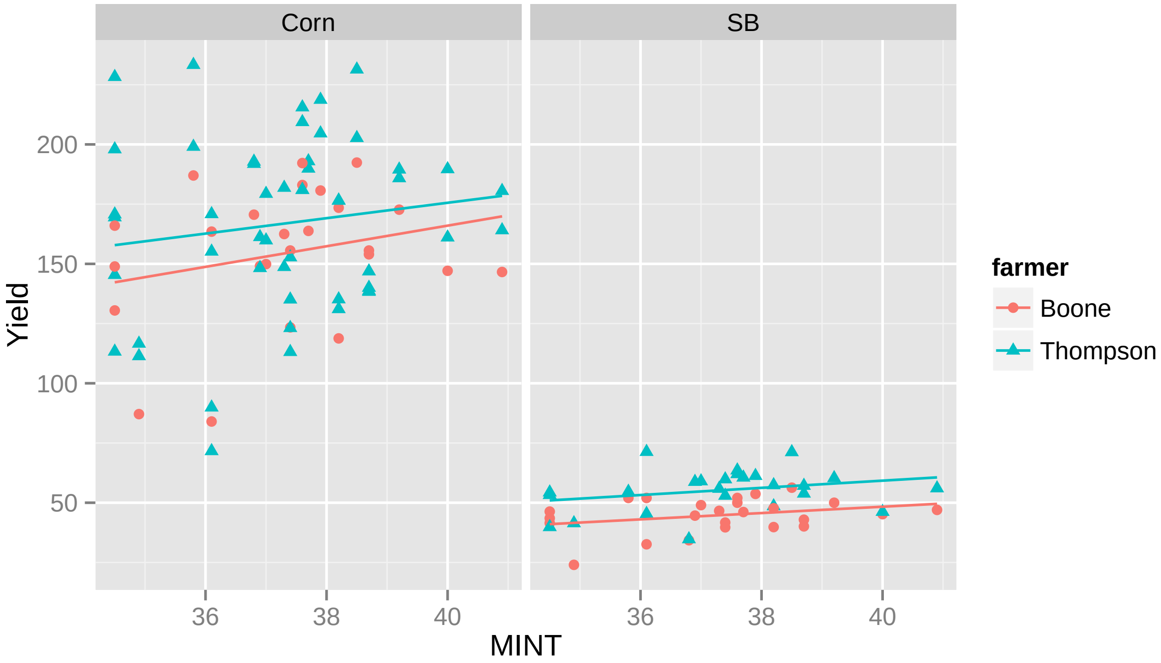
```
qplot(PREC, value, data=yields2, colour=farmer, shape=farmer,
      facets=~crop) + geom_smooth(method="lm", se=F) +
  ylab("Yield")
```



```
qplot(MAXT, value, data=yields2, colour=farmer, shape=farmer,
      facets=~crop) + geom_smooth(method="lm", se=F)+
  ylab("Yield")
```



```
qplot(MINT, value, data=yields2, colour=farmer, shape=farmer,
      facets=~crop) + geom_smooth(method="lm", se=F)+
  ylab("Yield")
```



Overall, pretty weak relationships between `Yield` for `Corn` and `SB` and `PREC`, `MAXT`, and `MINT`. Perhaps we need to look at monthly or seasonal weather data instead to see more of a trend.

Fields by year for Thompson that were used for `SB` or `corn`.

```
# pfi %>%
#   filter(item_type == "Unit Quantity", crop %in% c("Corn","SB"),
#          farmer == "Thompson") %>%
#   group_by(year, field_id) %>%
#   select(1:4)
```

Examine how farming practices relate to `Yield` by turning the `Expense` data into indicator variables:

- Each expense category is an indicator variable
- Variable is marked **TRUE** if the expense for that category for each particular **field.id** and **year** is greater than 0.
- **FALSE** if 0.

```
yields3 <- pfi %>%
  filter(item_type == "Expense", crop  %in% c("Corn", "SB")) %>%
  spread(item, value) %>%
  select(-c(5))
for (i in 5:36) {
  yields3[,i] <- yields3[,i] != 0
}
yields3$yield <- yields$value
yields3 <- yields3[,c(1:4,37,5:36)] # move 'yield' to 5th column
head(yields3)[1:6]
```

```
##   year   farmer field_id crop  yield Apply_NH4
## 1 1988    Boone        1 Corn  87.10      TRUE
## 2 1988    Boone        2   SB  24.00     FALSE
## 3 1988 Thompson        1 Corn 116.63     FALSE
## 4 1988 Thompson        2   SB  41.45     FALSE
## 5 1988 Thompson        4 Corn 111.39     FALSE
## 6 1989    Boone        1 Corn 130.50      TRUE
```

