# Global Crop Yield

Jo Hardin

9/1/2020

For a different project, I'm in need of a dataset to use to work through some inferential linear model ideas. Seems like maybe the crop yield dataset could be used?

Not sure how easy it is to create an inferential claim, but I'm going to try!

## The Data

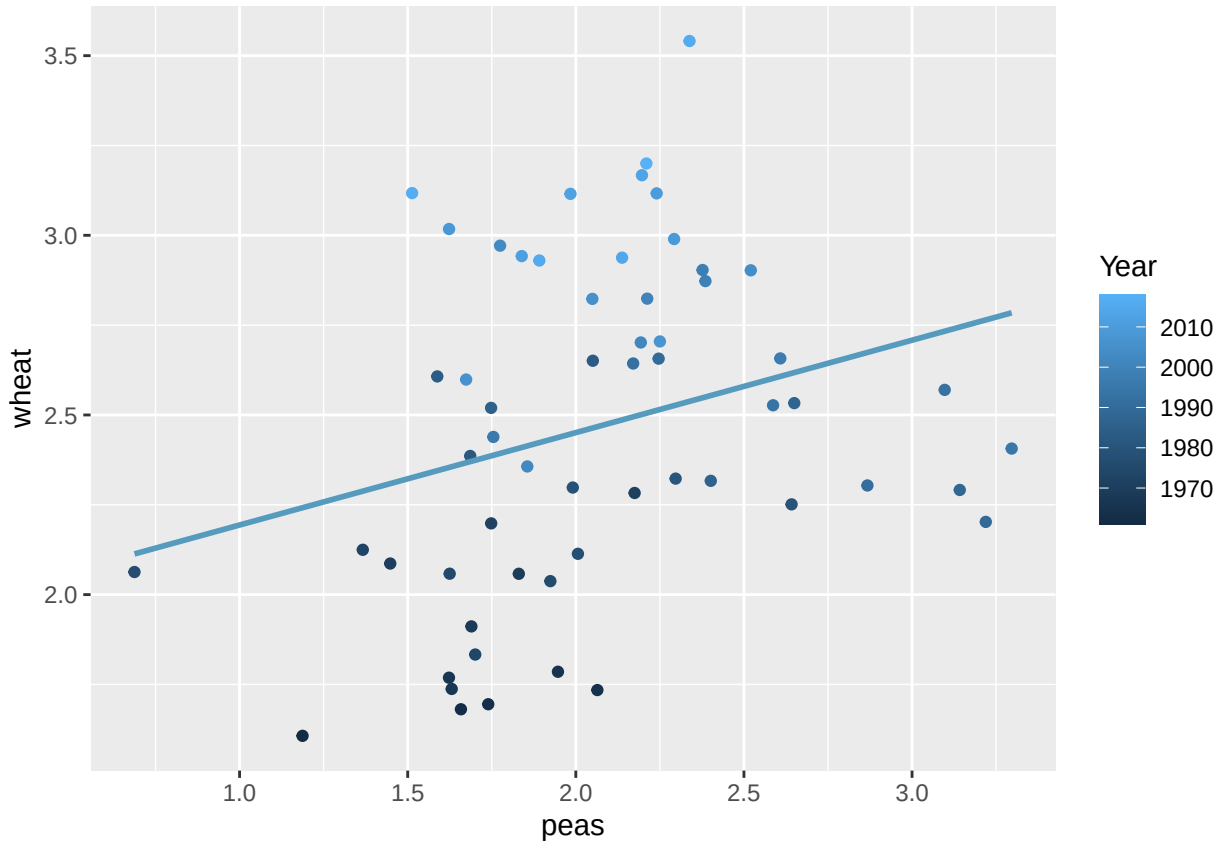
```
fertilizer <- readr::read_csv('cereal_crop_yield_vs_fertilizer_application.csv')
tractors <- readr::read_csv('cereal_yields_vs_tractor_inputs_in_agriculture.csv')
