Understanding the bias/variance tradeoff in ridge regression

Matthew Richey/Jaime Davila

4/13/2021

Introduction

We are interested in generating simulated data where we get to see the advantages of the ridge model over a simple linear regression. The simplest simulation we can get is by generating N points $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where the x_i are generated from a normal distribution with mean 0 and standard deviation 1. We will be generating the y_i by multiplying the x_i by b and then adding an error term (which we call ϵ). Our epsilon will also have a normal distribution with mean 0 and standard deviation σ.

One of the instances where ridge regression can outperform linear regression is in the case where the number of points in our dataset is very small. So for the sake of our simulation let's set $N = 3$ (the number of observations), $b = 1$ (the slope) and $\sigma = 2$ (the standard deviation of our error term)

```
N <- 3
sig <- 2
b <- 1
```

And let's write a function `build_sim` which will create the simulated data according to our equation. We will be using this function many times later on, so we will add a column `id` that will allow us to label each different simulation

```
build_sim <- function(id){
  x1 <- rnorm(N,0,1)
  y <- b*x1+rnorm(N,0,sig)
  tibble(id,x1,y)
}
set.seed(12345)
(sim.tbl.1 <- build_sim(1))
```

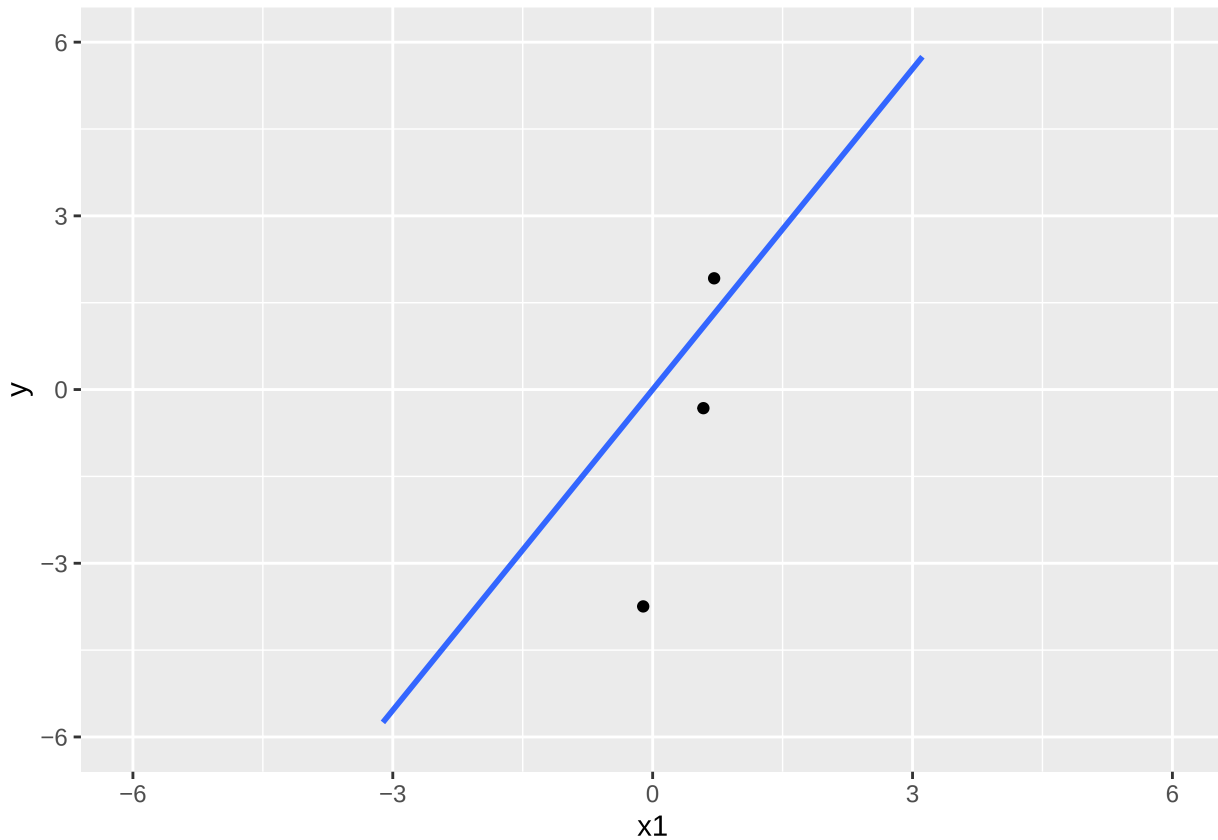
```
## # A tibble: 3 x 3
##      id     x1      y
##   <dbl>  <dbl>  <dbl>
## 1     1  0.586 -0.321
## 2     1  0.709  1.92
## 3     1 -0.109 -3.75
```

```
(sim.tbl.2 <- build_sim(2))
```