Eliminate variables that are all `TRUE` or all `FALSE`.

```
N <- nrow(yields3)
C <- ncol(yields3)
idx <- NULL
for (i in 6:C) {
  if ((sum(yields3[,i]) == N) | (sum(yields3[,i] == FALSE) == N)) {
    idx <- c(idx, i)
  }
}
yields3 <- yields3[-idx]
names(yields3)
```

```
##  [1] "year"             "farmer"           "field_id"
##  [4] "crop"             "yield"            "Apply_NH4"
##  [7] "Chop_StksCc"      "Corn_RSL"         "Cover_Crop"
## [10] "Crop_Ins"         "Cultivation"      "Drying_Cost"
## [13] "Fall_Tillage"     "Hedge_per_PL"     "Herbicides"
## [16] "Interest"           "Maunure_Charge"     "Mov_and_Stor_bales"
## [19] "Mow_per_Windrow"    "Purch_Pert"         "Rake"
## [22] "Rotary_Hoe"         "Shell_per_Grind"    "Spray_per_Walk"
## [25] "Spring_Tillage"     "Stack_Residues"     "Storage"
```

10 indicator variables eliminated so far. Break the dataset into yields for `Corn` and yields for `SB`. Elimate variables again that are all `TRUE` or all `FALSE`.

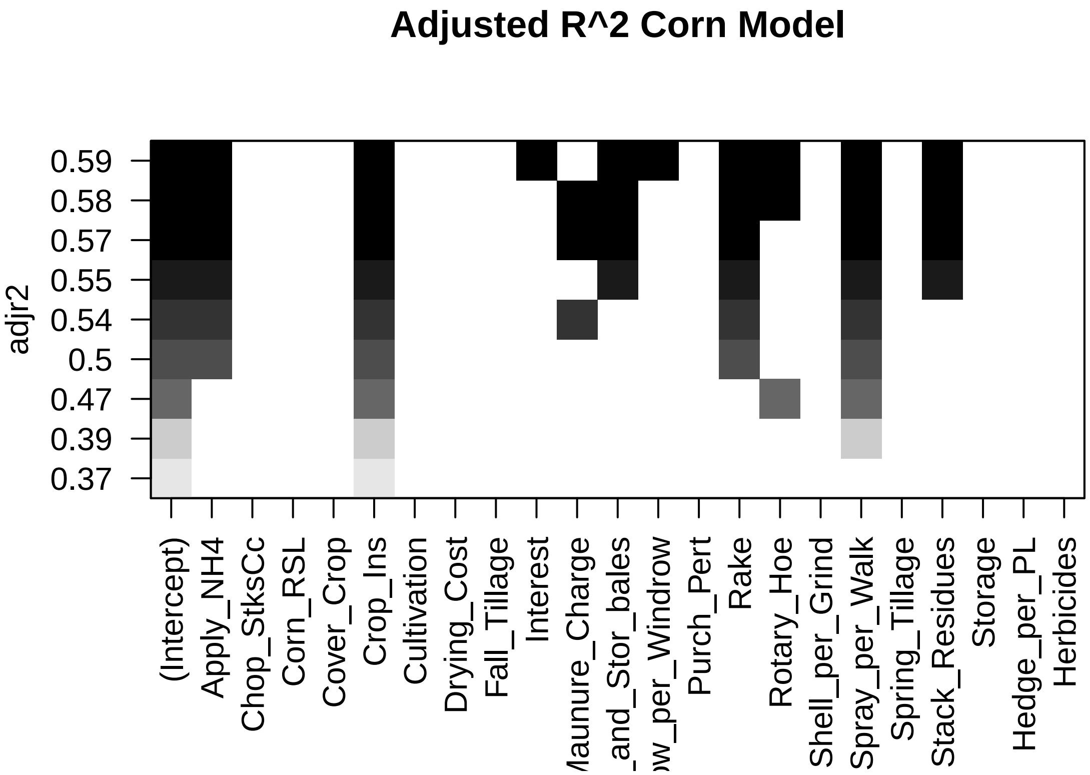
```
yields3_C <- subset(yields3, crop == "Corn")
yields3_SB <- subset(yields3, crop == "SB")

# yields3_C doesn't have any variables that are all T or all F

N <- nrow(yields3_SB)
C <- ncol(yields3_SB)
idx <- NULL
for (i in 6:C) {
  if ((sum(yields3_SB[,i]) == N) | (sum(yields3_SB[,i] == FALSE) == N)) {
    idx <- c(idx, i)
  }
}
yields3_SB <- yields3_SB[-idx]
```

