Problem Set 1

Huanyu Liu

Question 1

Review the basics of summation notation and covariance formulas. Show that:

1a) $\sum_{i=1}^{N}(Y_i - \bar{Y}) = 0$

$\bar{Y}$ is the sample mean:

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

$$N \times \bar{Y} = \sum_{i=1}^{N} Y_i$$

$$\sum_{i=1}^{N} Y_i - N \times \bar{Y} = 0$$

$$\sum_{i=1}^{N}(Y_i - \bar{Y}) = 0$$

1b) $\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N}(X_i - \bar{X})Y_i$

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$= \sum_{i=1}^{N}(X_i - \bar{X})Y_i - \sum_{i=1}^{N}(X_i - \bar{X})\bar{Y}$$

$$= \sum_{i=1}^{N}(X_i - \bar{X})Y_i - \bar{Y} \times \sum_{i=1}^{N}(X_i - \bar{X})$$

We have showed that in 1a): $\sum_{i=1}^{N}(X_i - \bar{X}) = 0$

$$\therefore \sum_{i=1}^{N}(X_i - \bar{X})Y_i - 0 = \sum_{i=1}^{N}(X_i - \bar{X})Y_i$$

Question 2

Define both and explain the difference between (a) the expectation of a random variable and (b) the sample average?

2a) The expectation of a random variable is the average value of the outcomes by theoretically infinity times or long run of the experiment. If it's a discrete random variable, the expectation is a weighted average of all the possible outcomes.

2b) A sample is a subset of the population, which we want to use to represent the characteristics of the population in practical. The average of the observations in the sample is the sample average.

The difference between Expectation of a random variable and Sample average

Expectation of a random variable	**Sample average**
Expectation of a random variable is based on the whole population of observations	**Sample average is the arithmetic mean of random sample values drawn from the population.**
To calculate the expectation of a random variable directly based on the population is very difficult	**It's easy to get the sample average**
The expectation of a random variable is accurate to describe the population	**The sample average is not accurate to represent the characteristics of the population**

Question 3

Review the normal distribution and the mean and variance of a linear combination of two normally distributed random variables. Let $X \sim \mathcal{N}(1, 2)$ and $Y \sim \mathcal{N}(2, 3)$. Note that the second parameter is variance. X and Y are independent. Compute:

3a) $\mathbb{E}[3X]$

$$\begin{aligned}\mathbb{E}[3X] &= 3 \times \mathbb{E}[X] \\ &= 3 \times \mu \\ &= 3 \times 1 \\ &= 3\end{aligned}$$

3b) $\text{Var}(3X)$

$$\begin{aligned}\text{Var}(3X) &= 3^2 \times \text{Var}(X) \\ &= 9 \times 2 \\ &= 18\end{aligned}$$

3c) $\text{Var}(2X - 2Y)$ and $\text{Var}(2X + 2Y)$

$$\begin{aligned}\text{Var}(2X - 2Y) &= 2^2 \times \text{Var}(X - Y) \\ &= 4 \times (\text{Var}(X) + \text{Var}(Y) - 2 \times cov(X, Y))\end{aligned}$$

$$\begin{aligned}
&\because \textbf{X and Y are independent.}\\
&\therefore cov(X,Y) = 0\\
&\therefore \text{Var}(2X - 2Y) = 4 \times (2+3) = 20
\end{aligned}$$

Similarly:

$$\begin{aligned}
\text{Var}(2X + 2Y) &= 4 \times (\text{Var}(X) + \text{Var}(Y) + 2 \times cov(X,Y))\\
&= 20
\end{aligned}$$

3d) Explain why in part (c) you get the same answer no matter whether you add or subtract. (Your answer should discuss both the coefficient on Y and why independence between X and Y is important.)

$$\begin{aligned}
&\because \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \times cov(X,Y)\\
&\therefore \text{Var}(X-Y) = \text{Var}(X + (-Y))\\
&\qquad = \text{Var}(X) + \text{Var}(-Y) + 2 \times cov(X,-Y)\\
&\because \text{Var}(aX) = a^2\text{Var}(X) \textbf{ and } cov(aX,bY) = ab \times cov(X,Y)\\
&\therefore \text{Var}(-Y) = \text{Var}(Y) \textbf{ and } cov(X,-Y) = -cov(X,Y)\\
&\because \textbf{X and Y are independent}\\
&\therefore cov(X,Y) = 0\\
&\therefore \textbf{Even though the sign of the covariance term is opposite, they both equal to 0.}\\
&\therefore \textbf{The variance in (c) is the same, no matter whether you add or subtract.}
\end{aligned}$$

