Problem Set 2
MFE 402: Econometrics
Professor Rossi

This problem set is designed to review material on the sampling distribution of least squares.

Question 1

a. Use the formulas for the least squares estimators to express the least squares intercept as a weighted sum (i.e., a linear combination) of the Y values in a similar way as is done in the lecture notes for the slope (see Ch1, pg 70–72).

$$
\begin{aligned}
b_1 &= \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} \\
b_0 &= \bar{Y} - b_i\bar{X} \\
\textbf{Suppose } D &= \sum(X_i - \bar{X})^2 = \sum(X_i - \bar{X})X_i - \bar{X}\sum(X_i - \bar{X}) \\
&= \sum X_i^2 - \bar{X}\sum X_i - \bar{X} \times 0 \\
&= \sum X_i^2 - N(\bar{X})^2 \\
b_0 &= \frac{1}{N}\sum Y_i - \frac{\bar{X}\sum(X_i - \bar{X})Y_i}{D} \\
&= \sum(\frac{1}{N} - \frac{X_i\bar{X} - \bar{X}^2}{D})Y_i \\
&= \sum(\frac{\frac{1}{N}\sum X_i^2 - \bar{X}^2 - X_i\bar{X} + \bar{X}^2}{D})Y_i \\
&= \sum(\frac{\frac{1}{N}\sum X_i^2 - X_i\bar{X}}{D})Y_i
\end{aligned}
$$

b. Use the formula in part (a) to show that b_0 is an unbiased estimator for β_0. That is, show $\mathbb{E}[b_0] = \beta_0$. You may not use the fact $\mathbb{E}[b_1] = \beta_1$ unless you explicitly prove it.

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
E(Y_i) &= \beta_0 + \beta_1 X_i + 0 \\
E(b_0) &= \sum(\frac{\frac{1}{N}\sum X_i^2 - X_i\bar{X}}{D})E(Y_i) \\
&= \sum(\frac{\frac{1}{N}\sum X_i^2 - X_i\bar{X}}{D})(\beta_0 + \beta_1 X_i) \\
&= \frac{1}{D}\sum(\frac{\beta_0}{N}\sum X_i^2 - \beta_0 X_i\bar{X} + \frac{\beta_1 X_i}{N}\sum X_i^2 - \beta_1 X_i^2\bar{X}) \\
&= \frac{1}{D}(\beta_0\sum X_i^2 - \beta_0 N\bar{X}^2 + \beta_1\bar{X}\sum X_i^2 - \beta_1\bar{X}\sum X_i^2) \\
&= \frac{\beta_0(\sum X_i^2 - N(\bar{X})^2)}{D} \\
&= \beta_0
\end{aligned}
$$

c. Use the formula in part (a) to show that $\text{Var}(b_0) = \sigma^2\left[\frac{1}{N} + \frac{\bar{X}^2}{(N-1)s_X^2}\right]$.

Note that parts (b) and (c) are somewhat challenging.

$$
\begin{aligned}
Var(Y_i|X_i) &= \sigma^2 \\
\textbf{let } c_i &= \frac{X_i - \bar{X}}{D} \\
\sum c_i &= \frac{\sum(X_i - \bar{X})}{D} = \frac{1}{D}\sum(X_i - \bar{X}) = 0 \\
\sum c_i^2 &= \sum \frac{(X_i - \bar{X})^2}{D^2} = \frac{1}{D^2}\sum(X_i - \bar{X})^2 = \frac{D}{D^2} = \frac{1}{D} \\
b_0 &= \sum(\frac{1}{N} - c_i\bar{X})Y_i \\
Var(b_0) &= Var(\sum(\frac{1}{N} - c_i\bar{X})Y_i|X) \\
&= \sum(\frac{1}{N} - c_i\bar{X})^2 Var(Y_i|X_i) \\
&= \sigma^2\sum(\frac{1}{N} - c_i\bar{X})^2 \\
&= \sigma^2\sum(\frac{1}{N^2} - \frac{2}{N}c_i\bar{X} + c_i^2\bar{X}^2) \\
&= \sigma^2(\frac{1}{N} - \frac{2}{N}\bar{X}\sum c_i + \bar{X}^2\sum c_i^2) \\
&= \sigma^2(\frac{1}{N} - 0 + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}) \\
&= \sigma^2[\frac{1}{N} + \frac{\bar{X}^2}{(N-1)s_x^2}]
\end{aligned}
$$

