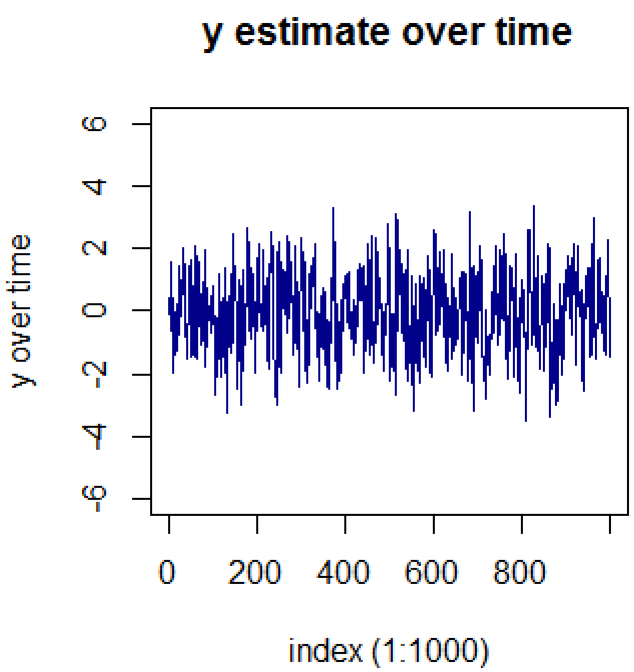
Assignment 5: Applied data science

ANSWERS:

1. Consider the AR(1) DGP presented in class: y_t=ρy_(t-1)+ϵ_t.
   a. For this exercise, set ρ=0.5. Generate this DGP using 1,000 Gaussian white-noise draws from N(0,1) by letting y1=ϵ1. Plot this DGP. Run a linear regression to get the least-squares estimate of ρ. (You should include a constant in the regression.) Does your 95% confidence interval include 0.5?

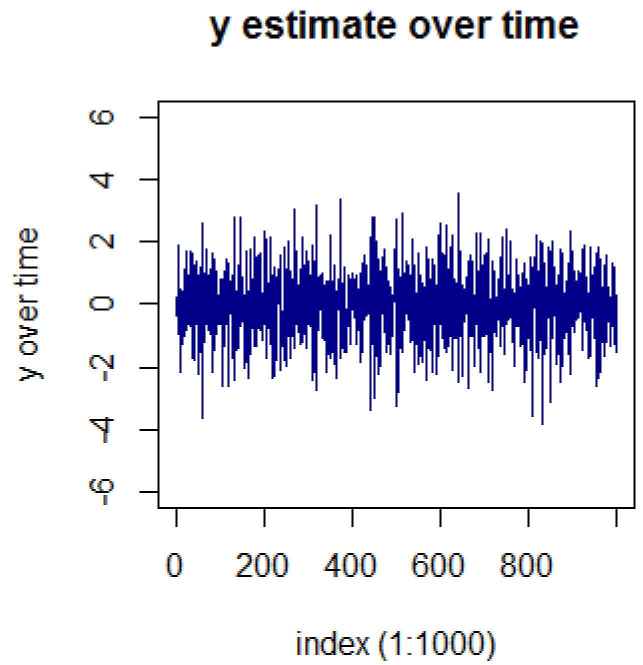



```
Call:
lm(formula = y[2:1000] ~ y[1:999])

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9222 -0.6388  0.0093  0.6259  3.1465 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.01994    0.03167   -0.63    0.529    
y[1:999]     0.54421    0.02658   20.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 997 degrees of freedom
Multiple R-squared:  0.296,	Adjusted R-squared:  0.2953 
F-statistic: 419.1 on 1 and 997 DF,  p-value: < 2.2e-16
```

```
> stargazer(model_linear, title="Linear regression results", typ
e="text", ci.level=0.95, ci=TRUE)

Linear regression results
===============================================
                        Dependent variable:
                    ---------------------------
                             y[2:1000]
-----------------------------------------------
y[1:999]                      0.544***
                           (0.492, 0.596)

Constant                       -0.020
                          (-0.082, 0.042)

-----------------------------------------------
Observations                    999
R2                             0.296
Adjusted R2                    0.295
Residual Std. Error      1.000 (df = 997)
F Statistic          419.139*** (df = 1; 997)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

We can see that 95% confidence interval include 0.5 (ranges between 0.492 and 0.596).

b. Repeat a. assuming ρ=-0.5.

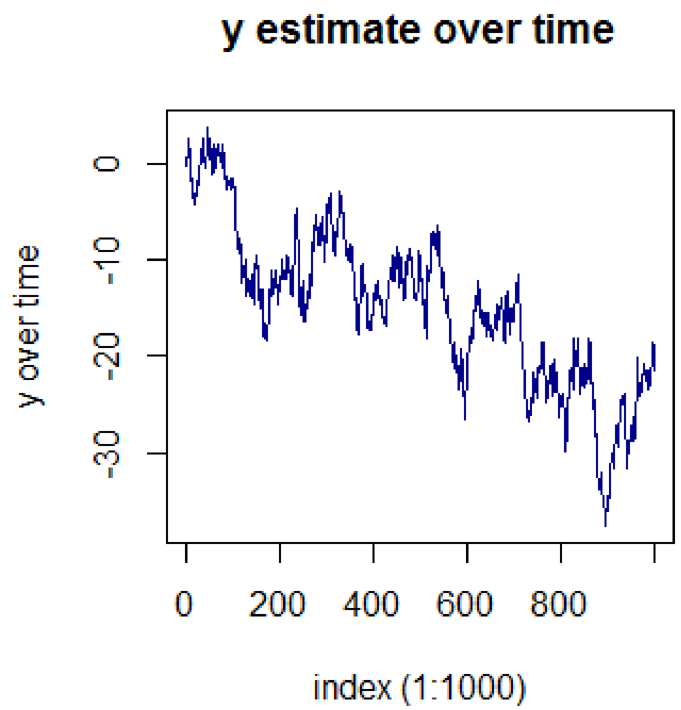



