WU #19 - Bias-Var Tradeoff II

Math 158 - Jo Hardin

Thursday 4/7/2022

Name: _____

Names of people you worked with: _____

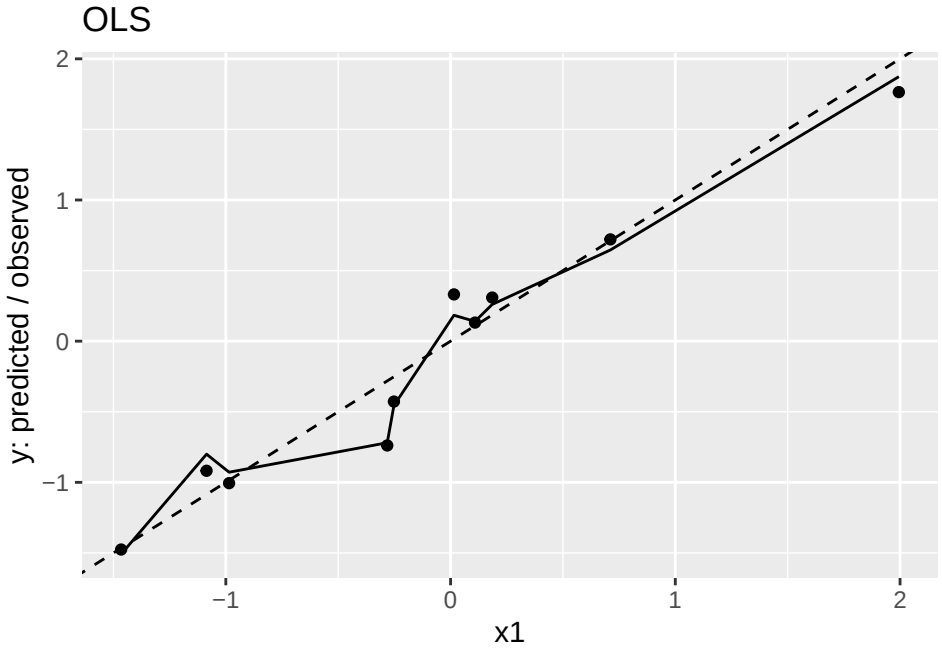
Consider an artificial data set comprising of ten observations on a response Y_i and eight covariates $X_{i,j}$. All covariate data are sampled from the standard normal distribution: $X_{i,j} \sim N(0,1)$. The response is generated by $Y_i = X_{i,1} + \epsilon_i$ with $\epsilon_i \sim N(0, 0.25)$. Hence, only the first covariate contributes to the response.

The regression model is fit to the artificial data using R.

- Which plot shows high variance and low bias? Explain.
- Which plot shows low variance and high bias? Explain.

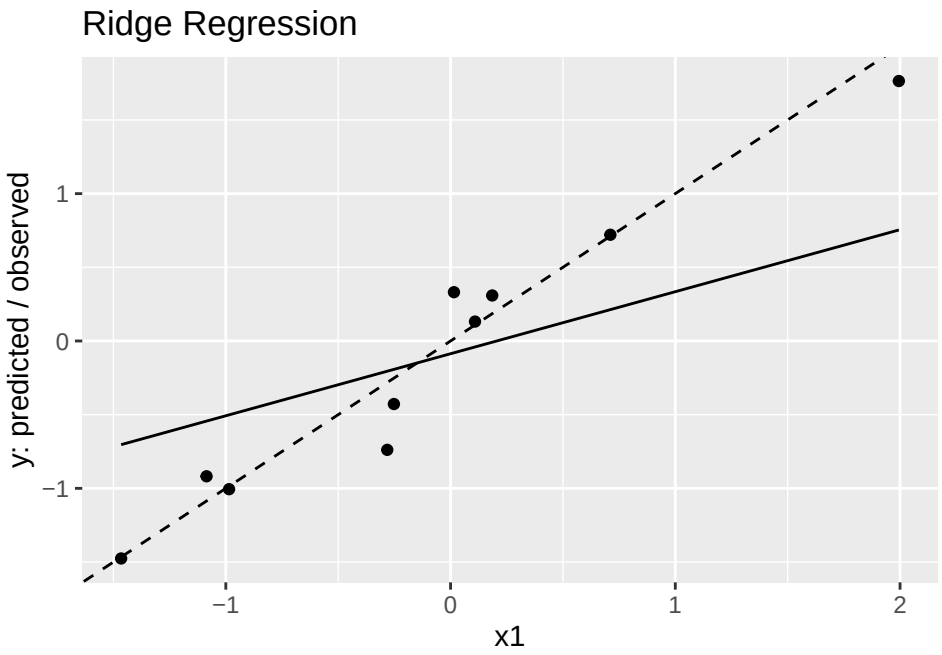
```
artif_lm <- data %>% lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = .)

artif_lm %>% tidy() %>% select(estimate) %>% pull()
```

```
## [1] -0.16623802  0.65929073 -0.17461427  0.06240072 -0.34454061
## [6]  0.12846580 -0.04326761  0.13458369  0.16035925
```



```r
artif_rec <- recipe(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = data) %>%
  step_normalize(all_numeric(), -all_outcomes())

ridge_spec <- linear_reg(mixture = 1, penalty = 0.5) %>%
  set_mode("regression") %>%
  set_engine("glmnet")

ridge_wf <- workflow() %>%
  add_recipe(artif_rec)

ridge_fit <- ridge_wf %>%
  add_model(ridge_spec) %>%
  fit(data = data)

ridge_fit %>% tidy() %>% select(estimate) %>% pull()
```

```
## [1] -0.1309584  0.4177448  0.0000000  0.0000000  0.0000000
## [6]  0.0000000  0.0000000  0.0000000  0.0000000
```



- The OLS plot shows low bias (i.e., it gets all the structure of the data) but high variance (that model would not be produced with a new dataset!).
- The ridge regression plot shows high bias (i.e., it gets none of the structure of the data) but low variance (the simple model could be reproduced in a different dataset).