Simple Linear Regression
Econometrics
Sai Sandeep Chandaluri

Question 1

a. We know that the intercept:

$$b_0 = \bar{Y} - b_1\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{X}\sum_{i=1}^{n} c_i Y_i$$

$$= \sum_{i=1}^{n}(\frac{1}{n} - c_i\bar{X})Y_i = \sum_{i=1}^{n} d_i Y_i$$

where $d_i = \frac{1}{n} - c_i\bar{X}$ and $c_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$

b. From (a), we know:

$$b_0 = \bar{Y} - b_1\bar{X}$$

Taking expectation on both sides, we get

$$E(b_0) = E(\bar{Y} - b_1\bar{X})$$

Since $E(\bar{Y}) = \beta_0 + \beta_1\bar{X}$, we get

$$E(b_0) = \beta_0 + \beta_1\bar{X} - E(b_1)\bar{X} \tag{1}$$

Let's try to derive $E(b_1)$. By Simple Regression model,

$$\bar{Y} = \beta_0 + \beta_1\bar{X} + \varepsilon$$

And

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1\bar{X} - \bar{\varepsilon})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Here, we know $\bar{\varepsilon} = 0$. Re-arranging the above equation,

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Taking expectation on both sides and using the fact that $E(\varepsilon_i) = 0$, we get

$$E(b_1) = \beta_1$$

Substituting this result in (1), we get

$$E(b_0) = \beta_0$$

c. From (a), We know that

$$b_0 = \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})Y_i$$

Re-arranging the above (Adding and subtracting $\hat{Y}$, the true fitted value),

$$b_0 = \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})(Y_i - \hat{Y} + \hat{Y})$$

$$= \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})(\varepsilon_i + \hat{Y}) = \bar{Y} - \bar{X}\hat{Y}\frac{\sum_{i=1}^{N}(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} + \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})\varepsilon_i$$

$$= \bar{Y} - \beta_1\bar{X} + \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})\varepsilon_i = \beta_0 + \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})\varepsilon_i$$

(Note that we have used the fact that $\beta_1 = \hat{Y}\frac{\sum_{i=1}^{N}(X_i-\bar{X})}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$)

Computing the variance (note that we can write the variance as sum of squares as all the terms are independent),

$$Var(b_0) = \sum_{i=1}^{n}(\frac{1}{N} - c_i\bar{X})^2\sigma^2$$

Expanding the above and using the properties $\sum c_i = 0$ and $\sum c_i^2 = \frac{1}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$, we get

$$Var(b_0) = [\frac{1}{N} + \frac{\bar{X}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}]\sigma^2 = \sigma^2[\frac{1}{N} + \frac{\bar{X}^2}{(N-1)s_X^2}]$$

Question 2

a. Writing the function:

```
simulate_simple_regression <- function(intercept, slope, X, err_sd){
  Y <- vector(length=length(X))
  for (i in seq(1, length(X), 1)){
  Y[i] = rnorm(1, mean = intercept + (slope * X[i]), sd = err_sd)
  }
  return(Y)
}
```

b. Given that $\beta_0 = 1$, $\beta_1 = 20$, and $\sigma = 1$.

