HW1 Part 2

5

a) Function to simulate y values from linear model

```
gety <- function(x) {
  y <- 2*x^2 + rnorm(length(x),0,2)
  return(y)
}
```

b) Simulation

```
##run simulation 100000 times
reps <- 100000

##create an array to store the estimated slope from each rep
store.y0 <- array(0, reps)
store.f_hat0 <- array(0, reps)

##time loop
start_time <- Sys.time() ##start time of loop

set.seed(4630)

for (i in 1:reps)

{
  ##generate the values of x
  x <- rep(seq(1,10,1),20)

  ##simulate values of y
  y <- gety(x)

  ##use least squares to obtain regression equation on simulated data
  result <- lm(y~x)

  ##store the important values from this rep
  store.y0[i] <- gety(c(7)) #y0
  store.f_hat0[i] <- predict(result, data.frame(x=c(7))) #y_hat0

  #track progress
  if(i %% 10000 == 0) print(paste("Iteration", i))
}
```

```
## [1] "Iteration 10000"
## [1] "Iteration 20000"
## [1] "Iteration 30000"
## [1] "Iteration 40000"
## [1] "Iteration 50000"
## [1] "Iteration 60000"
## [1] "Iteration 70000"
## [1] "Iteration 80000"
## [1] "Iteration 90000"
## [1] "Iteration 100000"
```

```
end_time <- Sys.time() ## end time of loop
end_time - start_time ##time taken by loop
```

```
## Time difference of 1.847002 mins
```

c) Calculate the expected test MSE at x0 = 7

```
exp_test_MSE <- (1/reps)*sum((store.y0-store.f_hat0)^2)
exp_test_MSE
```

```
## [1] 148.1069
```

d) Calculate f_bar(x0) at x0 = 7

```
f_bar_x0 <- mean(store.f_hat0)
f_bar_x0
```

```
## [1] 109.9993
```

e) Using your 100,000 values of y0 and f_hat(x0), calculate each of the three sources of error, report each of their values, and then add them up (and report the value when added up).

```
var_f_hat <- mean((store.f_hat0-f_bar_x0)^2)
var_f_hat
```

```
## [1] 0.02531016
```

```
bias_squared <- (98-f_bar_x0)^2
bias_squared
```

```
## [1] 143.9826
```

```
var_e <- mean((store.y0-98)^2)
var_e
```

```
## [1] 4.015478
```

```
var_f_hat + bias_squared + var_e
```

```
## [1] 148.0234
```

f) The third source of error, E[(y0 - f(x0))^2], should be close to 4 (it should be 4, theoretically). In one sentence, briefly explain why.

It should be 4 because that is the variance of the standard normal distribution from which we generated the error terms in the line: rnorm(length(x),0,2).

g) Based on your values from 5c and 5e, what is the difference between the LHS and RHS of (1)?

```
exp_test_MSE - (var_f_hat + bias_squared + var_e)
```

```
## [1] 0.08356216
```

h) Be sure to include the names of classmates you worked with on this question.

Grady Bartro