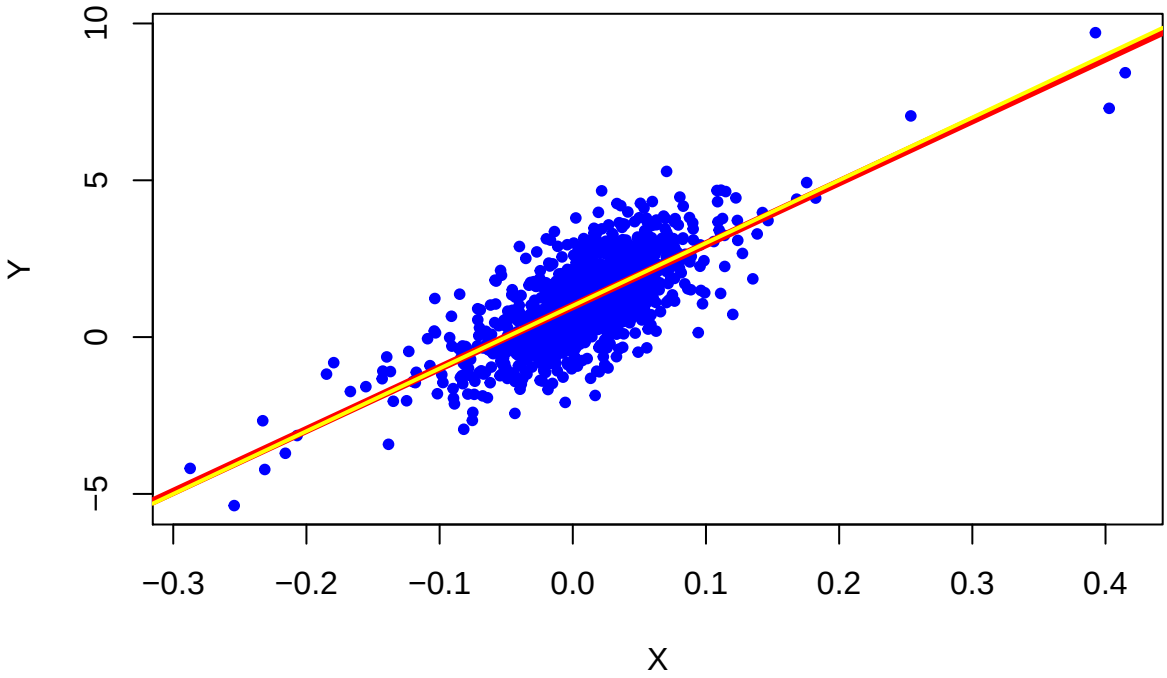
Question 2

a. Write a function in `R` (using `function()`) to simulate from a simple regression model. This function should accept as inputs: β_0 (intercept), β_1 (slope), X (a vector of values), and σ (error standard deviation). You will need to use `rnorm()` to simulate from the normal distribution. The function should return a vector of Y values.

```
regression = function(beta0,beta1,X,sd){
  error = rnorm(n = length(X),mean = 0,sd=sd)
  Y = beta0 + beta1 * X + error
  return(Y)
}
```

b. Simulate Y values from your function and make a scatterplot of X versus simulated Y. When simulating, use the `vwretd` data from the `marketRf` dataset as the X vector, and choose $\beta_0 = 1$, $\beta_1 = 20$, and $\sigma = 1$. Then add the fitted regression line to the plot as well as the true conditional mean line (the function `abline()` may be helpful).

```
library(DataAnalytics)
library(ggplot2)
data("marketRf")
X = marketRf[["vwretd"]]
Y = regression(beta0 = 1, beta1 = 20, X = X, sd = 1)
plot(X,Y,pch = 20,col = "blue")
abline(lm(Y~X), col="red",lwd=4)
abline(a=1,b=20, col = "yellow",lwd=2)
```