land_use <- readr::read_csv('land_use_vs_yield_change_in_cereal_production.csv')
arable_land <- readr::read_csv('arable_land_pin.csv')

key_crop_yields <- readr::read_csv('key_crop_yields.csv') %>%
  rename(wheat = `Wheat (tonnes per hectare)`,
         rice = `Rice (tonnes per hectare)`,
         maize = `Maize (tonnes per hectare)`,
         soybeans = `Soybeans (tonnes per hectare)`,
         potatoes = `Potatoes (tonnes per hectare)`,
         beans = `Beans (tonnes per hectare)`,
         peas = `Peas (tonnes per hectare)`,
         cassava = `Cassava (tonnes per hectare)`,
         barley = `Barley (tonnes per hectare)`,
         cocoa = `Cocoa beans (tonnes per hectare)`,
         bananas = `Bananas (tonnes per hectare)`)

crops_USA <- key_crop_yields %>%
  filter(Code == "USA")
```

## Goals for inference in linear regression:

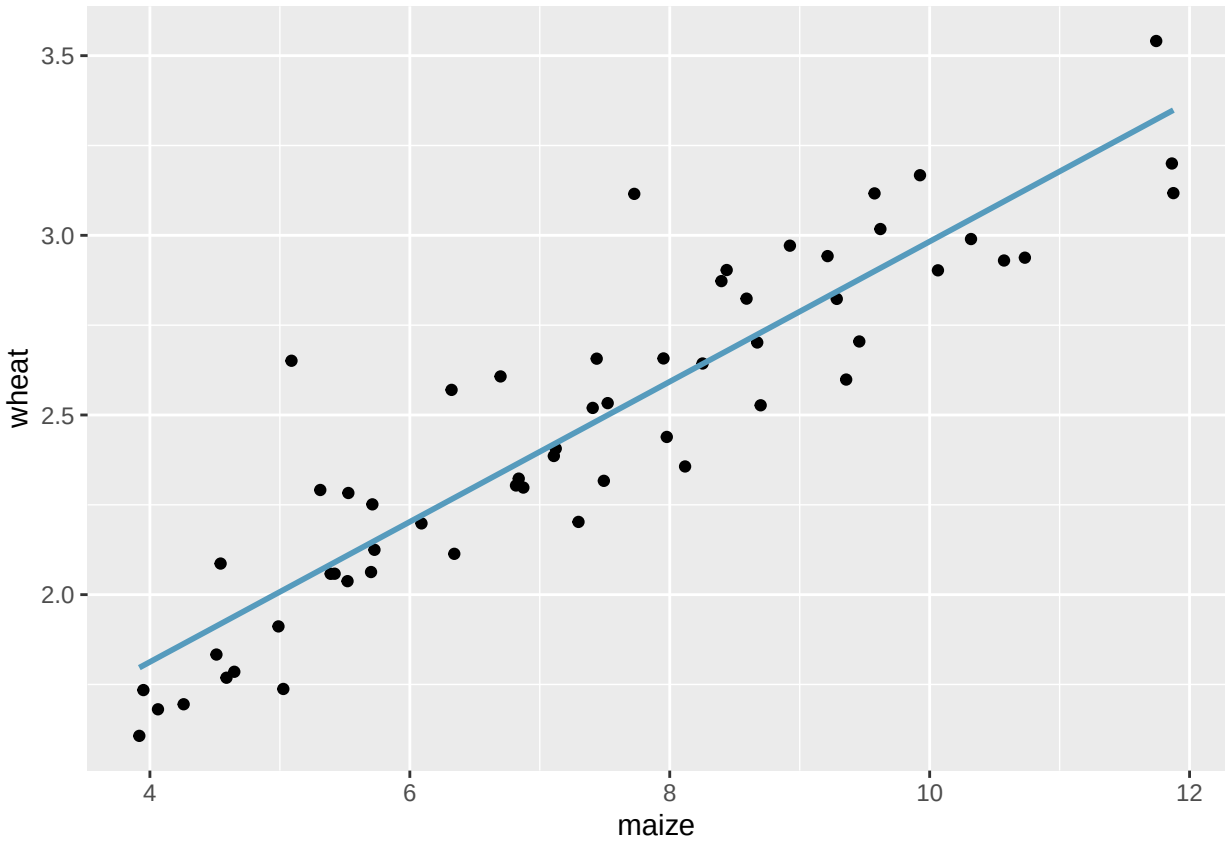
- make inferential claims about the model