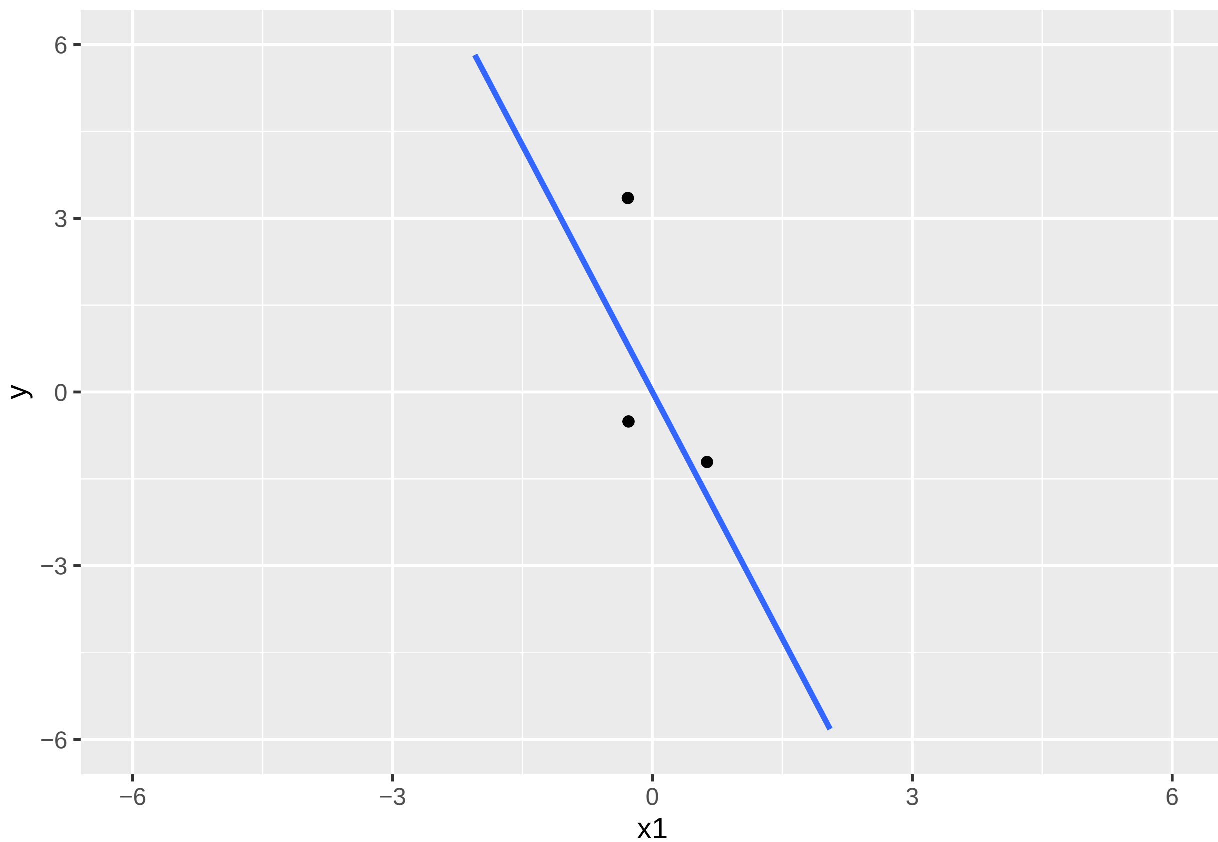
```
## # A tibble: 3 x 3
##      id     x1      y
##   <dbl>  <dbl>  <dbl>
## 1     2  0.630 -1.21
## 2     2 -0.276 -0.509
## 3     2 -0.284  3.35
```