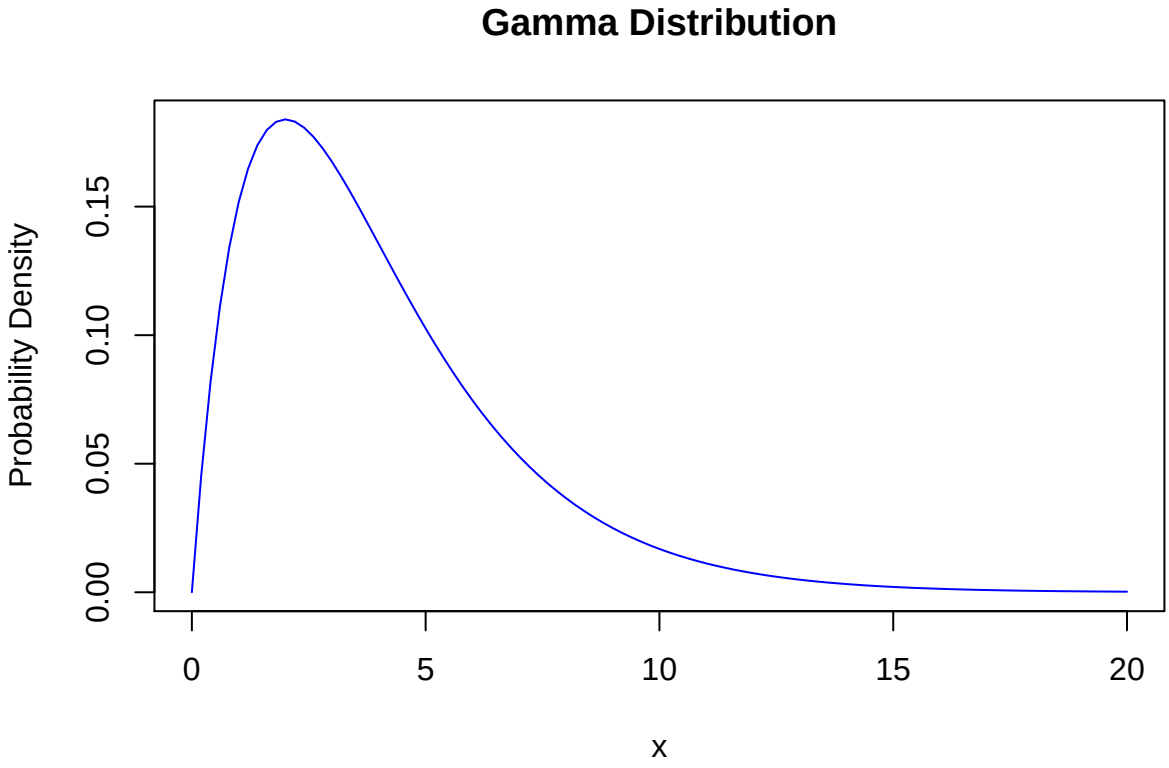
Question 4

4a) Describe the Central Limit Theorem as simply as you can.

Central limit theorem states that the sum of a large number of independent variables tends to normal distribution, no matter those independent variables are normally distributed or not.

4b) Let $X \sim \text{Gamma}(\alpha = 2,\ \beta = 2)$. For the Gamma distribution, α is often called the "shape" parameter, β is often called the "scale" parameter, and the $\mathbb{E}[X] = \alpha\beta$. Plot the density of X and describe what you see. You may find the functions `dgamma()` or `curve()` to be helpful.

```
curve(dgamma(x, shape=2, scale=2),from = 0,to = 20, col='blue',
      ylab = "Probability Density",main = "Gamma Distribution")
```