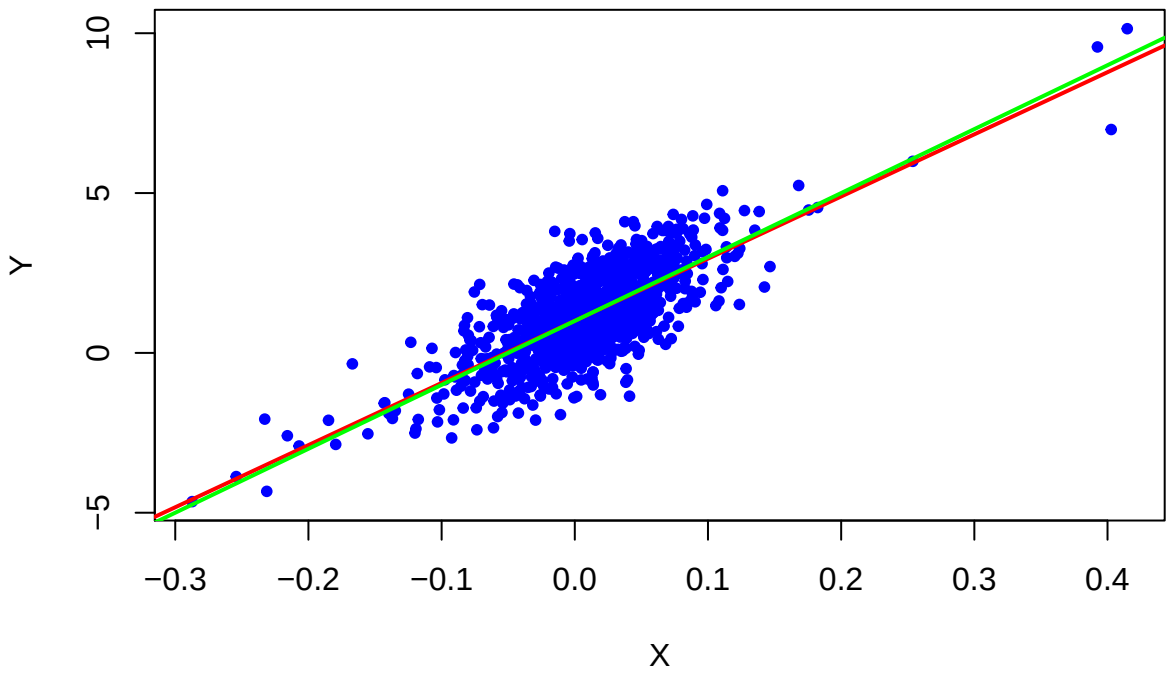
```
library(DataAnalytics)
data(marketRf, package="DataAnalytics")
X = marketRf$vwretd
Y = simulate_simple_regression(1, 20, X, 1)

plot(Y ~ X,pch=20, col="blue")
out_reg = lm(Y ~ X)
abline(coef(out_reg), col="red", lwd=2)

Y_true <- vector(length=length(X))
for (i in seq(1, length(X), 1)){
```

```
  Y_true[i] = 1 + (20 * X[i])  # True Conditional Mean Line is computed by finding expected value of Y
  }

out_true = lm (Y_true ~ X)
abline(coef(out_true), col="green", lwd=2)
```



We have plotted the scatterplot of X vs simulated Y. Red line indicates the fitted regression line and green line indicates the true conditional mean line.

Question 3

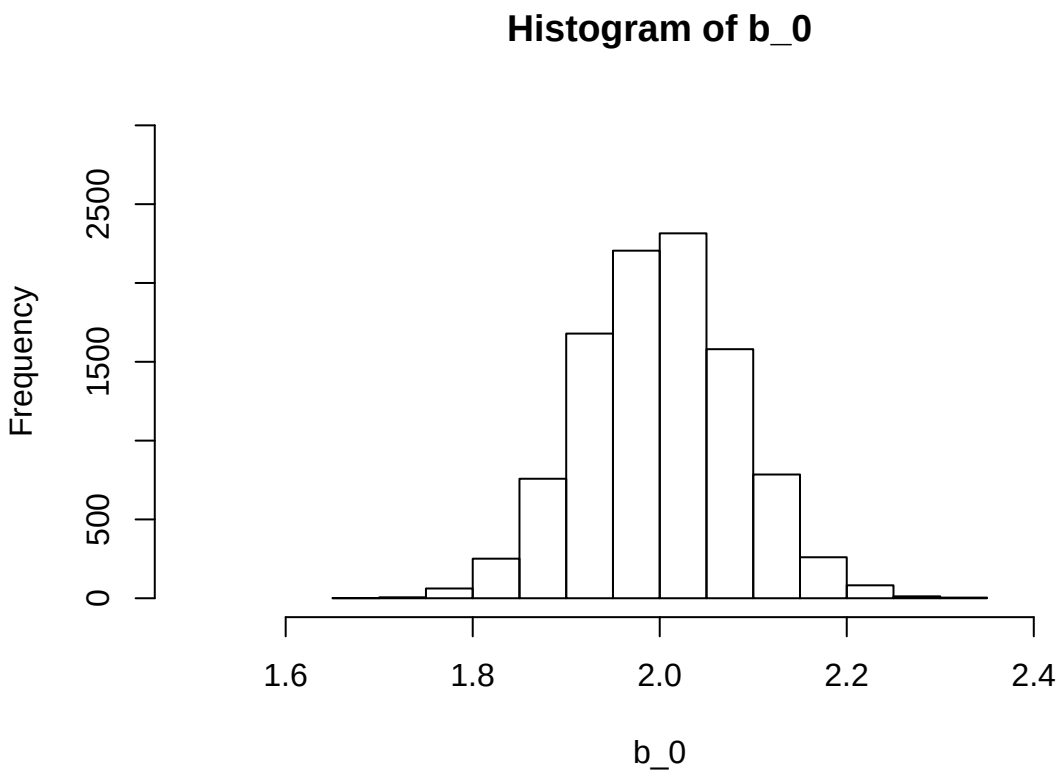
Given $Y = \beta_0 + \beta_1 X + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$. Let $\beta_0 = 2$, $\beta_1 = 0.6$, and $\sigma^2 = 2$.

