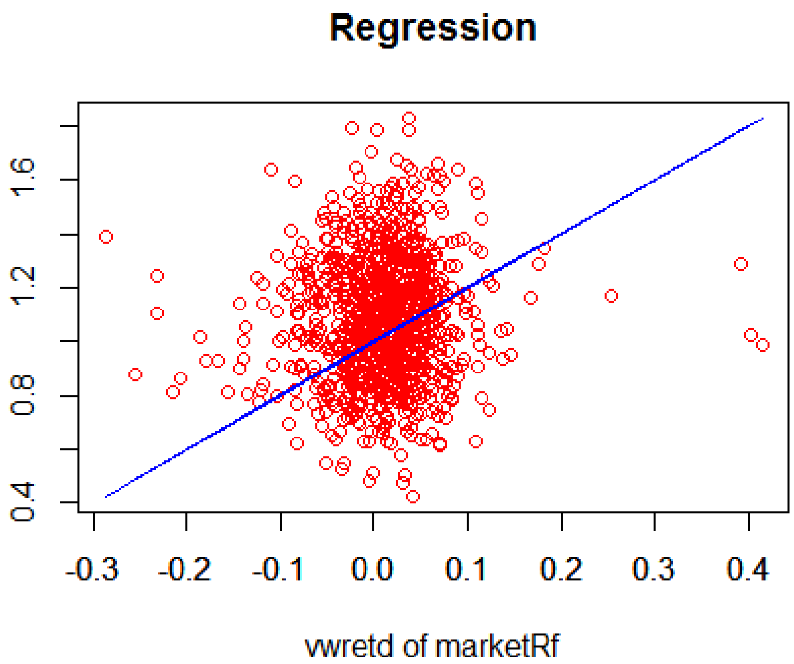
Homework 2

Nikhil Guruji

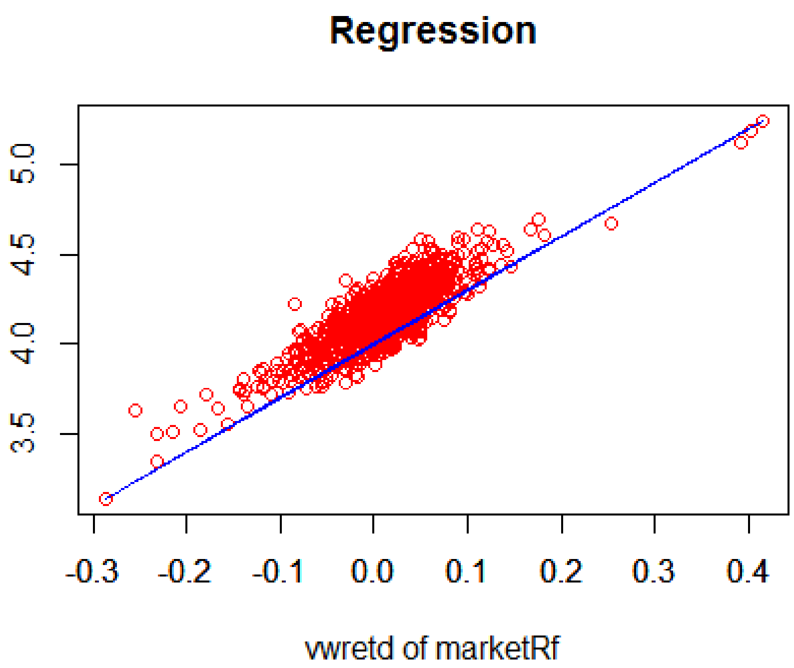
October 24, 2017

1- Regression function:

```
library(DataAnalytics)
data("marketRf")
regr<-function(beta0,beta1,x,sd){
  e<-rnorm(length(x),0,sd)
  yhat<-beta0+(beta1*x)
  y<-yhat+e
  plot(x,y,col="red",ylab="",xlab="",yaxt="n")
  par(new=TRUE)
  plot(x,yhat,type="l",col="blue",ylab="",xlab="vwretd of
marketRf",main="Regression")
}
asdf<-regr(1,2,marketRf$vwretd,5) #more dispersion
```




```
asd<-regr(4,3,marketRf$vwretd,0.1) #less dispersion
```



2-

We know from the slides that:

$b_1 = \frac{\sum(X_i - \overline{X})(Y_i)}{\sum(X_i - \overline{X})^2}$

we can find b_0 by substituting the above value in the equation:

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

substituting the value of b1,

$b_0 = \overline{Y} - \frac{\sum(X_i - \overline{X})(Y_i)}{\sum(X_i - \overline{X})^2}\overline{X}$

