Exercize 11 Section 7.9 ISL

In Section 7.7, it was mentioned that GAMs are generally fit using a backfitting approach. The idea behind backfitting is actually quite simple. We will now explore backfitting in the context of multiple linear regression. Suppose that we would like to perform multiple linear regression, but we do not have software to do so. Instead, we only have software to perform simple linear regression. Therefore, we take the following iterative approach: we repeatedly hold all but one coefficient estimate fixed at its current value, and update only that coefficient estimate using a simple linear regression. The process is continued until convergence - that is, until the coefficient estimates stop changing. We now try this out on a toy example.

(a) Generate a response Y and two predictors X_1 and X_2, with $n = 100$ as follows:

```
set.seed(123)
n = 100
x1 = rnorm(n)
x2 = rnorm(n)
y = 5+ 1*x1 + 3*x2 + rnorm(n)
```

(c) Initialize $\hat{\beta}_1$ to take on a value of your choice. It does not matter what value you choose.

```
## [1] -5
```

(d) Keeping $\hat{\beta}_1$ fixed, fit the model

$$Y - \hat{\beta}_1 X_1 = \beta_0 + \beta_2 X_2 + \epsilon$$

You can do this as follows

```
r = y - hatbeta1*x1
hatbeta2 = lm(r ~ x2)$coef[2]
```

```
##       x2
## 2.749494
```

(e) Keeping $\hat{\beta}_2$ fixed, fit the model

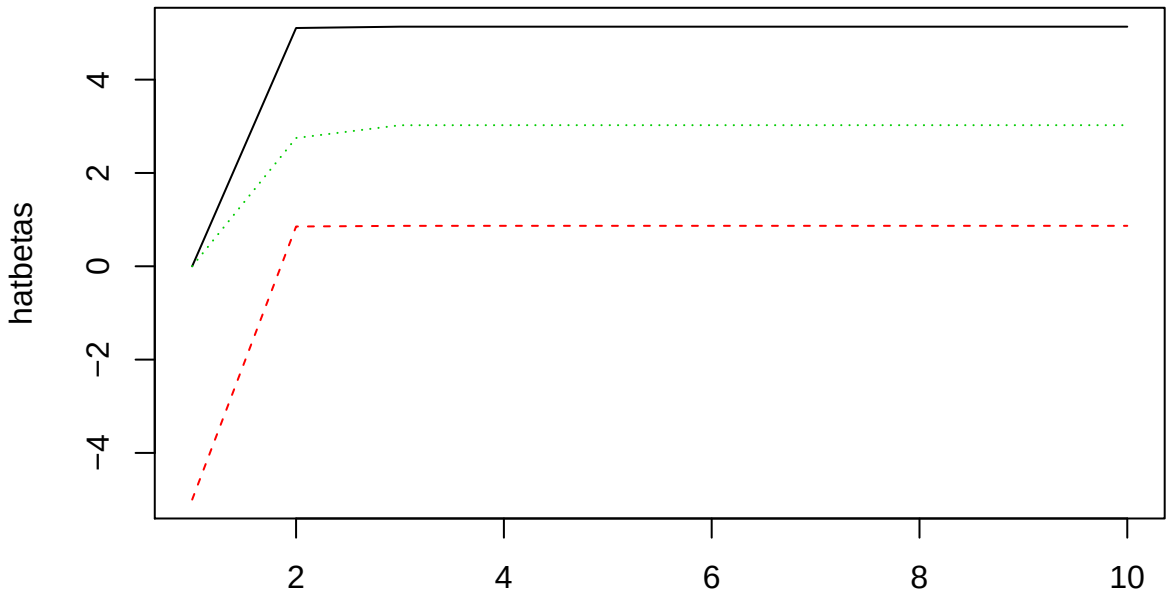
$$Y - \hat{\beta}_2 X_2 = \beta_0 + \beta_1 X_1 + \epsilon$$

You can do this as follows

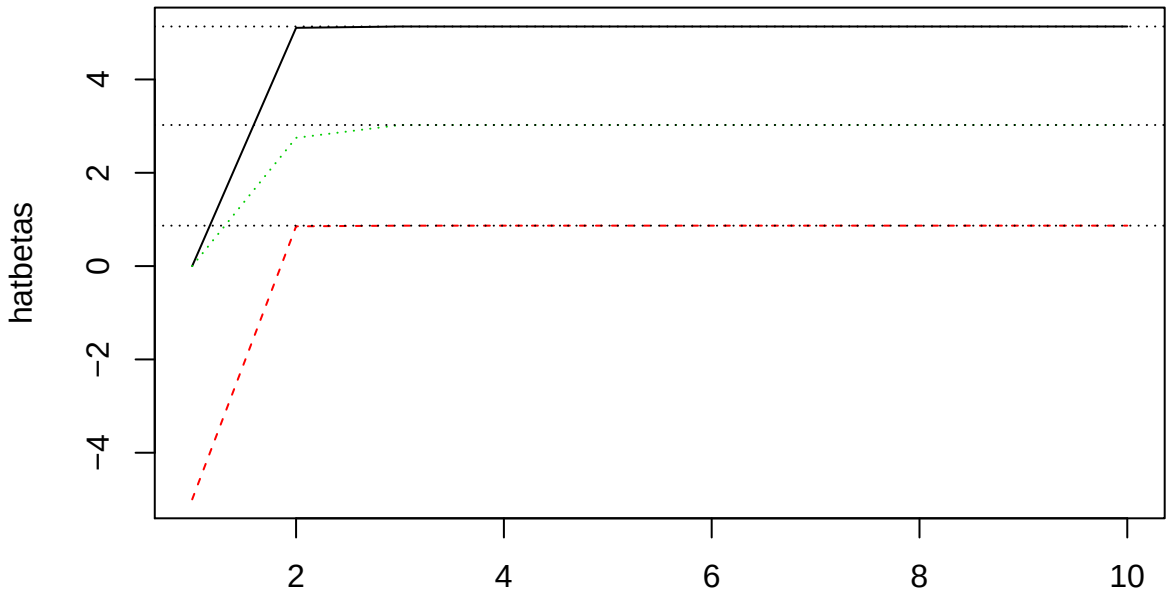
```
r = y - hatbeta2*x2
hatbeta1 = lm(r ~ x1)$coef[2]
```

```
##        x1
## 0.8524346
```

(e) Write a for loop to repeat (c) and (d) 1000 times. Report the estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ at each iteration of the for loop. Create a plot in which each of these values is displayed, with $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ each shown in a different color



(f) Compare your answer in (e) to the results of simply performing multiple linear regression to predict Y using X_1 and X_2. Use the `abline()` function to overlay those multiple linear regression coefficient estimates on the plot obtained in (e).



(g) On this data set, how many backfitting iterations were required in order to obtain a "good" approximation to the multiple regression coefficient estimate?