a. Using a sample size of 300 and 10000 samples (The 300 sample is taken from first 300 values of vwretd in the marketRf dataset):

```
library(DataAnalytics)
data(marketRf, package="DataAnalytics")
X1 = marketRf$vwretd[1:300]

count <- (1:10000)
b_0 <- c()

for(i in count){
Y = simulate_simple_regression(2, 0.6, X1, sqrt(2))
lmResult = lm (Y ~ X1)
b_0[i] <- coef(lmResult)["(Intercept)"]
}
hg <- hist(b_0, plot = FALSE)
plot(hg, ylim = c(0, 3000), xlim = c(1.5,2.5))
```



b. Empirical Value of $E(b_0)$ is the mean of all the values of b_0 obtained from the simulation.

Hence, calculating empirical value of b_0 in R gives:

```
b_0_emp <- mean(b_0)
b_0_emp #Empirical Value of b_0
```

```
## [1] 2.000511
```

From 1(b), we have theoretical value of $E(b_0) = \beta_0$, which implies that $E(b_0) = 2$.

Comparing the simulated value with theoretical value, we can see that empirical value ~ 2.0, which is the theoretical value. The slight difference is observed due to sampling error.

c. Empirical Value of $Var(b_0)$ is the variance of all the values of b_0 obtained from the simulation.

Hence, calculating empirical value of b_0 in R gives:

```
b_0_emp_var <- var(b_0)
b_0_emp_var #Empirical Value of variance of b_0
```

```
## [1] 0.006846087
```

From 1(c), we have theoretical value of $Var(b_0) = \sigma^2[\frac{1}{N} + \frac{\bar{X}^2}{(N-1)s_X^2}]$

Here, we have $N = 300$ and $\sigma^2 = 2$. Calculating mean and variance of X from the data, we have

```
X1_mean = mean(X1)
X1_mean # Mean of X1 is 0.00884
```

```
## [1] 0.008840307
```

```
X1_var = var(X1)
X1_var # Variance of X1 is 0.00627
```

```
## [1] 0.006272678
```

```
b_0_ther_var = 2 * (sum((1/300),(X1_mean^2/(299 * X1_var))))
b_0_ther_var #Theoretical value of variance of b_0 which is approximately 0.00675
```

```
## [1] 0.006750004
```

Comparing the simulated value with theoretical value, we can see that empirical value ~ 0.00675, which is the theoretical value. The slight difference is due to sampling error.