- create confidence intervals for the slope

## EDA

```
ggplot(crops_USA) +
  geom_point(aes(x = peas, y = wheat, color = Year)) +
  geom_smooth(aes(x = peas, y = wheat), method = "lm", se = FALSE, color = COL[1,1])
```




```
ggplot(crops_USA) +
  geom_point(aes(x = maize, y = wheat)) +
  geom_smooth(aes(x = maize, y = wheat), method = "lm", se = FALSE, color = COL[1,1])
```




### SLR

```
crops_USA %>%
  lm(wheat ~ peas, .) %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.94      0.253      7.67 2.73e-10
## 2 peas           0.257     0.119      2.16 3.50e- 2
```

```
crops_USA %>%
  lm(wheat ~ maize, .) %>%
  tidy()
```

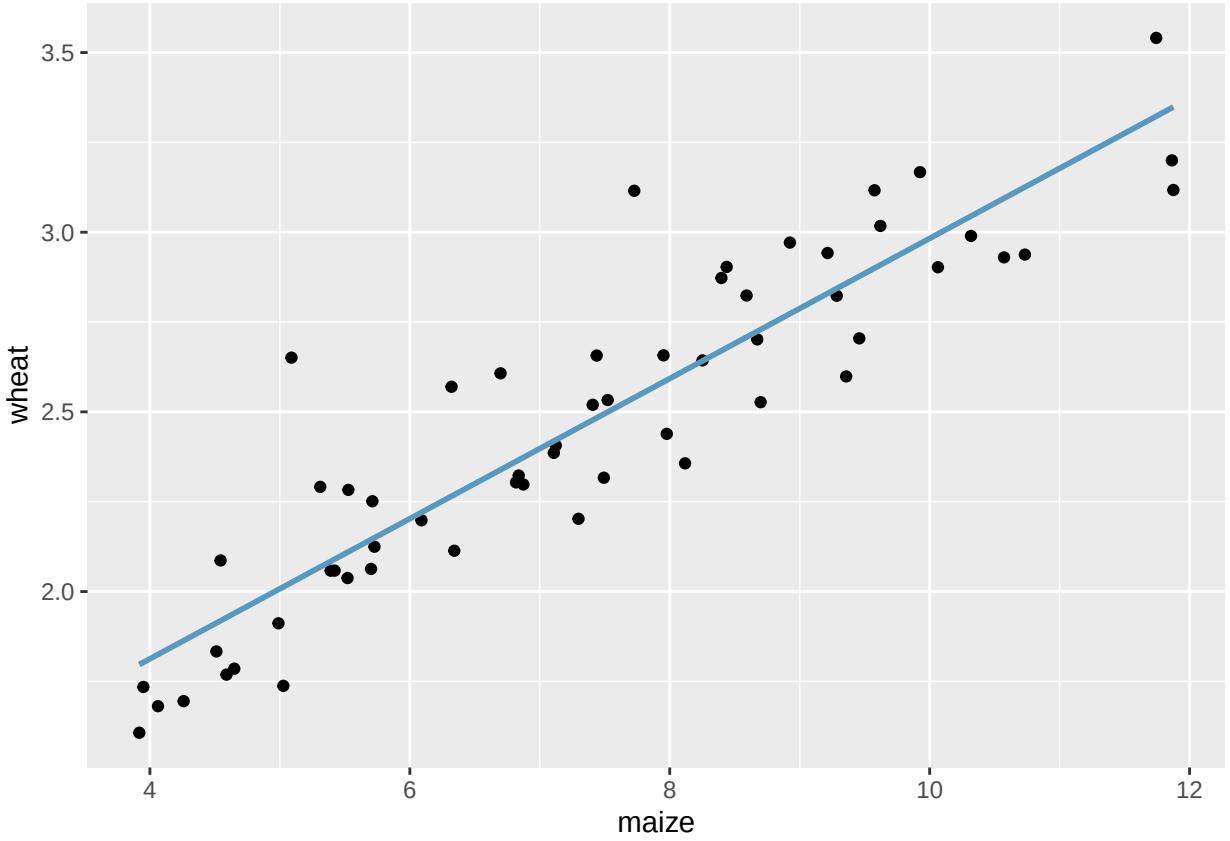
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.03     0.0912      11.3 4.13e-16
## 2 maize          0.195    0.0119      16.4 5.04e-23
```

### Sampling distribution

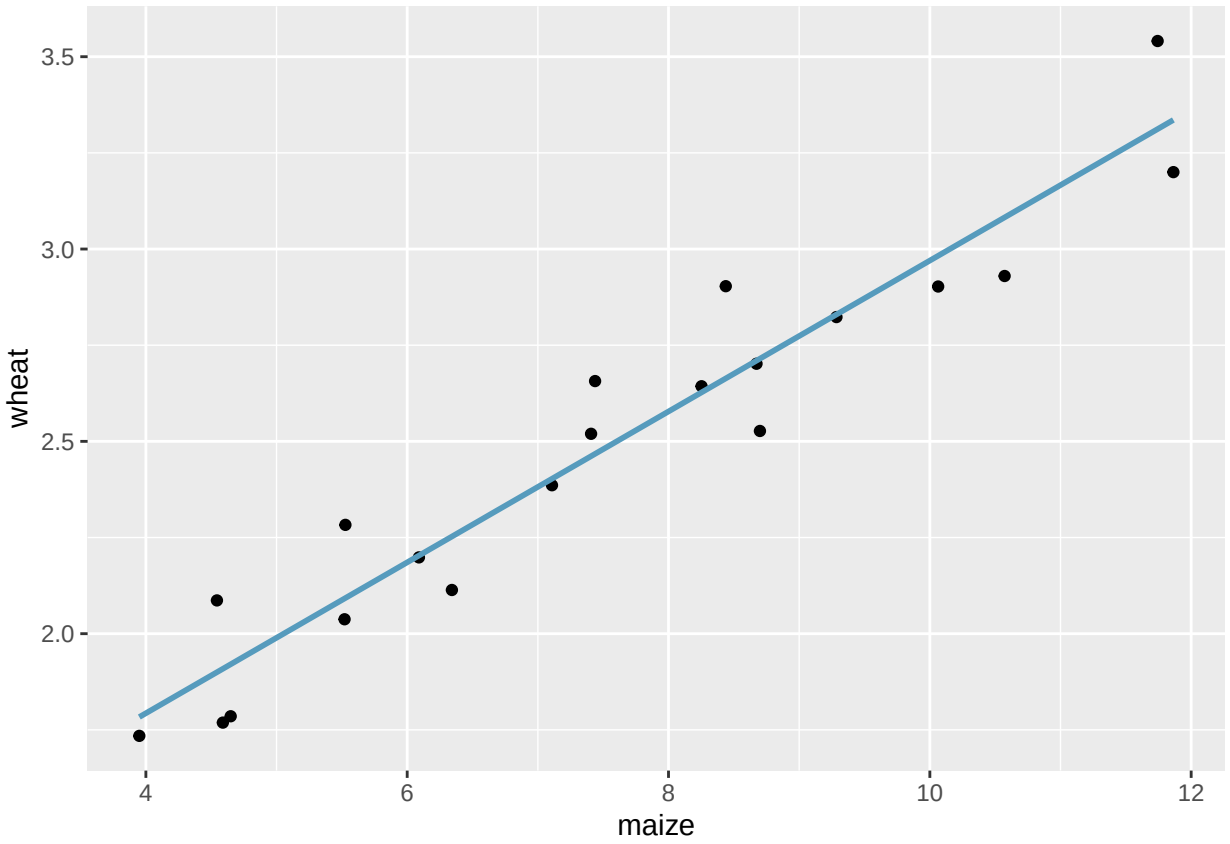
```
set.seed(4747)
crops2 <- crops_USA %>%
  sample_n(size=20)
```

```
crops3 <- crops_USA %>%