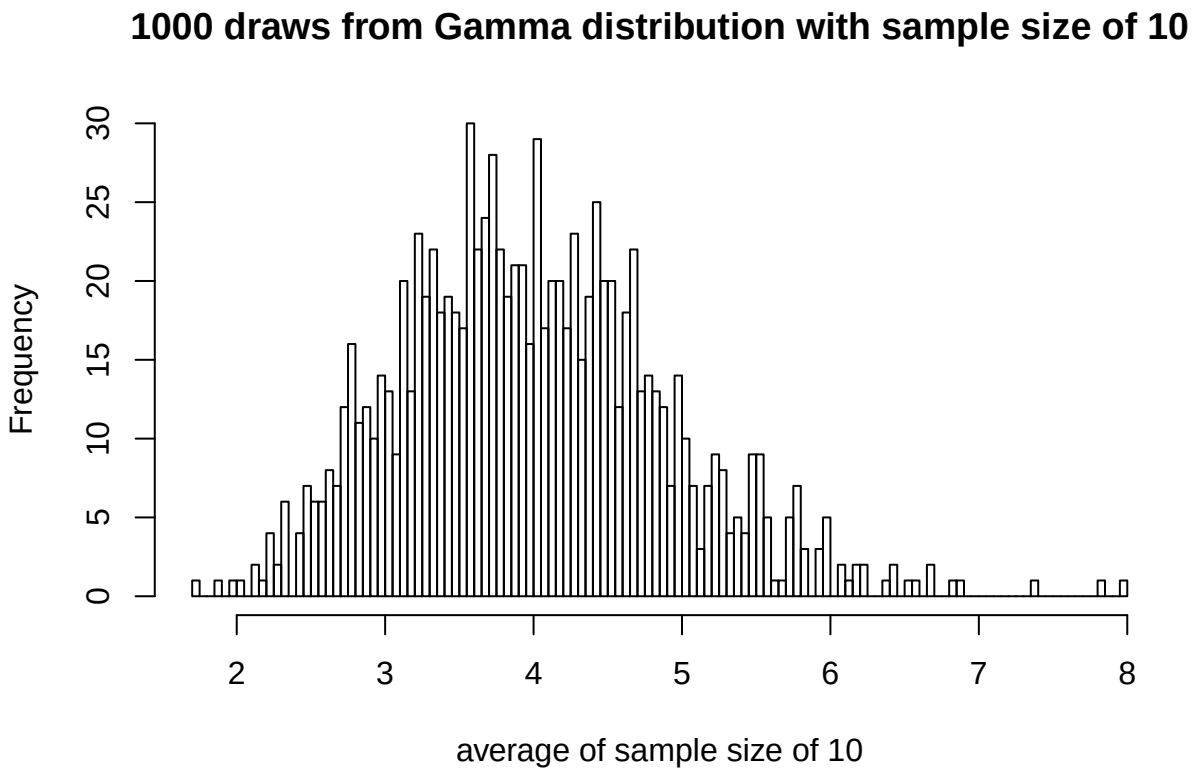


4c) Let n be the number of draws from that distribution in one sample and r be the number of times we repeat the process of sampling from that distribution. Draw an iid sample of size $n = 10$ from the Gamma(2,2) distribution and calculate the sample average; call this $\bar{X}_n^{(1)}$. Repeat this process r times where $r = 1000$ so that you have $\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(r)}$. Plot a histogram of these r values and describe what you see. This is the sampling distribution of $\bar{X}_{(n)}$.

```
myfunction = function(n){            # the sample mean with size n
  x = rgamma(n,shape = 2, scale = 2)
  return(sum(x) / n)
}
n = 10
myfunction(n)
```

```
## [1] 4.887708
```