Question 4

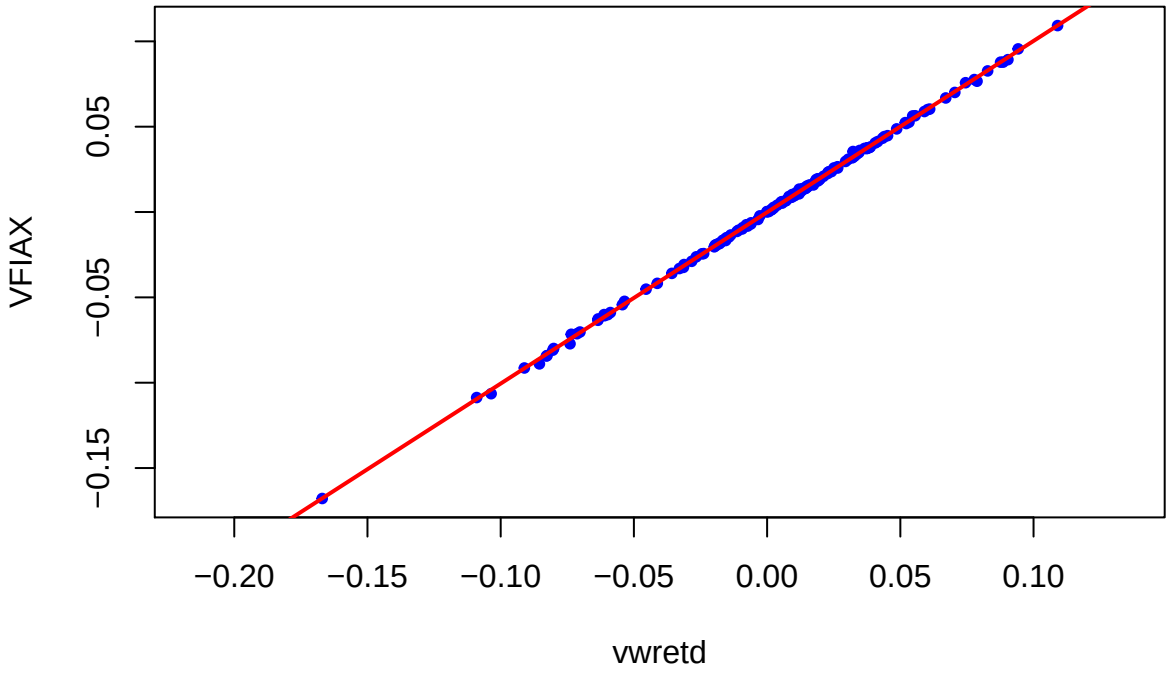
Fitting a regression between VFIAX of Vanguard dataset and vwretd of marketRf dataset:

```
library(DataAnalytics)
library(reshape2)
data(Vanguard)

Van=Vanguard[,c(1,2,5)]

V_reshaped=dcast(Van,date~ticker,value.var="mret")
data(marketRf, package = "DataAnalytics")
Van_mkt = merge(V_reshaped, marketRf, by="date")

plot(VFIAX ~ vwretd, data=Van_mkt, pch=20, col="blue")
out = lm(VFIAX ~ vwretd, data=Van_mkt)
abline(coef(out), col="red", lwd=2)
```



a. Given Null hypothesis $H_0^a : \beta_1 = 1$

We know that

$$t = \frac{b_1 - \beta_1^*}{s_{b_1}}$$

Computing b_1:

```r
b_1 <- coef(out)["vwretd"]
b_1 #Value of b_1
```

```
##   vwretd
## 1.003735
```

Computing s_{b_1}:

$$s_{b_1} = \sqrt{\frac{s^2}{(N-1)s_x^2}}$$

Here,

$$s = \sqrt{\frac{SSE}{N-2}}$$

```r
anova(out)
```

```
## Analysis of Variance Table
##
## Response: VFIAX
##            Df   Sum Sq  Mean Sq F value    Pr(>F)
## vwretd      1 0.304998 0.304998  485731 < 2.2e-16 ***
## Residuals 149 0.000094 0.000001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sd(as.numeric(unlist(out$model["vwretd"])))
```

```
## [1] 0.0449246
```

From the anova table, we have $SSE = 0.000094$ and $N-2 = 149$. Therefore, $s = 0.0007943$ and $s_x = 0.044925$.

Substituting these values, we have $s_{b_1} = 0.0014436$

Computing t-statistic from these values, we get

$$t = \frac{b_1 - \beta_1^*}{s_{b_1}} = \frac{1.003735 - 1}{0.0014436} = 2.587$$

For 0.05 significance level, we have

```r
qt(0.025, df = 149)
```

```
## [1] -1.976013
```

We can see that 2.587 is larger than the 95% critical value for t(149), which is 1.976. So we reject the null hypothesis: $H_0^a : \beta_1 = 1$, at 0.05 level of significance.

b. Given Null hypothesis $H_0^a : \beta_0 = 0$

We know that

$$t = \frac{b_0 - \beta_0^*}{s_{b_0}}$$

Computing b_0:

```r
b_0 <- as.numeric(coef(out)["(Intercept)"])
b_0 #Value of b_0
```

```
## [1] -0.0001313686
```

Computing s_{b_0}:

```
se_b_0 <- as.numeric(sqrt(diag(vcov(out)))["(Intercept)"])
se_b_0 #Value of Standard Error of b_0
```

```
## [1] 6.47522e-05
```

From the above, we have $b_0 = -0.0001314$ and $s_{b_0} = 6.47522 * 10^{-05}$

Computing t-statistic from these values, we get

$$t = \frac{b_0 - \beta_0^*}{s_{b_0}} = \frac{-0.0001314 - 0}{6.47522 * 10^{-05}} = -2.0293$$

We have corresponding p-value:

```
t = (b_0-0)/se_b_0
pvalue = 2*pt(-abs(t), df = 149)
pvalue # Coumputed P value
```

```
## [1] 0.04426054
```

We can see that P value is approximately 0.044, which is greater than 0.01. Hence, we cannot reject the Null Hypothesis $H_0^b : \beta_0 = 0$, at 0.01 significance level.