  sample_n(size=20)
crops_many <- crops_USA %>%
  rep_sample_n(size = 20, replace = FALSE, reps = 50)
```

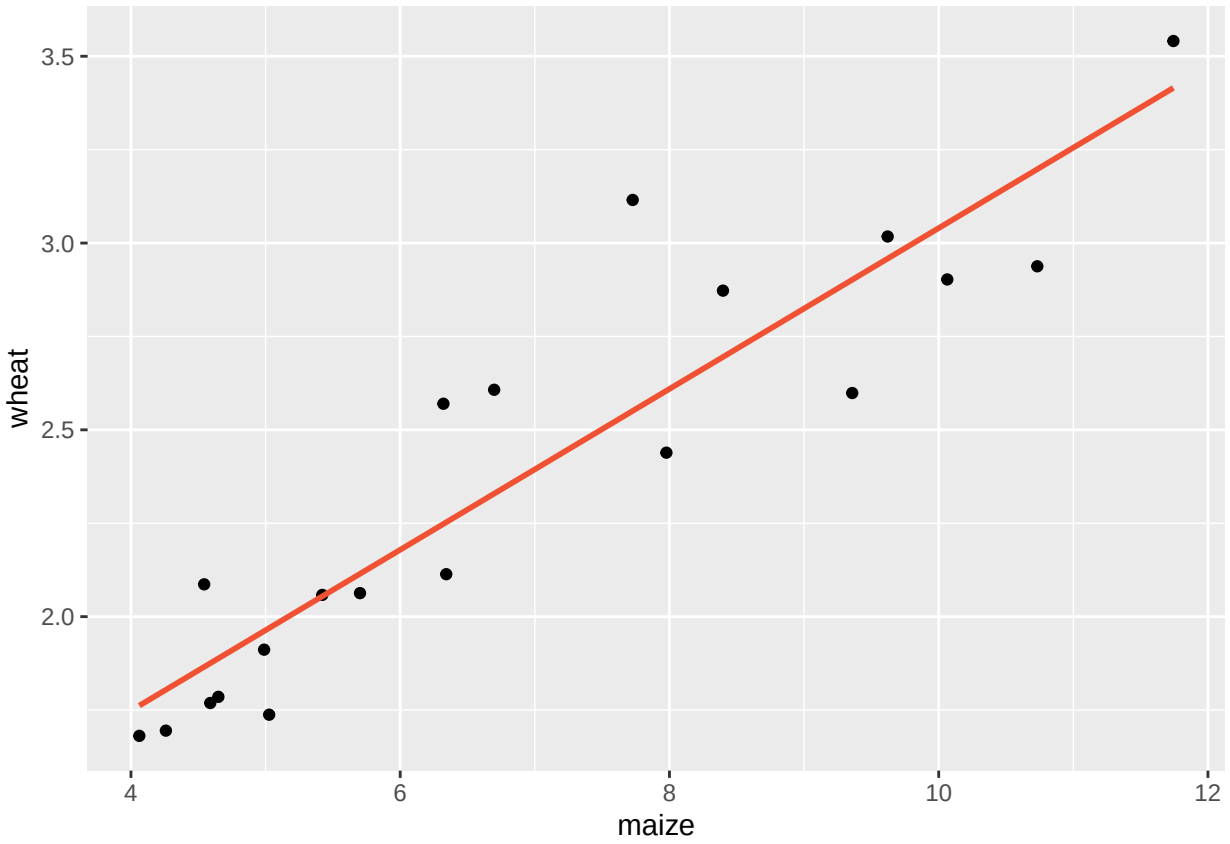
```
ggplot(crops_USA) +
  geom_point(aes(x = maize, y = wheat)) +
  geom_smooth(aes(x = maize, y = wheat), method = "lm", se = FALSE, color = COL[1,1])
```



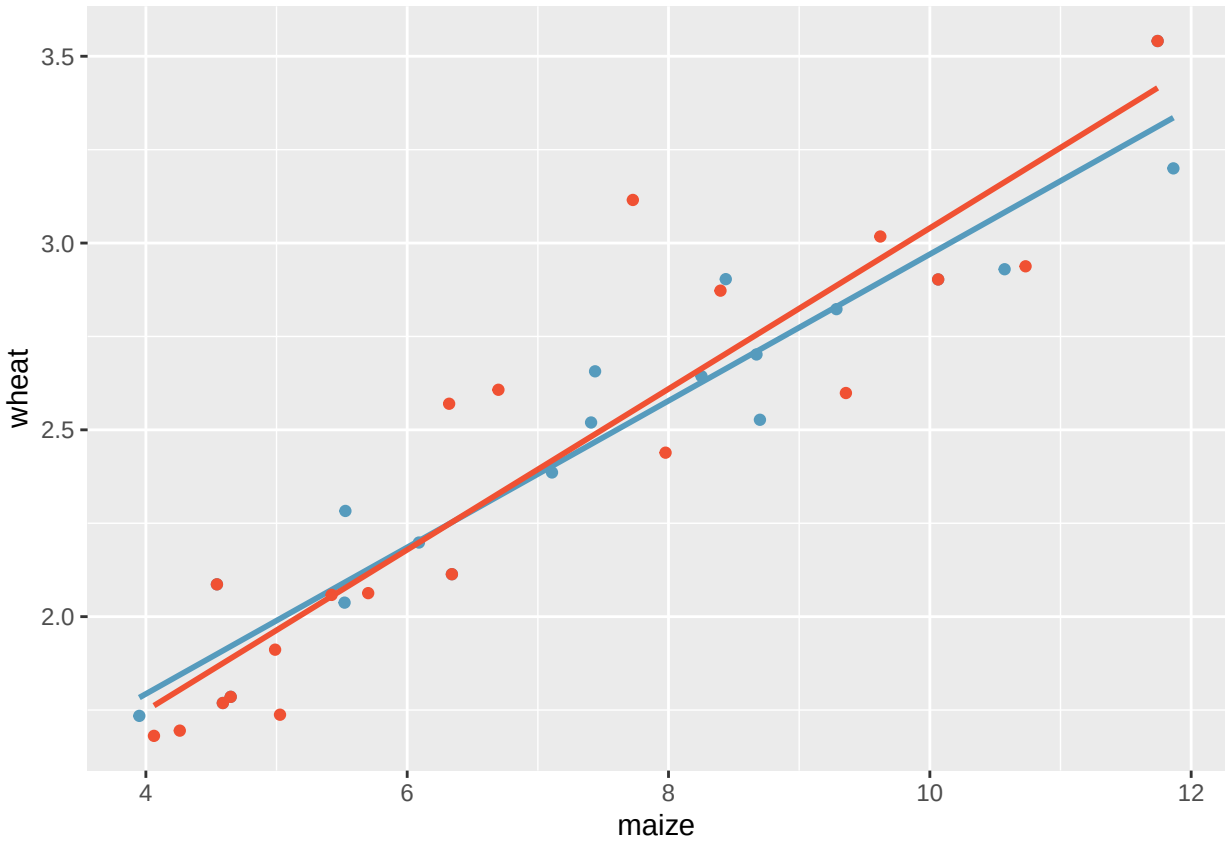

A subset of size 20 items shows a similar positive trend between maize and wheat, despite having fewer observations on the plot.