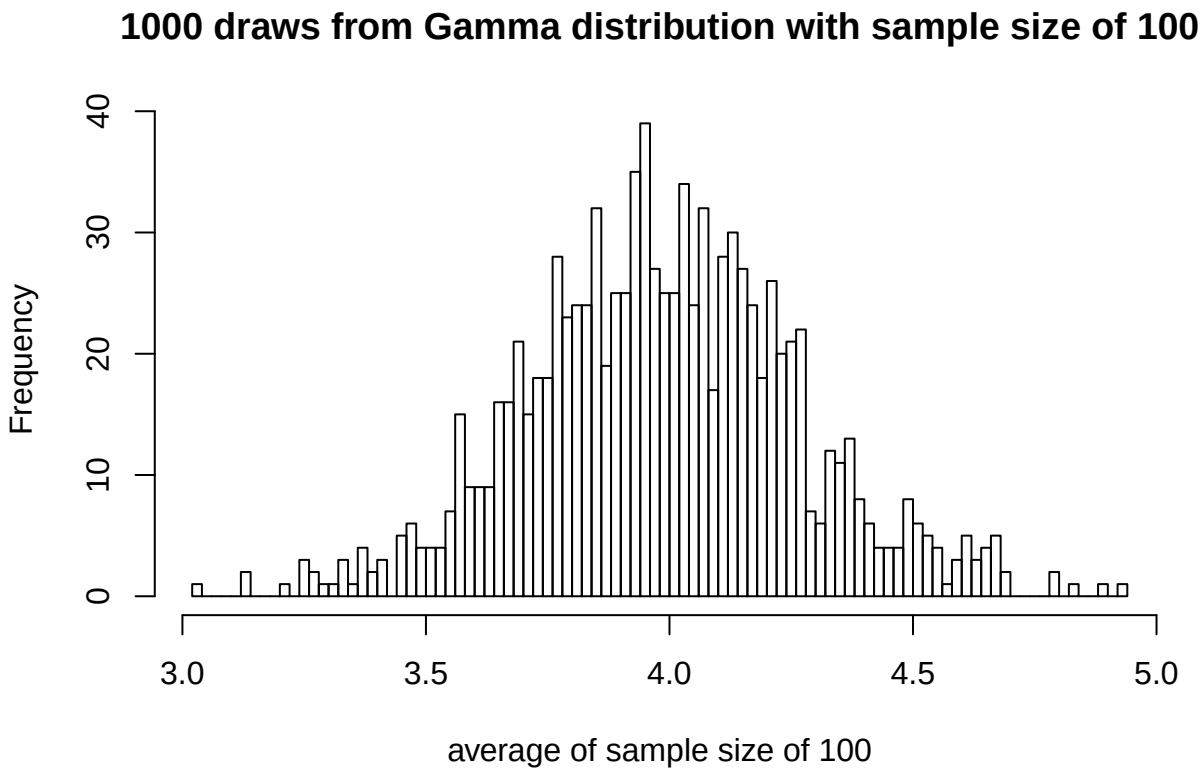
```
r = 1000
average_list = numeric()
for (i in 1:1000){
  average_list[i] = myfunction(n)
}
hist(average_list,breaks = 100,
     main = "1000 draws from Gamma distribution with sample size of 10",
     xlab = "average of sample size of 10")
```



The shape of the histogram is similar to the shape of the pdf of gamma distribution

4d) Repeat part (c) but with $n = 100$. Be sure to produce and describe the histogram.

```
n = 100
for (i in 1:1000){
  average_list[i] = myfunction(n)
}
hist(average_list,breaks = 100,
     main = "1000 draws from Gamma distribution with sample size of 100",
     xlab = "average of sample size of 100")
```



The shape of the histogram is similar to the shape of normal distribution

4e) Let's say you were given a dataset for 2,000 people with 2 variables: each person's height and weight. What are the values for n and r in this "real world" example?