And let's plot our two simulated tables and their linear trends.

```
ggplot(sim.tbl.1, aes(x1,y)) +
      geom_point() +
      xlim(-6,6)+
      ylim(-6,6)+
      geom_smooth(method=lm, formula = y~0+x,
                  se=FALSE, fullrange=TRUE)
```




```
ggplot(sim.tbl.2, aes(x1,y)) +
      geom_point() +
      xlim(-6,6)+
      ylim(-6,6)+
      geom_smooth(method=lm, formula = y~0+x,
                  se=FALSE, fullrange=TRUE)
```



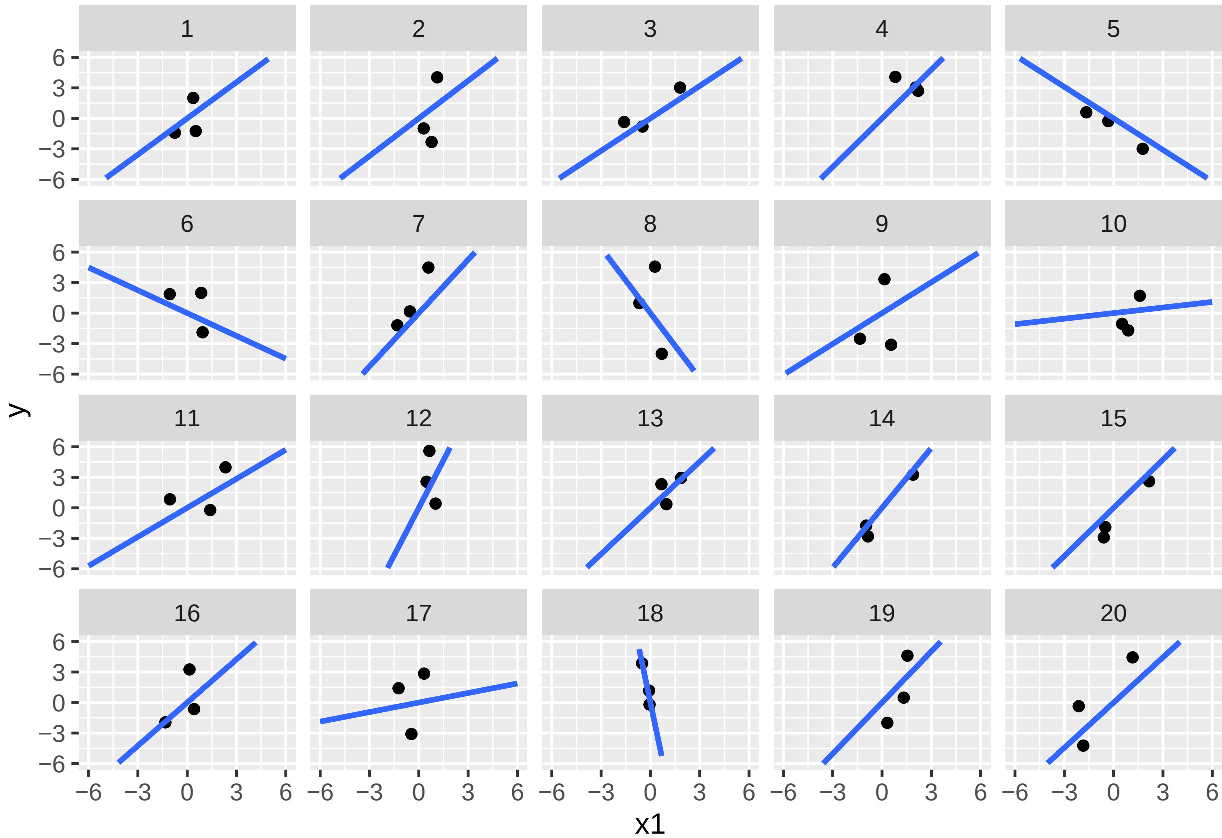
Notice how our trend lines look very different in these two instances. Notice that is partly due to the fact that we have a small number of points, so the slope estimates can change by a lot.