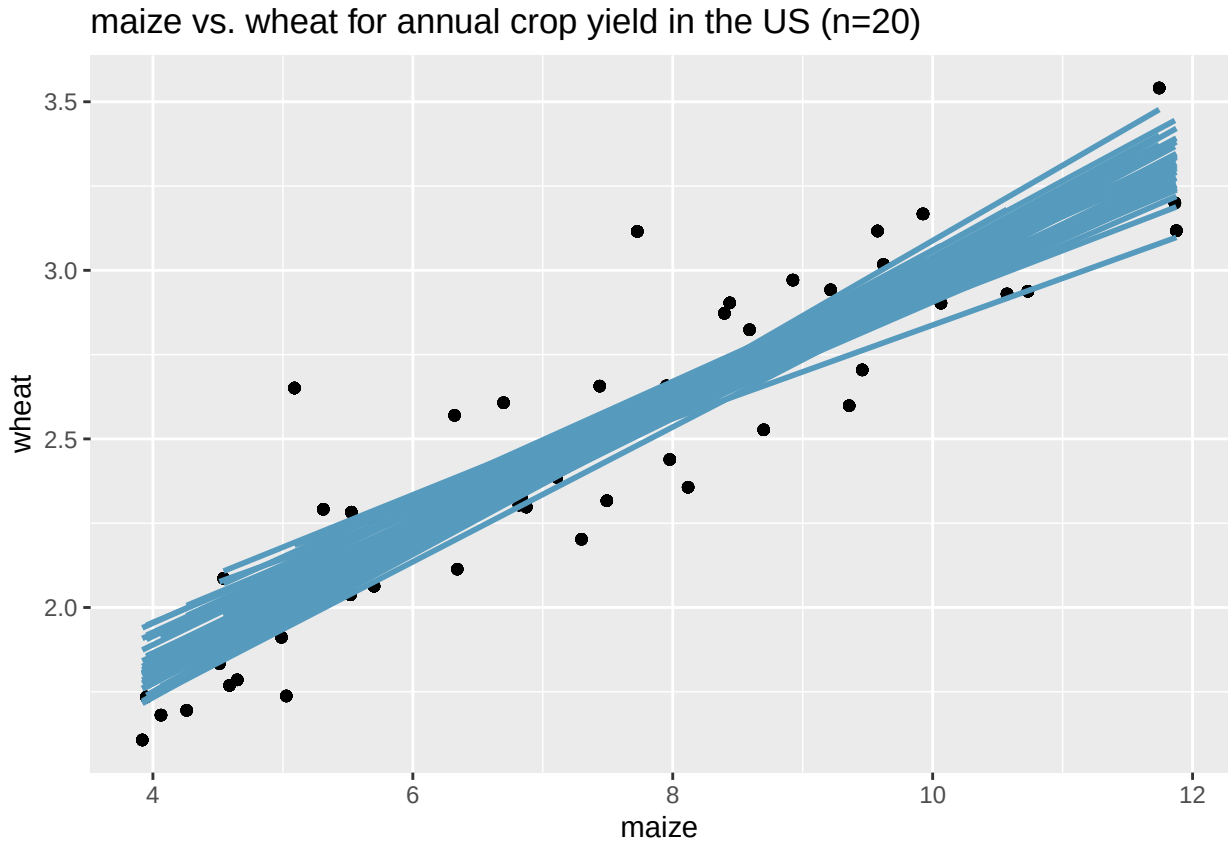

A second sample of size 20 also shows a positive trend!