Let $c_i = \frac{(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}$

Then, $b_0 = \overline{Y} - (\sum c_i Y_i)\overline{X}$

$Var(b_0) = Var\left(\overline{Y} - (\sum c_i Y_i)\overline{X}\right)$

$Var(b_0) = Var(\overline{Y}) + (\overline{X})^2(Var(b_1)) + 2(\overline{Y})(\overline{X})(b_1)cov(b_1\overline{X}, \overline{Y})$

$Var\left(\overline{Y}\right) = \frac{1}{N}\sum Var(Y_i)$

$$Var(b_0) = \frac{\sigma^2}{N} + (\overline{X})^2 Var(\sum c_i Y_i | X)$$

$$Var(b_0) = \frac{\sigma^2}{N} + (\overline{X})^2 \frac{\sigma^2}{\sum(X_i - \overline{X})^2}$$

$$Var(b_0) = \frac{\sigma^2}{N} + (\overline{X})^2 \frac{\sigma^2}{(N-1)s_x^2}$$

$$E[b_0] = E[\overline{Y}] - E[\sum c_i Y_i] E[\overline{X}]$$

$= \overline{Y} - E[b_1]\overline{X}$

$$E[b_1] = E\left[\frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}\right]$$

$= E\left[\frac{\sum(X_i - \overline{X})(\beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 - \beta_1 \overline{X} - \overline{\epsilon})}{\sum(X_i - \overline{X})^2}\right]$

$= \beta_1 + \frac{\sum(X_i - \overline{X})(\epsilon_i - \overline{\epsilon})}{\sum(X_i - \overline{X})^2}$

$= \beta_1 + \frac{\sum(X_i - \overline{X})(\epsilon_i)}{\sum(X_i - \overline{X})^2}$

Since $E[\epsilon_i] = 0$,

$E[b_1] = \beta_1$

and,

$E[b_0] = \overline{Y} - \beta_1 \overline{X}$

$= \beta_0 + \beta_1 \overline{X} - \beta_1 \overline{X}$

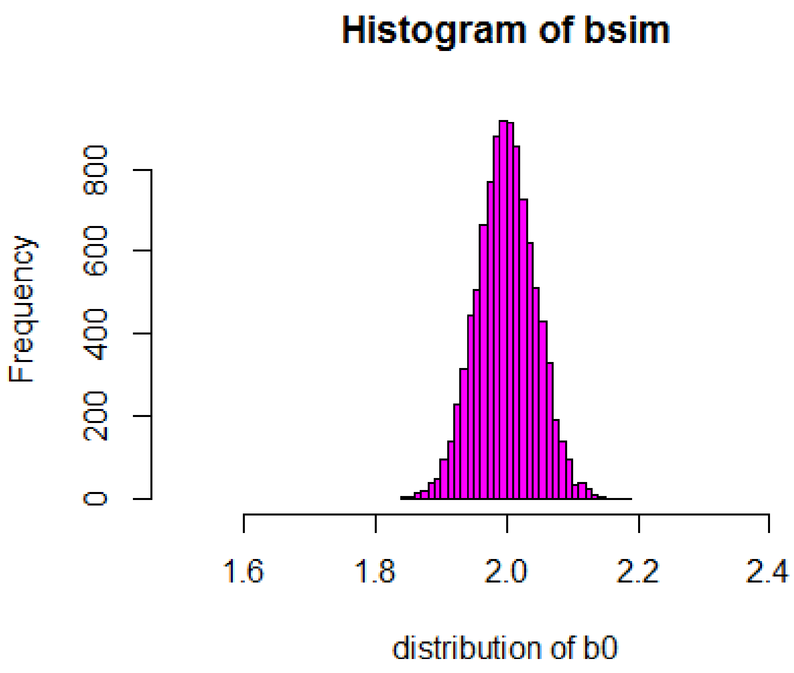
$E[b_0] = \beta_0$

3-

```
simreg=function(beta0,beta1,sigma,x){
  y=beta0+beta1*x+rnorm(length(x),sd=sigma)
}

x=marketRf$vwretd
beta0=2
beta1=0.6
sigma=sqrt(2)
nsample=10000
bsim=double(nsample)
for(i in 1:nsample){
  y=simreg(beta0,beta1,sigma,x)
```

```
  bsim[i]=lm(y~x)$coef[1]
}
hist(bsim,breaks=40,col="magenta",xlim=c(1.5,2.5),xlab="distribution of b0")
```




```
sd(bsim)

## [1] 0.04434364
```

From Question 2, we got the following results:

$E[b_0] = \beta_0$

and

$Var(b_0) = \frac{\sigma^2}{N} + (\overline{X})^2 \frac{\sigma^2}{(N-1)s_x^2}$

so, theoretical values are:

$E[b_0] = 2$ (because β_0=2)

```
sigma=sqrt(2)
n=length(x)
x=marketRf$vwretd
modif<-rep(0,length(x))
for(i in 1:length(x)){
  modif[i]<-(x[i]-mean(x))^2
}
```

```r
varb0<-((sigma^2)/n)+((mean(x))^2 * sigma^2)/((n-1)*var(x))
sqrt(varb0)

## [1] 0.04439179
```

We can see that the standard deviation from the theoretical calculations matches the standard deviation from the sampling distribution of b_0 = 4.4%

4- Hypothesis Tests:

```r
out<-lm(VFIAX~Van_mkt$vwretd,data=Van_mkt)
b1est<-summary(out)$coefficients[2,1]
b1se<-summary(out)$coefficients[2,2]
b1t<-summary(out)$coefficients[2,3]
b1p<-summary(out)$coefficients[2,4]

b0est<-summary(out)$coefficients[1,1]
b0se<-summary(out)$coefficients[1,2]
b0t<-summary(out)$coefficients[1,3]
b0p<-summary(out)$coefficients[1,4]

#part a

#null is beta1=1 at 0.05 level of signifance
t0_05<-qt(0.025,length(x)-2) #double tailed test
ttest1<-(b1est-1)/(b1se)

if(abs(ttest1)>abs(t0_05)){
  print(cat("reject the null (beta1=1) since t test is out of limits","t-
statistic=",ttest1,"p-value=",b1p))
} else print(cat("accept or fail to reject the null (beta1=1)","t-
statistic=",ttest1,"p-value=",b1p))

## reject the null (beta1=1) since t test is out of limits t-statistic=
2.593157 p-value= 1.176664e-263NULL

#part b

#null is beta0=0 at 0.10 level of signifance
t0_10<-qt(0.05,length(x)-2) #double tailed test
ttest0<-(b0est-0)/(b0se)

if(abs(ttest0)>abs(t0_10)){
  print(cat("reject the null (beta0=0) since t test is out of limits","t-
statistic=",ttest0,"p-value=",b0p))
} else print(cat("accept or fail to reject the null (beta0=0)","t-
statistic=",ttest0,"p-value=",b0p))

## reject the null (beta0=0) since t test is out of limits t-statistic= -
2.02879 p-value= 0.04426054NULL
```

5-

a) Standard error essentially gives the accuracy of the estimates of the true population parameteres using the sample parameteres. The standard deviation, on the other hand, gives the dispersion of values around the mean. Sampling error is the error in estimation of actual values. Mathematically, it is the difference between the actual random error terms (ϵ) and the estimated error terms (e).

b) I would suggest him to judge the model based on the value of the standard error. Compared to the mean, the standard error should be as small as possible. This shows that the statistical test he conducted is accurate.

c) I would ask him for a t-value and the p-value. To be significant, the values obtained from the statistical test should be different from zero (which will be the null hypothesis).

If the t statistic of a coefficient is too large (or larger than a critical value depending on the confidence level we choose, for eg 95%), I would suggest him to reject the null hypothesis and conclude that the values obtained from the statistical test are significant and reliable.

If the t statistic doesn't exceed the critical t-value, the model cannot be relied upon since we fail to reject the null that the values obtained from the test are zero.

In the case of p-value, I would suggest him to reject the null and consider the test to be reliable if the p-value is lower than the significance level (like 5%). If p-value is too large, he should not rely upon the test he conducted.

6-

a) Let's first run the regression of VGHCX vs vwretd and estimate b0 and b1

```
out6<-lm(VGHCX~Van_mkt$vwretd,data=Van_mkt)
b1<-summary(out)$coefficients[2,1]
b0<-summary(out)$coefficients[1,1]

# part a
# to find conditional mean return of HCX fund when market goes up by 5%,

hcxmeanret<-b0+b1*0.05
hcxmeanret

## [1] 0.05005536
```

b) The conditional standard deviation of a regression is the estimated standard error of regression which is independent of how X behaves (ie $Cov(X_i, \epsilon) = 0$)

The standard error of regression is given by:

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

where $SSE = \sum e_i^2$ is the sum of squared errors

```
#part b
#conditional standard deviation of returns of HCX given market was up by 10%
or any other number

anovatable<-anova(out6)
sse<-anovatable[2,2]
N<-length(Van_mkt$vwretd)
std_err_reg<-sqrt(sse/(N-2))
std_err_reg

## [1] 0.02504834
```