Weekly Assignment 01

Xiang Li

2024/2/12

Question 1

```r
set.seed(519)
gen_sample_set = function(n) {
    x = runif(n, min = -3, max = 3)
    epsilon = rnorm(n, mean = 0, sd = 1)
    y = 8 * sin(x) + epsilon
    data = data.frame(x = x, y = y)
    return(data)
}
test_set = gen_sample_set(n = 10000)
cal_mse = function(n_train, test, degree) {
    n_reps = 100
    mse_v = replicate(n_reps, 0)
    y_hat_mat = matrix(0, nrow = n_reps, ncol = nrow(test))
    rownames(y_hat_mat) = 1:n_reps
    colnames(y_hat_mat) = 1:nrow(test)
    for (i in 1:n_reps) {
        train = gen_sample_set(n = n_train)
        model = lm(y ~ poly(x, degree), train)
        y_hat = predict(model, newdata = test)
        y_hat_mat[i, ] = y_hat
        mse_v[i] = mean((test$y - y_hat)^2)
    }
    mse = mean(mse_v)
    bias = mean((apply(y_hat_mat, 2, mean) - test$y)^2)
    var_ = mean(apply(y_hat_mat, 2, var))
    return(list(mse = mse, bias = bias, var = var_))
}
re_50_3 = cal_mse(n_train = 50, test = test_set, degree = 3)
re_50_15 = cal_mse(n_train = 50, test = test_set, degree = 15)
re_1w_3 = cal_mse(n_train = 10000, test = test_set, degree = 3)
re_1w_15 = cal_mse(n_train = 10000, test = test_set, degree = 15)
mse_df = data.frame(row.names = c("training set size = 50", "training set size = 10000"),
    degree3 = c(re_50_3$mse, re_1w_3$mse), degree15 = c(re_50_15$mse, re_1w_15$mse))
mse_df
```

```
##                             degree3     degree15
## training set size = 50    1.309112 98870.592796
## training set size = 10000 1.199095     1.009384
```

The best possible prediction rule of f can be the model with degree 15 and training set of size 10000. And based on the MSE results, this model actually has the lowest MSE. It's not a surprised because a larger training set can decrease the prediction error and a higher degree model is more similar to the true model($y = 8sin(x)$).

Question 2

```
mse_df
```

```
##                             degree3     degree15
## training set size = 50    1.309112 98870.592796
## training set size = 10000 1.199095     1.009384
```

```
bias_df = data.frame(row.names = c("training set size = 50", "training set size = 10000"),
    degree3 = c(re_50_3$bias, re_1w_3$bias), degree15 = c(re_50_15$bias, re_1w_15$bias))
bias_df
```

```
##                             degree3    degree15
## training set size = 50    1.200663 1535.418412
## training set size = 10000 1.198563    1.007696
```

```
var_df = data.frame(row.names = c("training set size = 50", "training set size = 10000"),
    degree3 = c(re_50_3$var, re_1w_3$var), degree15 = c(re_50_15$var, re_1w_15$var))
var_df
```

```
##                                degree3     degree15
## training set size = 50    0.1095448240 9.831836e+04
## training set size = 10000 0.0005372688 1.705113e-03
```

In first table, we can see the 4 obtained MSE. And I also calculate the bias and variance of predict data on test set. We can get some conclusions:

1.With the size of training set becoming bigger, MSE becomes smaller. Because the variance becomes smaller.

2.When training set is big and the degree of polynomial model increases, MSE is decreasing. Because the bias is decreasing.

3.When training set is small and the degree increases, MSE is greatly increasing. Because the bias and variance are all largely increasing, which indicates that there is possibly an overfitting problem.