Question 5

Standard errors and p-values.

a. Standard Error of a sample statistic or an estimator:

Many times, we do not fully know about the complete population and often deal with samples of original population. While dealing with these samples, every sample statistic or estimator will have an uncertainty across different samples. Standard error is the estimated standard deviation of a statistic or an estimate. It is different from standard deviation as it is the estimated value of standard deviation.

Standard deviation is more broader concept where it is defined as the measure of spread of the data and is applicable to whole population or any sample. Each sample statistic/estimator can have different standard deviations and the estimated value of all such sample standard deviations is Standard Error.

b. Sampling Error and Standard Error:

When the sample data that is chosen from a population do not completely represent the entire population in terms of its results, characteristics etc., we often end up with a statistical error called the Sampling Error.

Standard Error captures the sampling error as it is an estimated value of standard deviation (how the statistic/estimator is spread across samples). Hence, we try to express any statistic/estimator with a confidence interval around its expected value to capture Sampling error, with a standard deviation equal to standard error. This can be understood because otherwise, statistic/estimator would have been always equal for all samples and it should be the same irrespective of the sample taken.

c. Recommendation to Steven:

The parameter estimates and standard errors help Steven in building the model to predict Y (predicted output) given X (input) and also in understanding how well are X and Y correlated to each other. But this model needs a metric to evaluate how well it is estimating true parameters. So, using the parameter estimates and standard errors, I would suggest Steven to give estimates of the parameters with confidence intervals and the level of significance that they can be accepted (maybe by using t-values or p-values).

d. Recommendation to Xingua

With the test statistic and p-value that Xingua has in her output, she knows the minimum significance level at which she can reject the Null Hypothesis. So if p-value for her Null Hypothesis is less than the significance level that she has in mind, she has to reject the Null. Otherwise, if p-value is greater than the significance level she has in mind, she has to accept the Null.

$$p - value < \alpha \implies Reject\ null$$
$$p - value \geq \alpha \implies Accept\ null$$

Question 6

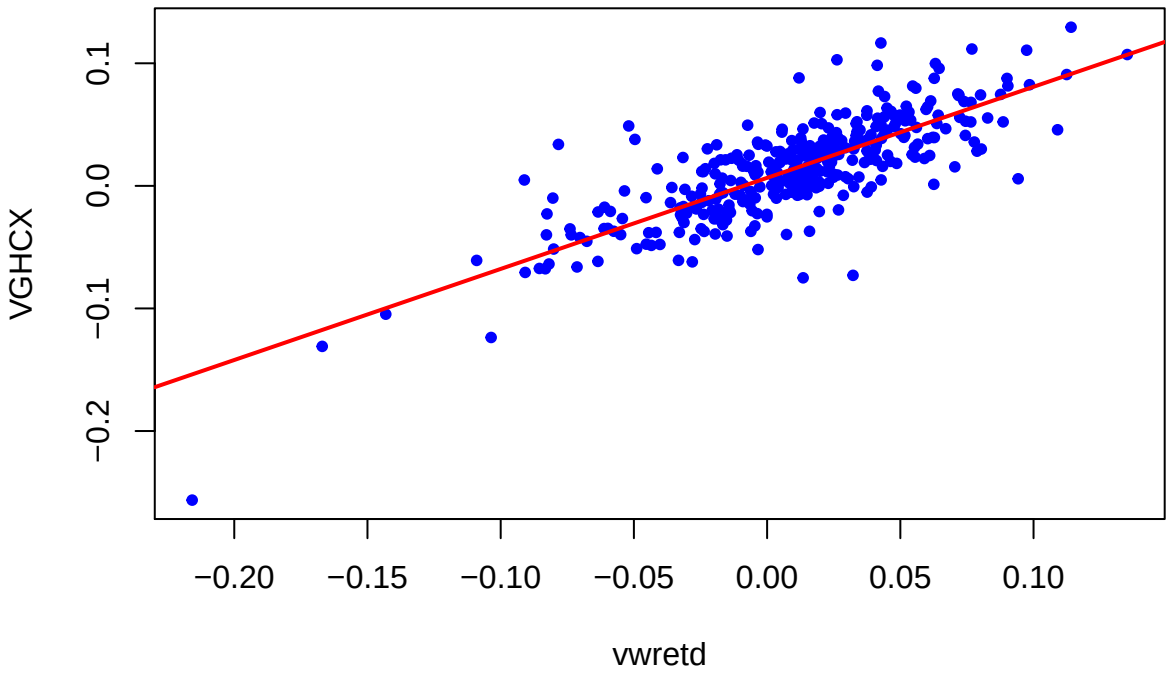
Fitting a regression between VGHCX of Vanguard dataset and vwretd of marketRf dataset:

```
library(DataAnalytics)
library(reshape2)
data(Vanguard)

Van=Vanguard[,c(1,2,5)]

V_reshaped=dcast(Van,date~ticker,value.var="mret")
data(marketRf, package = "DataAnalytics")
Van_mkt = merge(V_reshaped, marketRf, by="date")

plot(VGHCX ~ vwretd, data=Van_mkt, pch=20, col="blue")
reg_out = lm(VGHCX ~ vwretd, data=Van_mkt)
abline(coef(reg_out), col="red", lwd=2)
```