Run regsubets to find best model for `Corn` yields.

```
regsubsets.out <- regsubsets(x = as.matrix(yields3_C[,6:27]),
                             y = yields3_C[,5])
```

```
## Reordering variables and trying again:
```

```
# Variables with black boxes at the highest y-axis label should be included
plot(regsubsets.out, scale = "adjr2", main = "Adjusted R^2 Corn Model")
```



```
summary(lm(data=yields3_C, yield~Apply_NH4+Crop_Ins+Interest+
             Mov_and_Stor_bales+Mow_per_Windrow+Rake+Rotary_Hoe+
             Spray_per_Walk+Stack_Residues))
```

```
##
## Call:
## lm(formula = yield ~ Apply_NH4 + Crop_Ins + Interest + Mov_and_Stor_bales +
##     Mow_per_Windrow + Rake + Rotary_Hoe + Spray_per_Walk + Stack_Residues,
##     data = yields3_C)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -71.83 -12.08  -0.58  14.00  50.68
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              204.63      17.19   11.91  < 2e-16 ***
## Apply_NH4TRUE            -51.89      14.39   -3.61  0.00060 ***
## Crop_InsTRUE             -41.95       9.97   -4.21  8.1e-05 ***
## InterestTRUE              23.66      10.89    2.17  0.03345 *
## Mov_and_Stor_balesTRUE    -8.60      10.48   -0.82  0.41500
## Mow_per_WindrowTRUE       33.49      10.67    3.14  0.00254 **
## RakeTRUE                  14.37      17.28    0.83  0.40877
## Rotary_HoeTRUE           -53.41      13.62   -3.92  0.00021 ***
## Spray_per_WalkTRUE        24.48       9.11    2.69  0.00914 **
## Stack_ResiduesTRUE        -2.07      17.72   -0.12  0.90744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.1 on 65 degrees of freedom
## Multiple R-squared:  0.601,  Adjusted R-squared:  0.545
## F-statistic: 10.9 on 9 and 65 DF,  p-value: 4.27e-10
```

Note that when NH4 is applied, the expected yield decreases by 51.892 when everything else is held constant. Why is crop insurance such a good predictor? Maybe the farmer buys crop insurance when they think it's going to be a bad year. Or it's just because Thompson stopped using crop insurance after 2002, and has had consistently higher yields than the Boone average.