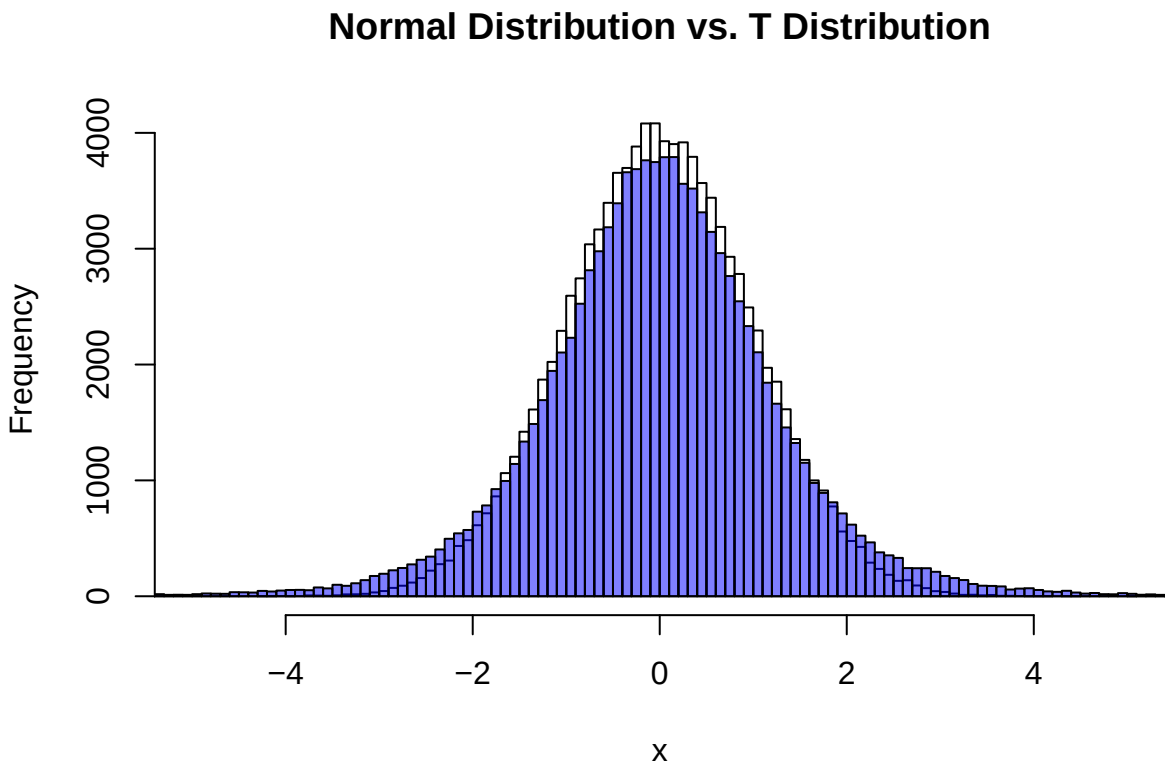
In this real world example, $n = 1000$**,** $r = 1$

Question 5

The normal distribution is often said to have "thin tails" relative to other distributions like the t-distribution. Use random number generation in R to illustrate that a $\mathcal{N}(0, 1)$ distribution has much thinner tails than a t-distribution with 5 degrees of freedom.

A few coding hints: `rnorm()` and `rt()` are the functions in R to draw from a normal distribution and a t-distribution. The option `add=TRUE` for the `hist()` command can be used to overlay a second histogram on top of another histogram, and after installing the `scales` package, you can make a blue histogram 50% transparent with the option `col=scales::alpha("blue",0.5)`. Alternatively, you can put two plots side-by-side by first setting the plotting parameter with the code `par(mfrow=c(1,2))`. You can set the range of the x-axis to go from -5 to 5 with the plotting option `xlim=c(-5,5)`.

```
hist(rnorm(100000,mean = 0,sd = 1),breaks = seq(-50,50,0.1),
     xlim = c(-5,5),main = "Normal Distribution vs. T Distribution", xlab = "x")
hist(rt(n=100000,df=5),breaks = seq(-50,50,0.1),
     col = scales::alpha("blue",0.5),add=T,xlim = c(-5,5))
```



From the histogram above, we can see that the tail of the t-distribution is fatter than that of normal distribution

Question 6

6a) From the Vanguard dataset, compute the standard error of the mean for the `VFIAX` index fund return.

```
library(DataAnalytics)
data("Vanguard")
return = Vanguard[Vanguard$ticker == "VFIAX","mret"]
std_err = function(x) {return(sd(x) / sqrt(length(x)))}
std_err(return)
```

```
## [1] 0.003670128
```

6b) For this fund, the mean and the standard error of the mean are almost exactly the same. Why is this a problem for a financial analyst who wants to assess the performance of this fund?

$$stderr(\bar{x}) = \frac{s_x}{\sqrt{n}} \approx \bar{x}$$
$$s_x \approx \bar{x} \times \sqrt{n}$$

The standard deviation of the fund is too large to estimate the performance of this fund.

6c)Calculate the size of the sample which would be required to reduce the standard error of the mean to 1/10th of the size of the mean return.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the population does not change with the changing of sample size, so to reduce the standard error of the mean to 1/10th of the mean value, we should increase the sample size.