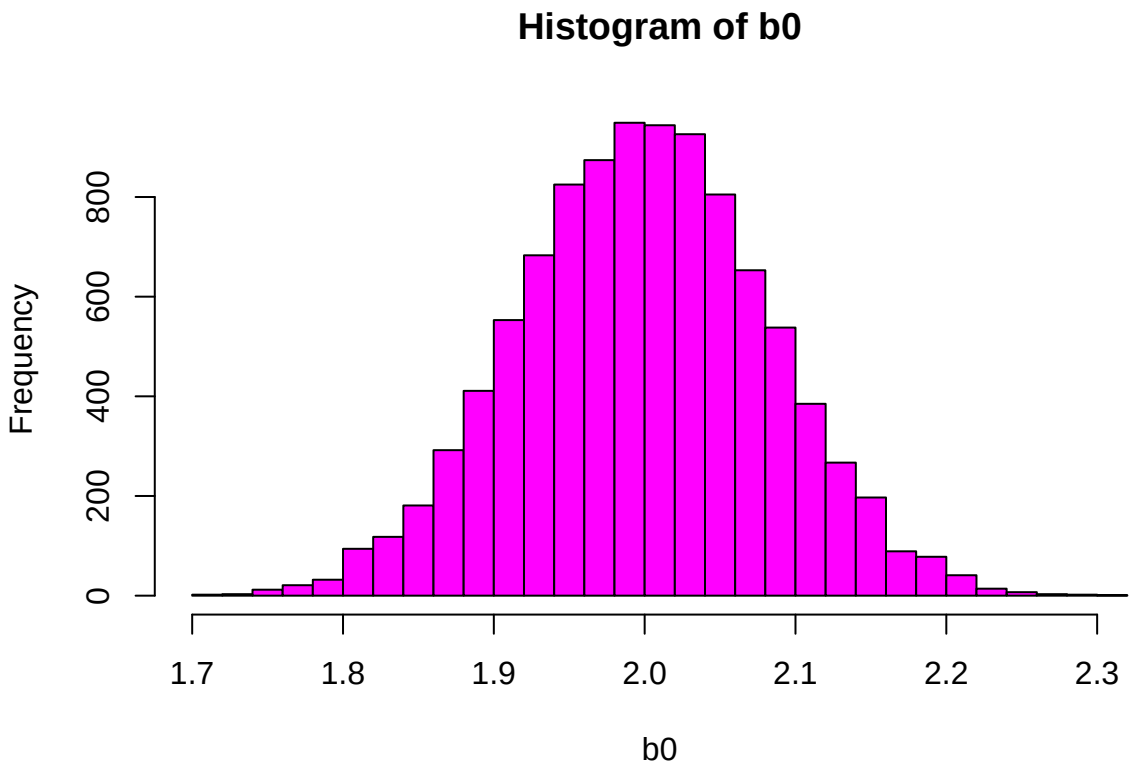
Question 3

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

a. Use your R function from question 1 to simulate the sampling distribution of the intercept. Use a sample size of 300 and calculate b_0 for 10,000 samples. Plot a histogram of the sampling distribution of b_0. You may find slide 75 of Chapter 1 of the course notes to be helpful.

```
simreg = function(beta0,beta1,sigma,x){
  y = beta0 + beta1 * x + rnorm(length(x),mean = 0, sd = sigma)
  return(y)
}
beta0 = 2
beta1 = 0.6
sigma = sqrt(2)
nsample = 10000
b0 = double(nsample)
x = rnorm(300)
for(i in 1:nsample){
  y = simreg(beta0 = beta0, beta1 = beta1, sigma = sigma, x = x)
  b0[i] = lm(y~x)$coef[1]
```

```
}
hist(b0,breaks = 40,col = 'magenta')
```



b. Calculate the empirical value for $\mathbb{E}[b_0]$ from your simulation and provide the theoretical value for $\mathbb{E}[b_0]$ (you might find question 1b to be helpful here). Compare the simulated and theoretical values.

```
expectation_b0 = mean(b0)
theoretical_b0 = beta0
cat("The simulated value of beta0 is",expectation_b0,
    ".\nAnd the theoretical value is ",theoretical_b0)
```

```
## The simulated value of beta0 is 1.998417 .
## And the theoretical value is  2
```

c. Calculate the empirical value for $\text{Var}(b_0)$ from your simulation and provide the theoretical value for $\text{Var}(b_0)$ (you might find question 1c to be helpful here). Compare the simulated and theoretical values.

```
empirical_var = var(b0)
empirical_var
```

```
## [1] 0.006790927
```

```
x_bar = mean(x)
sx_2 = var(x)
```

```
theoretical_var = sigma^2 * (1/300 + x_bar^2 / (300 - 1) / sx_2)
theoretical_var
```

```
## [1] 0.006666733
```