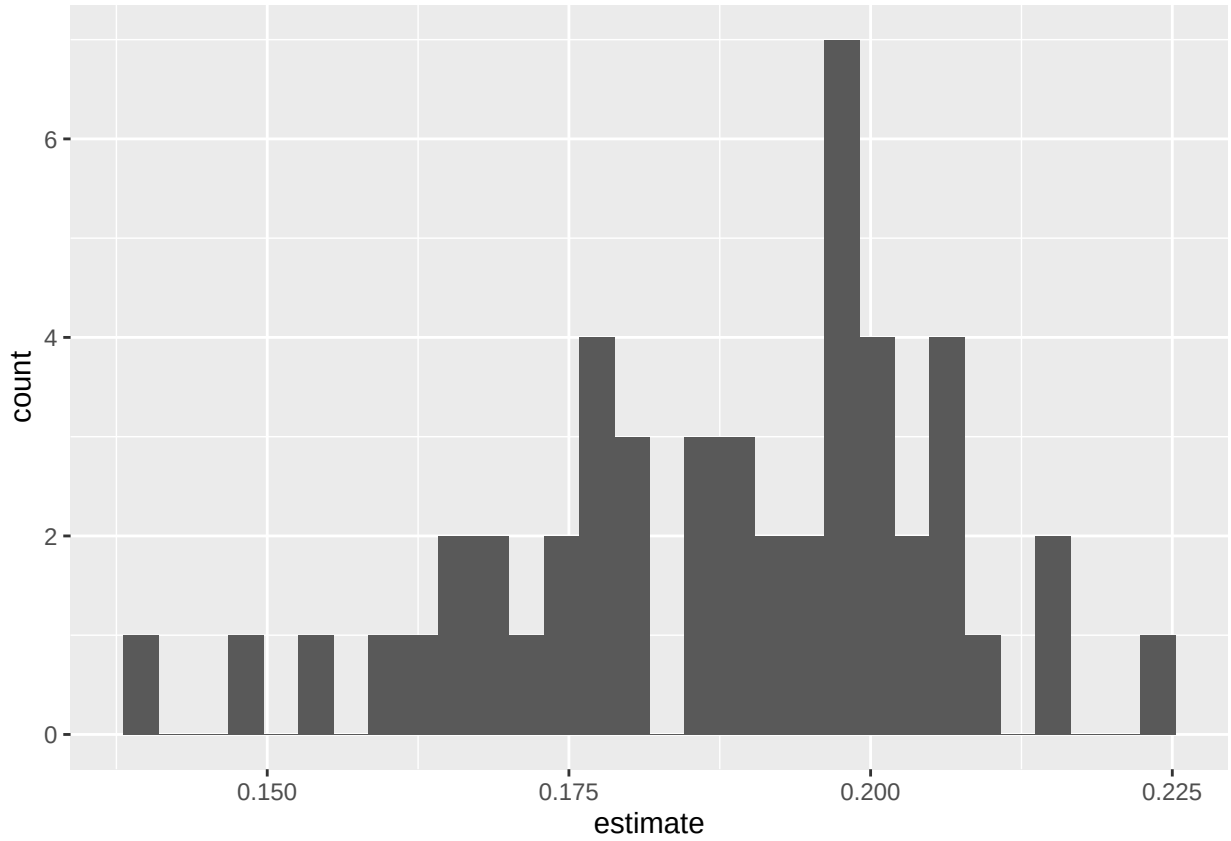



But the line is slightly different!




That is, there is variability in the regression line from sample to sample. The concept of the sampling variability is something you've seen before, but in this lesson, you will focus on the variability of the line instead of the variability of a single statistic.


maize vs. wheat for annual crop yield in the US (n=20)
wheat
3.5
3.0
2.5
2.0
4
6
8
10
12
maize


count
6
4
2
0
0.150
0.175
0.200
0.225
estimate