```
Call:
lm(formula = y[2:1000] ~ y[1:999])
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.9783 -0.6297  0.0019  0.6240  3.1791

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02123    0.03167   -0.67    0.503
y[1:999]    -0.46266    0.02808  -16.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 997 degrees of freedom
Multiple R-squared:  0.214, Adjusted R-squared:  0.2133
F-statistic: 271.5 on 1 and 997 DF,  p-value: < 2.2e-16

> stargazer(model_linear, title="Linear regression results", ty
pe="text", ci.level=0.95, ci=TRUE)

Linear regression results
===============================================
                        Dependent variable:
                    ---------------------------
                             y[2:1000]
-----------------------------------------------
y[1:999]                     -0.463***
                          (-0.518, -0.408)

Constant                      -0.021
                           (-0.083, 0.041)

-----------------------------------------------
Observations                    999
R2                             0.214
Adjusted R2                    0.213
Residual Std. Error      1.001 (df = 997)
F Statistic          271.522*** (df = 1; 997)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

We can see that 95% confidence interval includes -0.5 (ranges between -0.408 and -0.518).

c. Repeat a. assuming ρ=1. (This is called a random walk or unit root.)

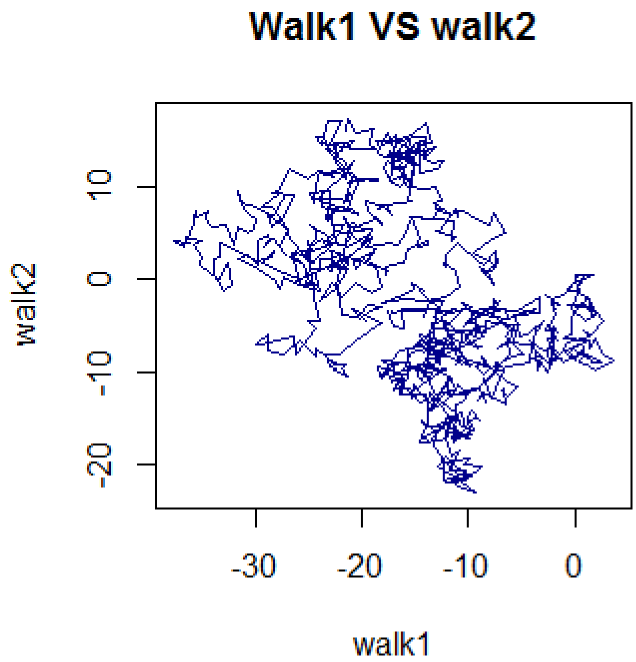



```
Call:
lm(formula = y[2:1000] ~ y[1:999])

Residuals:
     Min       1Q   Median       3Q      Max 
-3.03277 -0.62796  0.01084  0.61970  3.15372 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.151911   0.065431  -2.322   0.0204 *  
y[1:999]     0.991329   0.003816 259.765   <2e-16 ***
---
Signif. codes:  0 ‘**’ 0.001 ‘*’ 0.01 ‘ ’ 0.05 ‘ ’ 0.1 ‘ ’ 1

Residual standard error: 0.9992 on 997 degrees of freedom
Multiple R-squared:  0.9854,Adjusted R-squared:  0.9854 
F-statistic: 6.748e+04 on 1 and 997 DF,  p-value: < 2.2e-16

> stargazer(model_linear, title="Linear regression results", ty
pe="text", ci.level=0.95, ci=TRUE)

Linear regression results
===============================================
                        Dependent variable:    
                    ---------------------------
                             y[2:1000]         
     -----------------------------------------------
     y[1:999]                  0.991***        
                            (0.984, 0.999)     
```

```
Constant                      -0.152**
                          (-0.280, -0.024)

-----------------------------------------------
Observations                    999
R2                             0.985
Adjusted R2                    0.985
Residual Std. Error      0.999 (df = 997)
F Statistic         67,478.100*** (df = 1; 997)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Now the result shows interesting point. We can see that 95% confidence interval does not include 1 (ranges between 0.984 and 0.999). It shows that having ρ equal to 1 means that y[t] in respect to y[t-1] can be completely random (also shown in the graph above.)

