HW15 – Q3

Nick Huo

2022-10-21

Question 3

Set-up

```r
set.seed(500)

X.true <- matrix(NA, 1000, 3)
X.true[,1] <- 1
X.true[,2] <- rnorm(1000, 0, 1)
X.true[,3] <- rnorm(1000, 0, 1)
beta.true <- matrix(rnorm(3, 0, 1), 3, 1)
sigma2.true <- 0.05

Y <- rnorm(X.true %*% beta.true, sigma2.true)
X.big <- matrix(rnorm(1000*800,0,1), 1000, 800)

r2.store <- rep(NA, 800)
r2adj.store <- rep(NA,800)
```

Run Simulation

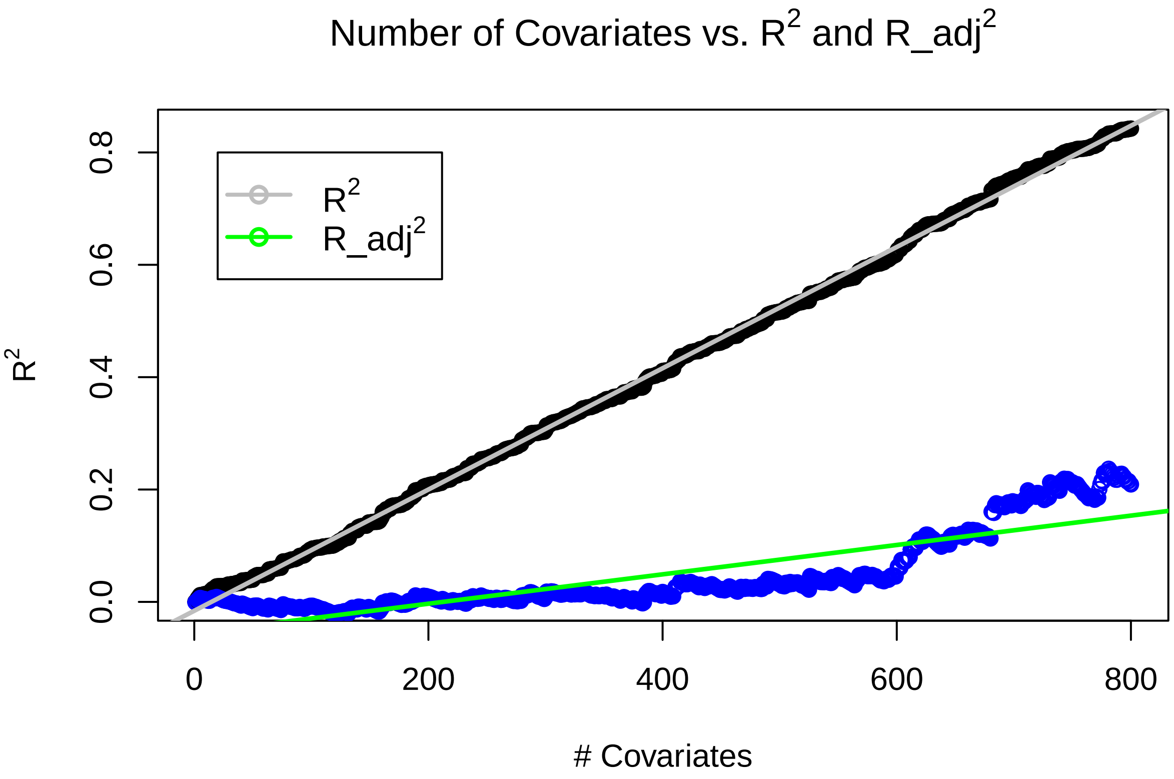
```r
for (i in 1:800) {
  this_model <- lm(Y ~ X.big[,1:i])
  r2.store[i] <- summary(this_model)$r.squared
  r2adj.store[i] <- summary(this_model)$adj.r.squared
}
```

Part a.

```r
num_covar <- c(1:800)

plot(num_covar, r2.store, xlab="# Covariates", ylab=expression(R^2),
     main=expression(paste("Number of Covariates vs. ",R^2," and ",R_adj^2)))
points(num_covar, r2adj.store, col="blue")
abline(lm(r2.store ~ num_covar), col="grey", lwd=2.3)
abline(lm(r2adj.store ~ num_covar), col="green", lwd=2.3)
```

```
legend(20,0.8, legend=c(expression(R^2),expression(R_adj^2)), pch = c(1,1),
       col=c("grey","green"), lty=1, lwd=2, cex=1.1)
```



Part b.

We see that as the number of covariates increase, R^2 also increased with it linearly; while for R^2_{adj}, as as the number of covariates increase, it is also increasing, but at a much slower rate compared to R^2. Also we notice that, for the R^2_{adj}, it increased really slowly with the number of covariates, but after the number of covariates increased to more than 600, R^2_{adj} increased at a faster rate.

The problem is that, all the X data are randomly generated. So there shouldn't be any relationship between the X covariates and our randomly simulated Y. We notice the phenomenon that R^2 not being an accurate descriptor of the model's power because it is affected by the large amount of predictors we have. This is why we want to use R^2_{adj} instead of R^2 when there are many predictors.