We would like to generate 20 simulation and plot all of them at the same time. In our first step we will create our 20 simulations by using the function `map_dfr()`. `map_dfr(x,f)` works by applying the function `f` for every element of `x` and then putting the results in dataframe. In particular it assumes that the results of `f(x)` are a dataframe (that's why the suffix `_dfr` in `map_dfr`)

```
(sim.tbl <- map_dfr(1:20, build_sim))
```

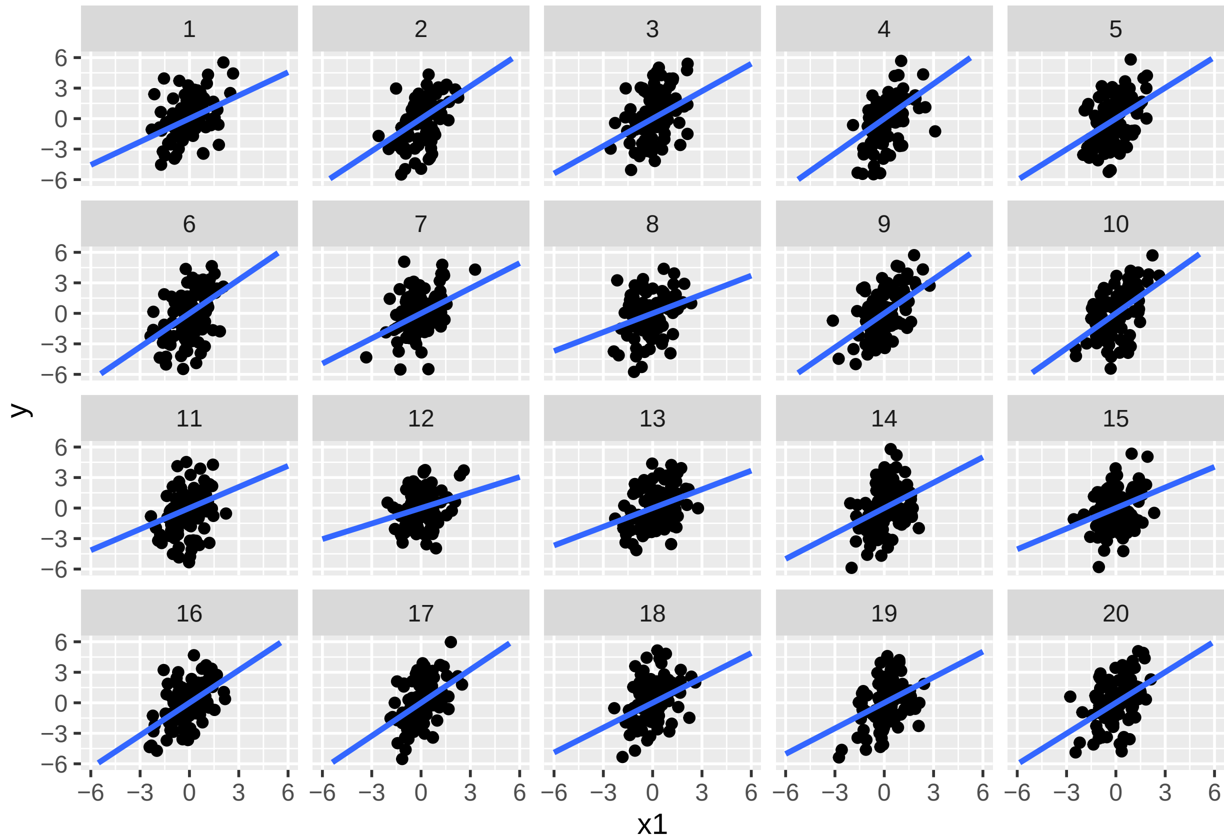
Finally we can show graphically the results of those simulations below. Notice how the slope varies according to each simulation

```
ggplot(sim.tbl,aes(x1,y)) +
      geom_point() +
      xlim(-6,6)+
      ylim(-6,6)+
      geom_smooth(method=lm, formula = y~0+x,
                  se=FALSE, fullrange=TRUE)+
      facet_wrap(vars(id))
```



Let's also point out that if the number of points N increases we get a more uniform set of models

```
N <- 100
map_dfr(1:20, build_sim)  %>%
  ggplot(aes(x1,y)) +
      geom_point() +
      xlim(-6,6)+
      ylim(-6,6)+
      geom_smooth(method=lm, formula = y~0+x,
                  se=FALSE, fullrange=TRUE)+
      facet_wrap(vars(id))
```