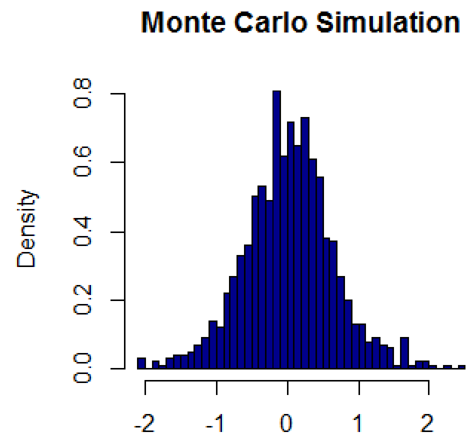
2. Consider the AR(1) DGP presented in class: y_t=ρy_(t-1)+ϵ_t.
   a. Following 1c. above, generate two independent random walks of 1,000 observations, calling them Walk1 and Walk2. Fit the bivariate linear model that relates Walk1 to Walk2 and report your regression results.




```
Call:
lm(formula = walk2 ~ walk1)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6905  -5.5889  -0.2076   5.8936  19.0279 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.18099    0.54254  -16.92   <2e-16 ***
walk1       -0.49803    0.03163  -15.74   <2e-16 ***
---
Signif. codes:  0 ·**·0.001 ·*·0.01 ··0.05 ··0.1 ··
1

Residual standard error: 8.286 on 998 degrees of freedom
Multiple R-squared:  0.1989, Adjusted R-squared:  0.1981 

F-statistic: 247.9 on 1 and 998 DF,  p-value: < 2.2e-16

> stargazer(model_linear, title="Linear regression resul
ts", type="text", ci.level=0.95, ci=TRUE)

Linear regression results
===============================================
```

```
                    Dependent variable:
                ---------------------------
                          walk2
-----------------------------------------------
walk1                   -0.498***
                     (-0.560, -0.436)

Constant                -9.181***
                     (-10.244, -8.118)

-----------------------------------------------
Observations              1,000
R2                        0.199
Adjusted R2               0.198
Residual Std. Error    8.286 (df = 998)
F Statistic         247.852*** (df = 1; 998)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

b. Recall the Monte Carlo simulation exercise in HW 4, Question 3c. Using a similar approach, repeat a. above 1,000 times, each time recording the estimated value of the slope coefficient of the bivariate regression. Generate a histogram of your 1,000 replications. How do these results compare to those you found in HW 4, Question 3c?




```
> mean(betas)
[1] 0.02731522
> sd(betas)
[1] 0.6327763
```

We could see from the histogram above that the distribution of beta (line ar regression) of each random walk iteration forms another normal distribu tion at 0 as its center (mean = 0.027 and standard deviation = 0.632).

c. Consider your results in 2b as well as the dispersion in your histogram. Is there something about independent unit roots that may lead one to find correlation when there is none?

We could see that both histograms are centered to 0, meaning it shows that the beta is 0 which means that although in single iteration it might show that it is related, the monte carlo simulation confirmed that the it is not related.

In the case of equation used in this assignment, I think independent unit roots being introduced cannot add more correlations as long as it is constant value. However the value could introduce correlation if Ro is dependent to time element. It is also possible to have “drifts” in addition to random noise, and trends which might not be independent to time element, for instance:

$$y_t = \hat{\alpha} + \hat{\gamma}t + \hat{\beta}y_{t-1} + \hat{e}_t$$

In this case, y might have some correlation to y(t-1).