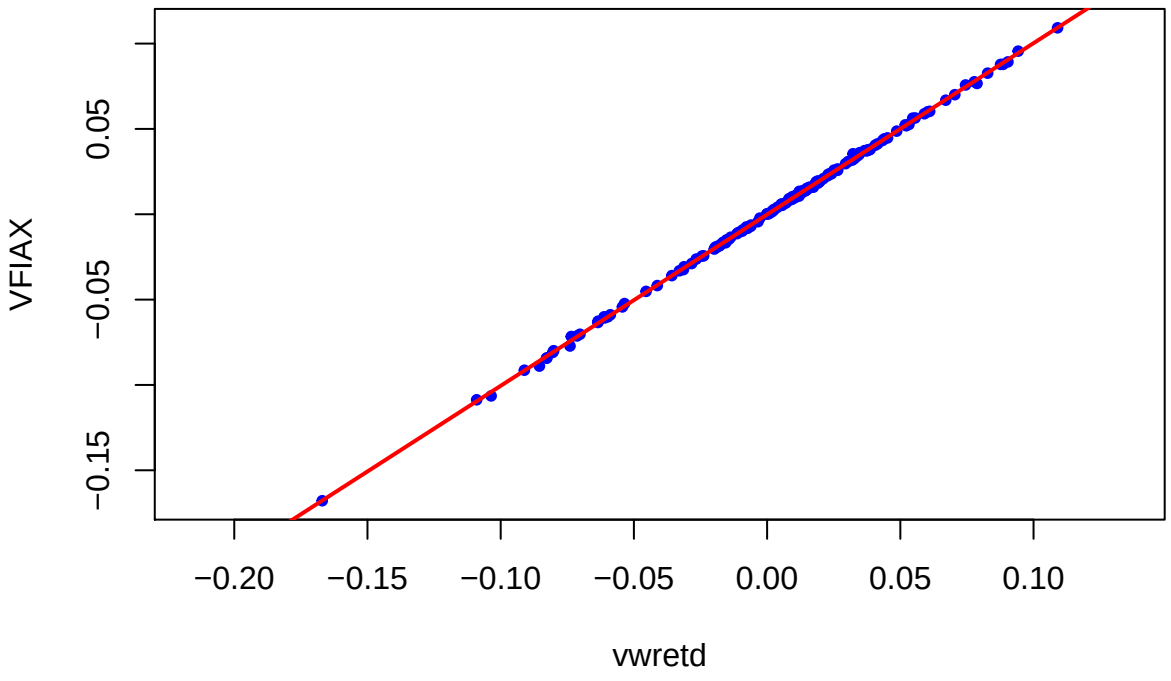
```
cat("The simulated variance of b0 is ", empirical_var,
    "\nWhile the theoretical value is ", theoretical_var)
```

```
## The simulated variance of b0 is  0.006790927
## While the theoretical value is  0.006666733
```

Question 4

Fit a regression of the Vanguard 500 Index Fund returns (VFIAX in the `Vanguard` dataset from the `DataAnalytics` package) on the `vwretd` series (from the `marketRF` dataset in the `DataAnalytics` package).

```
library(reshape2)
data("Vanguard")
data("marketRf")
Van = Vanguard[,c(1,2,5)]
V_reshaped=dcast(Van,date~ticker,value.var="mret")
Van_mkt=merge(V_reshaped,marketRf,by="date")
reg = lm(Van_mkt$VFIAX~Van_mkt$vwretd)
with(Van_mkt,
     plot(vwretd,VFIAX,pch=20,col="blue")
)
abline(reg,col="red",lwd=2)
```



a.

Test the hypothesis $H_0^a : \beta_1 = 1$ at the 0.05 level of significance using t-statistics. Report your decision (accept or reject the null hypothesis).

```r
estimated_beta = reg$coef[2]
std_error = lmSumm(reg)$coef[2,2]
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 151 observations
##
## lm(formula = Van_mkt$VFIAX ~ Van_mkt$vwretd)
##
## Coefficients:
##                   Estimate Std Error t value p value
## (Intercept)    -0.0001314 6.475e-05   -2.03   0.044
## Van_mkt$vwretd  1.0040000 1.440e-03  696.94   0.000
## ---
## Standard Error of the Regression:  0.0007924
## Multiple R-squared:  1  Adjusted R-squared:  1
## Overall F stat: 485730.9 on 1 and 149 DF, pvalue= 0
```

```r
t = (estimated_beta - 1) / std_error
critical_value = qt(0.025,df=length(which(!is.na(Van_mkt$VFIAX))) - 2)
cat("t =",t, "> critical value = ", abs(critical_value),
    "\nso we reject the null hypothesis: H0: beta1 = 1")
```

```
## t = 2.593507 > critical value =  1.976013
## so we reject the null hypothesis: H0: beta1 = 1
```