Before we continue our exercises, let's set up N equal to 3 again

```
N <- 3
```

1. Create a function `calc_slope_lm()` that creates a simulated dataset using the function `build_sim()` and returns a tibble with the `id` and the slope of the linear model with 0 intercept.

```
calc_slope_lm <- function(id) {
    tibble (id=id,x=slope)
}
```

2. Use `map_dfr` on `calc_slope_lm()` to create a table with the values of the slope for 100 simulations. Do a histogram of the slope and calculate its mean and standard deviation. We can define the bias as the mean of the slope minus the actual slope value (1). What is the bias of the simulated data?

Notice that the bias is small, however there is a quite a bit of variation.

3. Create a new simulated dataset by setting up $N = 10$ and $N = 100$. How does the slope histogram changed when compared with exercise 2?

Comparing linear regression and ridge

We would like to compare the estimated slope that we would get from our dataset using both ridge and linear models. In the `glmnet` implementation the ridge model needs to have at least two explanatory variables x_1 and x_2 so we will be modifying our simulation function to create an output of the form $b \cdot x_1 + b \cdot x_2 + \epsilon$. Our first step is to create a `build_sim2d` function as below

```
N <- 3
build_sim2d <- function(id){
  x1 <- rnorm(N,0,1)
  x2 <- rnorm(N,0,1)
  y <- b*x2+b*x1+rnorm(N,0,sig)

  tibble(id,x1,x2,y)
}
set.seed(123)
build_sim2d(1)
```

```
## # A tibble: 3 x 4
##      id     x1     x2      y
##   <dbl>  <dbl>  <dbl>  <dbl>
## 1     1 -0.560 0.0705  0.432
## 2     1 -0.230 0.129  -2.63
## 3     1  1.56  1.72    1.90
```

And we will be creating a `calc_coefs` function that outputs the estimates of the coefficients for a given model.

```
calc_coefs <- function(id, model) {
  # We generate a simulated dataset
  sim.tbl <- build_sim2d(id)

  # We create a workflow for our model and fit it on the simulated data
  recipe <- recipe(y ~ 0+x1+x2, data=sim.tbl) %>%
    step_normalize(all_predictors())
  wflow <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(model)
  fit <- fit(wflow, sim.tbl)

  # We pull the coefficients from x1 and x2 and put them in a tibble
  x1 <- tidy(fit) %>%
    filter(term=="x1") %>%
    pull(estimate)
  x2 <- tidy(fit) %>%
    filter(term=="x2") %>%
    pull(estimate)
  tibble (id=id,x1=x1, x2=x2)
}
```

Notice that using `calc_coefs` we can calculate the coefficients for a linear model as follows

```
lm.model <- linear_reg() %>%
    set_engine("lm")

calc_coefs(1,lm.model)
```

```
## # A tibble: 1 x 3
##      id    x1    x2
##   <dbl> <dbl> <dbl>
## 1     1 0.658  4.25
```

Or we can calculate the coefficients for a ridge model with penalty equal to 1.

```
ridge.model <-
  linear_reg(mixture = 0, penalty=1) %>%
  set_mode("regression") %>%
  set_engine("glmnet")

calc_coefs(1,ridge.model)
```

```
## # A tibble: 1 x 3
##       id     x1    x2
##    <dbl>  <dbl> <dbl>
## 1     1 -0.693 0.232
```

3. Use the functions `map_dfr` and `calc_coefs` to obtain the coefficients `x1` and `x2` from the linear model and from the ridge model with penalty 1 for 100 simulations. Do a histogram of the values of the coefficients for each type of model. Calculate the mean, bias and standard deviation. How do those values compare across models?

4. Use penalty values of $0.1, 1, 10, 100$ and see what is the effect on the mean and standard deviation of the coefficients for x1 and x2 when using a ridge model.

The Bias-Variance Trade-off

Imagine a fixed prediction value $x_0 := (x_{0,1}, x_{0,2})$\$ and an observed value $y_0 = f(x_0) + \epsilon$. And let $\hat{f}(x)$ be that value that our model predicts.

The bias-variance trade-off equation says:

$$E[(y_0 - \hat{f}(x_0))^2] = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

Let's look closely at each of these components.