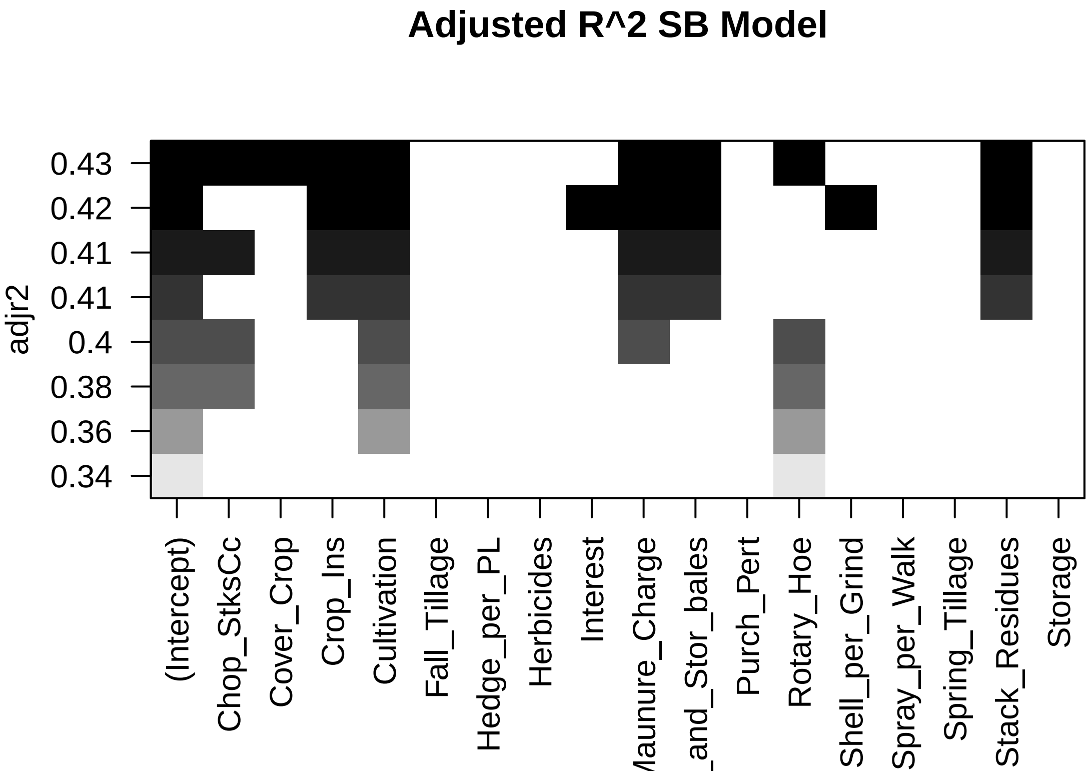
Here's a model that makes more "sense":

```
summary(lm(data=yields3_C, yield~Apply_NH4+
             Mow_per_Windrow+Rotary_Hoe+Spray_per_Walk))
```

```
##
## Call:
## lm(formula = yield ~ Apply_NH4 + Mow_per_Windrow + Rotary_Hoe +
##     Spray_per_Walk, data = yields3_C)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -71.94 -12.90  -1.65  15.85  58.46
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           191.55      13.68   14.00  < 2e-16 ***
## Apply_NH4TRUE         -59.39      15.09   -3.94  0.00019 ***
## Mow_per_WindrowTRUE    50.10       9.35    5.36  1.0e-06 ***
## Rotary_HoeTRUE        -66.78      14.92   -4.48  2.9e-05 ***
## Spray_per_WalkTRUE     26.89      10.11    2.66  0.00970 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.6 on 70 degrees of freedom
## Multiple R-squared:  0.428,  Adjusted R-squared:  0.396
## F-statistic: 13.1 on 4 and 70 DF,  p-value: 5.09e-08
```

Run regsubsets for best model for SB yields.

```
regsubsets.out2 <- regsubsets(x = as.matrix(yields3_SB[,6:22]),
                              y = yields3_SB[,5])
plot(regsubsets.out2, scale = "adjr2", main = "Adjusted R^2 SB Model")
```



```
summary(lm(data=yields3_SB, yield~Chop_StksCc+Cover_Crop+Crop_Ins+
             Cultivation+Maunure_Charge+Mov_and_Stor_bales+Rotary_Hoe+
             Stack_Residues))
```

```
##
## Call:
## lm(formula = yield ~ Chop_StksCc + Cover_Crop + Crop_Ins + Cultivation +
##     Maunure_Charge + Mov_and_Stor_bales + Rotary_Hoe + Stack_Residues,
##     data = yields3_SB)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18.55  -3.28   1.25   3.74  12.57
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)               40.16       5.29    7.59  2.5e-09 ***
## Chop_StksCcTRUE           -4.40       2.95   -1.49    0.144
## Cover_CropTRUE            -4.53       3.44   -1.32    0.195
## Crop_InsTRUE               9.81       4.55    2.16    0.037 *
## CultivationTRUE           -7.42       3.30   -2.25    0.030 *
## Maunure_ChargeTRUE         9.45       5.14    1.84    0.073 .
## Mov_and_Stor_balesTRUE    -8.49       4.93   -1.72    0.093 .
## Rotary_HoeTRUE             6.66       4.71    1.41    0.165
## Stack_ResiduesTRUE        10.62       6.56    1.62    0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.22 on 41 degrees of freedom
## Multiple R-squared:  0.521,  Adjusted R-squared:  0.428
## F-statistic: 5.58 on 8 and 41 DF,  p-value: 8.48e-05
```

Crop insurance again is found to be a significant predictor.

Here is a model that makes more “sense”:

```
summary(lm(data=yields3_SB, yield~Cultivation+Maunure_Charge+Rotary_Hoe))
```

```
##
## Call:
## lm(formula = yield ~ Cultivation + Maunure_Charge + Rotary_Hoe,
##     data = yields3_SB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.044  -2.465   0.589   4.472  15.506
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)           48.71       2.85   17.08   <2e-16 ***
## CultivationTRUE       -6.73       3.40   -1.98    0.054 .
## Maunure_ChargeTRUE     6.31       4.56    1.38    0.173
## Rotary_HoeTRUE         7.49       4.52    1.66    0.104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.54 on 46 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.376
## F-statistic: 10.8 on 3 and 46 DF,  p-value: 1.67e-05
```