b. Test the hypothesis $H_0^b : \beta_0 = 0$ at the 0.01 level of significance using p-values. Report your decision (accept or reject the null hypothesis).

```r
estimated_beta0 = reg$coef[1]
std_error0 = lmSumm(reg)$coef[1,2]
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 151 observations
##
## lm(formula = Van_mkt$VFIAX ~ Van_mkt$vwretd)
##
## Coefficients:
##                   Estimate Std Error t value p value
## (Intercept)    -0.0001314 6.475e-05   -2.03   0.044
## Van_mkt$vwretd  1.0040000 1.440e-03  696.94   0.000
## ---
## Standard Error of the Regression:  0.0007924
## Multiple R-squared:  1  Adjusted R-squared:  1
## Overall F stat: 485730.9 on 1 and 149 DF, pvalue= 0
```

```r
t0 = (estimated_beta0 - 0)/std_error0
pvalue = 2 * pt(-abs(t0),df = length(which(!is.na(Van_mkt$VFIAX))) - 2)
cat("p =",pvalue,"> 0.01",
    "\nso we cannot reject at 0.01 level of significance")
```

```
## p = 0.04425341 > 0.01
## so we cannot reject at 0.01 level of significance
```

You may **not** use the `summary()` command or a similar command that "automatically" computes t and p values. You must compute the t and p values "by hand". You may, however, use `qt()`, `pt()`, or similar commands.

Question 5

Standard errors and p-values.

a. What is a standard error of a test statistic? How is a standard error different from a standard deviation?

Standard error measures how far the sample mean is away from the true population mean. While a standard deviation measures the dispersion of the sample data from its mean.

b. What is sampling error? How does the standard error capture sampling error?

Sampling error incures when the sample statistics are used to estimate the population statistics. It is the difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter.

Standard error measures how large the sampling error would be.

c. Your friend Steven is working as an investment analyst and comes to you with some output from some statistical method that you've never heard of. Steven tells you that the output has both parameter estimates and standard errors. He then asks, "how do I interpret and use the standard errors?" What do you say to Steven to help him even though you don't know what model is involved?

Standard errors is how far your estimated parameter away from the true parameter. The smaller the standard error, the better those parameters estimates can describe the true population parameters

d. Your friend Xingua works with Steven. She also needs help with her statistical model. Her output reports a test statistic and the p-value. Xingua has a Null Hypothesis and a significance level in mind, but she asks "how do I interpret and use this output?" What do you say to Xingua to help her even though you don't know what model is involved?

The p-value is the minimum significance level at which you can reject the null

If your p-value is larger than the significance level, you cannot reject the null hypothesis.

Question 6

Use the fitted regression of `VGHCX` (in the `Vanguard` dataset from the `DataAnalytics` package) on `vwretd` (from the `marketRF` dataset in the `DataAnalytics` package) to answer the following questions. You may not use the `predict()` command. You must perform the calculations "by hand". Note that the data has values like 0.003, which is a positive return of 0.3%. a. Compute an estimate of the conditional mean of the

Vanguard HCX fund's return given that the market is up by 5%.

```
Van = Vanguard[,c(1,2,5)]
V_reshaped=dcast(Van,date~ticker,value.var="mret")
Van_mkt=merge(V_reshaped,marketRf,by="date")
reg = lm(Van_mkt$VGHCX~Van_mkt$vwretd)
beta0 = reg$coef[1]
beta1 = reg$coef[2]
X_f = 0.05
HCX_return = beta0 + beta1 * X_f
HCX_return
```

```
## (Intercept)
##  0.04367826
```

b. Compute an estimate of the conditional standard deviation of the Vanguard HCX fund's return given that the market is up by 10%.

$$Var(Y|X=x) = Var(\epsilon) = \sigma_\epsilon^2$$
$$\sigma_\epsilon = \sigma_{Y|X}$$
$$\sigma_\epsilon \approx s = \sqrt{\frac{SSE}{N-2}}$$

```
estimated_Y = beta0 + beta1 * Van_mkt$vwretd
SSE = sum((estimated_Y - Van_mkt$VGHCX)^2)
sigma = sqrt(SSE / (length(which(!is.na(Van_mkt$VGHCX))) - 2))
sigma
```

```
## [1] 0.02504834
```

c. Compute an estimate of the prediction error (s_{pred}) for a prediction of the Vanguard HCX fund's return given that the market is up by 15%.

```
std_err = sigma
N = length(which(!is.na(Van_mkt$VGHCX)))
sx_2 = var(Van_mkt$vwretd)
s_pred = std_err * sqrt(1 + 1/N + 0.15^2/(N-1)/sx_2)
s_pred
```

```
## [1] 0.02548823
```