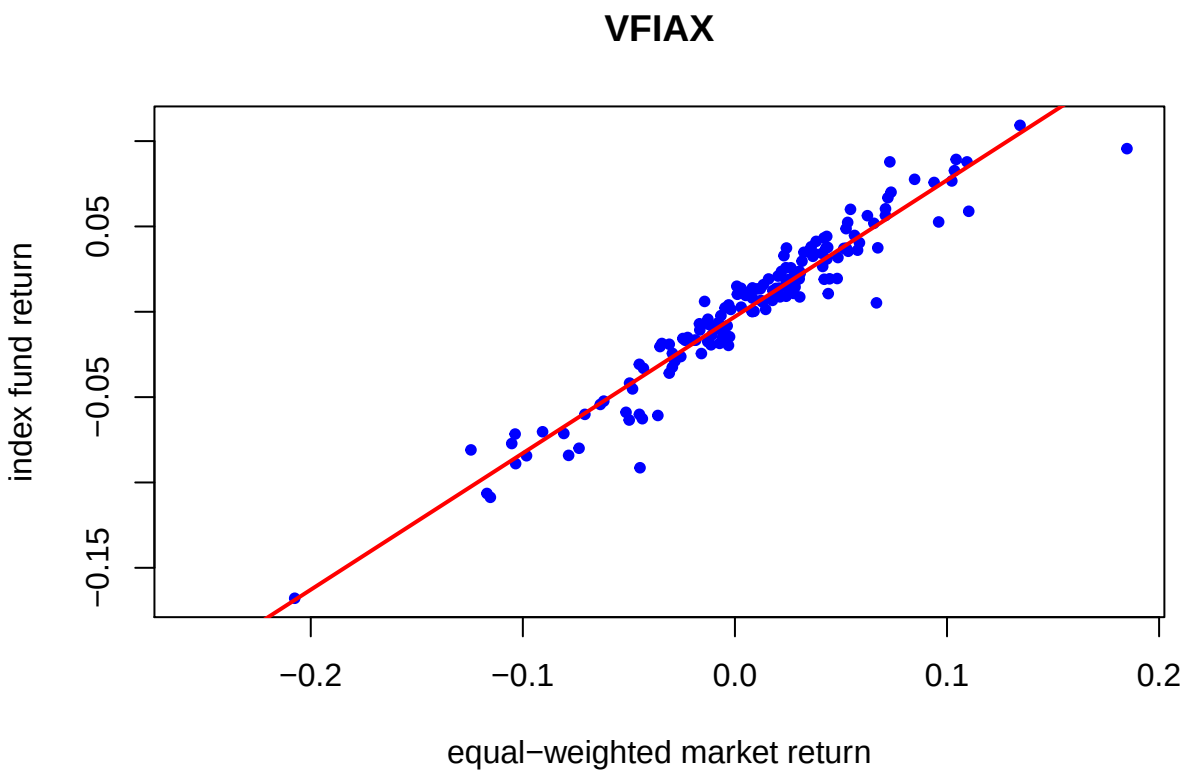
```
std_err_new = 0.1 * mean(return)
n = as.integer((sd(return) / std_err_new)^2)
n
```

```
## [1] 12970
```

Question 7

7a) Plot the `VFIAX` index fund return (as the Y variable) against the `ewretd` (equal-weighted market return, as the X variable) and add the fitted regression line to the plot. You might find the function `abline()` to be helpful.

```
data("marketRf")
library(reshape2)
Van = Vanguard[,c(1,2,5)]
V_reshaped = dcast(Van,date~ticker,value.var = "mret")
Van_mkt = merge(V_reshaped,marketRf,by="date")
with(Van_mkt, plot(ewretd,VFIAX,pch=20,col="blue",
                   main = "VFIAX",
                   ylab = "index fund return",
                   xlab = "equal-weighted market return"))
out = lm(VFIAX~ewretd,data = Van_mkt)
abline(out$coef,col="red",lwd=2)
```



7b) Provide the regression output using the `lmSumm()` function from the `DataAnalytics` package.

```
lmSumm(out)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 151 observations
##
## lm(formula = VFIAX ~ ewretd, data = Van_mkt)
##
## Coefficients:
##              Estimate Std Error t value p value
## (Intercept) -0.002855  0.001014   -2.82   0.006
## ewretd       0.799900  0.018520   43.19   0.000
## ---
## Standard Error of the Regression:  0.01231
## Multiple R-squared:  0.926  Adjusted R-squared:  0.926
## Overall F stat: 1865.32 on 1 and 149 DF, pvalue= 0
```