- $E[(y_0 - \hat{f}(x_0)^2]$ is the expected (average) value $(y_0 - \hat{f}(x_0))^2$. This a gauge of how close our prediction comes to the actual value. We also know this as the **Mean Squared Error (MSE)**.
- $\text{Var}(\hat{f}(x_0))$ is the variance of the values $\hat{f}(x_0)$ generated from each training set. This is a gauge of how "wiggly" the prediction is as we build $\hat{f}(x)$ with different training data sets.
- $\text{Bias}(\hat{f}(x_0))$ is the expected value of $\hat{f}(x_0) - f(x_0)$ over the training data, i.e., $\text{Bias}(\hat{f}(x_0)) = E[\hat{f}(x_0) - f(x_0)]$. This is a gauge how the predicted values misses the actual value. Note that the bias could be zero but we still miss by a lot!
- σ^2 is the "noise'.The noise reflects all the influences that our $\hat{f}(x)$ misses. Generally, we assume that the noise has mean zero and $\text{Var}(\epsilon) = \sigma^2$.

Let's create a function `calc_errors` that calculates the mse, variance and bias for a given model.

```
calc_errors <- function(id, model) {
  # Simulate our testing/training dataset
  train.tbl <- build_sim2d(id)
  test.tbl <- build_sim2d(id)

  # Create a workflow and fit using your training
  recipe <- recipe(y ~ 0+x1+x2, data=train.tbl)
  wflow <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(model)
  fit <- fit(wflow, train.tbl)
```

```r
  # Calculate the predictions on the testing dataset
  predict.tbl <- augment(fit, test.tbl)

  # Calculate mse, var, bias and put everything on a tibble
  mse = mean((predict.tbl$y-predict.tbl$.pred)^2)
  var = mean(predict.tbl$.pred^2)
  bias = mean(predict.tbl$y-predict.tbl$.pred)^2
  tibble(mse=mse, var=var, bias=bias)
}
calc_errors(1,lm.model)
```

Let's calculate those parameters on the linear model

```r
calc_errors(1,lm.model)
```

```
## # A tibble: 1 x 3
##     mse   var  bias
##   <dbl> <dbl> <dbl>
## 1  2.65  3.22  1.99
```

```r
set.seed(12345)
errors.lm <- map_dfr(1:20, calc_errors, lm.model)

errors.lm %>%
  pivot_longer(1:3) %>%
  group_by(name) %>%
  summarize(mean = mean(value))
```

```
## # A tibble: 3 x 2
##   name   mean
##   <chr> <dbl>
## 1 bias   13.4
## 2 mse   118.
## 3 var   106.
```

And let's calculate those values on different ridge models with different penalties.

```r
for (penalty in c(0.1,1,10,100,1000)){
  ridge.model <-
    linear_reg(mixture = 0, penalty=penalty) %>%
    set_mode("regression") %>%
    set_engine("glmnet")

  set.seed(12345)
  errors.ridge <- map_dfr(1:20, calc_errors, ridge.model)
  errors.tbl <- errors.ridge %>%
    pivot_longer(1:3) %>%
    group_by(name) %>%
    summarize(mean = mean(value))

  print(penalty)
  print(errors.tbl)
}
```

```
## [1] 0.1
## # A tibble: 3 x 2
```

```
##   name   mean
##   <chr> <dbl>
## 1 bias   7.94
## 2 mse   20.7
## 3 var   17.8
## [1] 1
## # A tibble: 3 x 2
##   name   mean
##   <chr> <dbl>
## 1 bias   4.53
## 2 mse    9.93
## 3 var    6.39
## [1] 10
## # A tibble: 3 x 2
##   name   mean
##   <chr> <dbl>
## 1 bias   2.99
## 2 mse    6.71
## 3 var    2.08
## [1] 100
## # A tibble: 3 x 2
##   name   mean
##   <chr> <dbl>
## 1 bias   3.12
## 2 mse    6.69
## 3 var    1.88
## [1] 1000
## # A tibble: 3 x 2
##   name   mean
##   <chr> <dbl>
## 1 bias   3.16
## 2 mse    6.72
## 3 var    1.89
```

Notice how although ridge it has more bias than linear regression, it also has less variation, which results in a smaller overall mse.