a. Given that the market is up by 5%. Hence X = 0.05

True Conditional Mean of VGHCX is obtained by:

$$Return_{VGHCX} = \beta_0 + \beta_1(Return_{market})$$

But since we do not know β_0 and β_1, we use the estimates b_0 and b_1. Hence, estimate of the conditional mean:

$$Return_{VGHCX} = b_0 + b_1(Return_{market})$$

```
b_0_estimate <- as.numeric(coef(reg_out)["(Intercept)"])
b_0_estimate #Value of b_0_estimate
```

```
## [1] 0.006527817
```

```
b_1_estimate <- as.numeric(coef(reg_out)["vwretd"])
b_1_estimate #Value of b_1_estimate
```

```
## [1] 0.7430089
```

```
sum(b_0_estimate, b_1_estimate * 0.05)
```

```
## [1] 0.04367826
```

Using these values, we get estimate of conditional mean of Vanguard HCX's fund for market return of 5% as $E[Y|X = 0.05] = 4.368\%$.

b. Given that the market is up by 10%. Hence X = 0.1

Estimate of the conditional standard deviation of Vanguard HCX fund is simply the estimate of σ_ε, which is the standard error of regression.

We know,

$$s = \sqrt{\frac{SSE}{N-2}}$$

```
anova(reg_out)
```

```
## Analysis of Variance Table
##
## Response: VGHCX
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## vwretd      1 0.38142 0.38142  607.92 < 2.2e-16 ***
## Residuals 347 0.21771 0.00063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sd(as.numeric(unlist(reg_out$model["vwretd"])))
```

```
## [1] 0.04455729
```

From the anova table, we have $SSE = 0.21771$ and $N - 2 = 347$. Therefore, $s = 0.02505$.

Note that this estimate of conditional standard deviation is independent of market return.

c. Given that the market is up by 15%. Hence X = 0.15

We know that the prediction error s_{pred} is given by

$$s_{pred} = s(1 + \frac{1}{N} + \frac{(X_f - \bar{X})^2}{(N-1)s_x^2})^{\frac{1}{2}}$$

From (b), we know $s = 0.02505$. Given $X_f = 0.15$.

Computing the mean and variance of *vwretd*, we get

```r
mean(as.numeric(unlist(reg_out$model["vwretd"])))
```

```
## [1] 0.009961438
```

```r
var(as.numeric(unlist(reg_out$model["vwretd"])))
```

```
## [1] 0.001985352
```

$\bar{X} = 0.00996$, $s_x^2 = 0.00198$ and $N = 349$. Substituting these values in above equation, we get

```r
0.02505 * sqrt(sum(1, 1/349, (0.15-0.009961438)^2/(348*0.001985352)))
```

```
## [1] 0.02543839
```

$$s_{pred} = 2.54\%$$