P-Set 1 Adv Econometrics

Ethan Grant uni: erg2145

October 6, 2016

1. (a)

average value for each vector comes out to:
E = 0.014003698, 0.054311176, 0.004744255, -0.046776593,0.044269689
This is in line with what i predicted as all the values are close to zero.

Var(X) =
[,1] [,2] [,3] [,4] [,5]
[1,] 0.9177694 0.9598261 1.017785 1.017376 1.017693
[2,] 0.9598261 1.9786326 2.010952 2.035977 2.079679
[3,] 1.0177846 2.0109515 2.989345 3.007466 3.063471
[4,] 1.0173761 2.0359771 3.007466 4.045282 4.071903
[5,] 1.0176934 2.0796790 3.063471 4.071903 5.147865

this is also what I predicted as variance covariance matrix increases as you move toward bottom element

(b) see code for implementation

(c) Var(reg$betas_n_1000)
[,1] [,2] [,3] [,4] [,5]
[1,] 4.492060e-03 -2.109167e-03 -3.271328e-04 9.6336e-05 2.7132e-05
[2,] -2.109167e-03 4.305546e-03 -2.177376e-03 3.7844e-05 -4.2901e-05
[3,] -3.271328e-04 -2.177376e-03 4.299288e-03 -1.9269e-03 -4.9570e-05
[4,] 9.633620e-05 3.784497e-05 -1.926927e-03 3.8629e-03 -1.9412e-03
[5,] 2.713261e-05 -4.290154e-05 -4.957046e-05 -1.9412e-03 1.9570e-03

σ^2=2 as defined by creation of variables

thus $2 * (X'X)^{-1}$=
[,1] [,2] [,3] [,4] [,5]
[1,] 4.451362e-03 -1.955891e-03 -2.560439e-04 -4.1618e-05 9.5782e-05
[2,] -1.955891e-03 4.0460e-03 -1.9976e-03 1.866394e-05 -7.537309e-05
[3,] -2.560439e-04 -1.997684e-03 4.0690e-03 -1.8713e-03 -8.352393e-05
[4,] -4.161810e-05 1.866394e-05 -1.8713e-03 3.7301e-03 -1.833926e-03
[5,] 9.578261e-05 -7.537309e-05 -8.352393e-05 -1.8339e-03 1.8992e-03

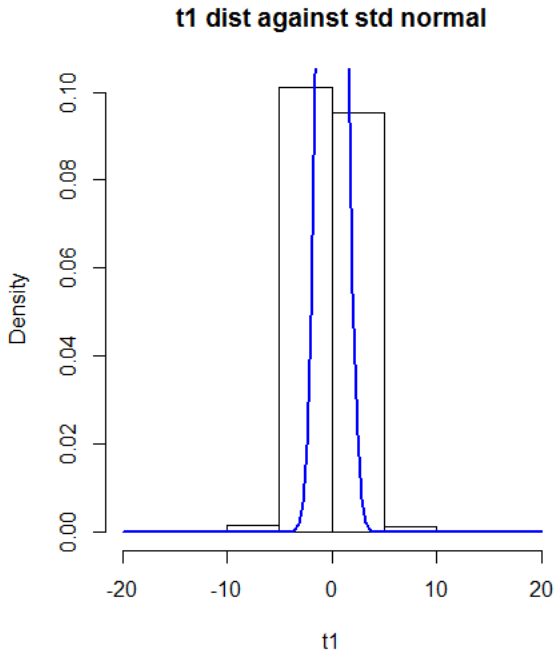
These values are very similar as is expected

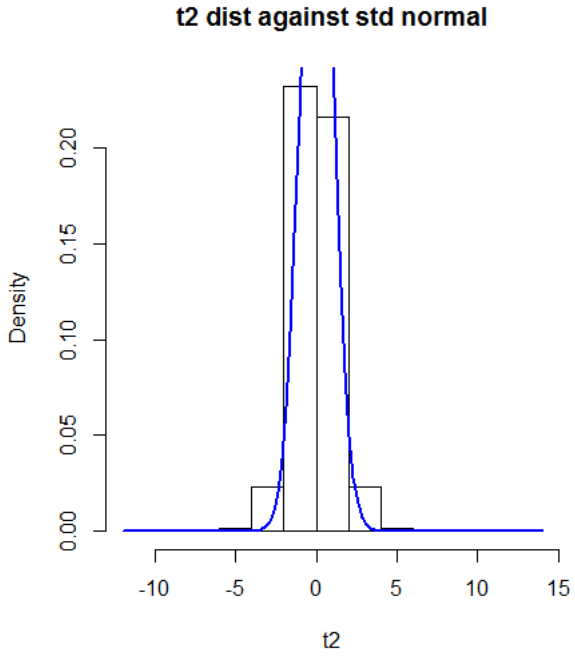
(d) N=10
mean($\hat{\sigma}^2_{1,b}$)= 0.9895263
mean($\hat{\sigma}^2_{2,b}$)= 1.979053
N=20
mean($\hat{\sigma}^2_{1,b}$)= 1.4700532
mean($\hat{\sigma}^2_{2,b}$)= 1.960071
N=100
mean($\hat{\sigma}^2_{1,b}$)= 1.8840889
mean($\hat{\sigma}^2_{2,b}$)= 1.983251
N=1000

mean($\hat{\sigma}^2_{1,b}$)= 1.9884902
mean($\hat{\sigma}^2_{2,b}$)= 1.998483

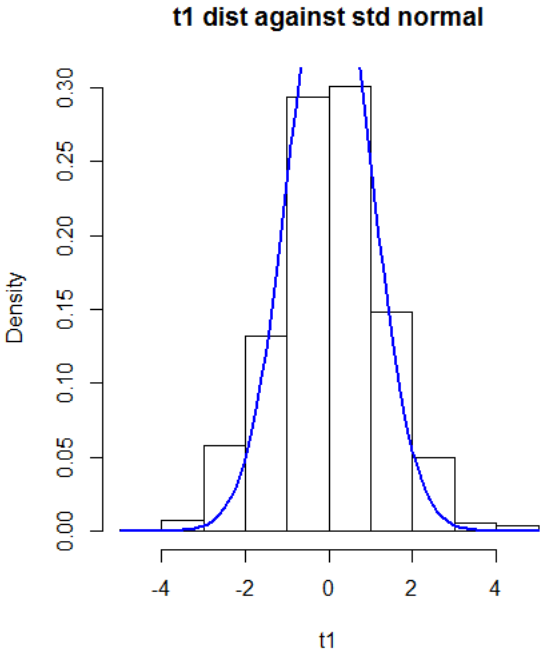
Both formulas eventually converge upon the true average (2), but the adjusted sigma that accounts for bias ($\hat{\sigma}^2_{2,b}$) gets close to the true value even when n=10 while the other is very far away at that point

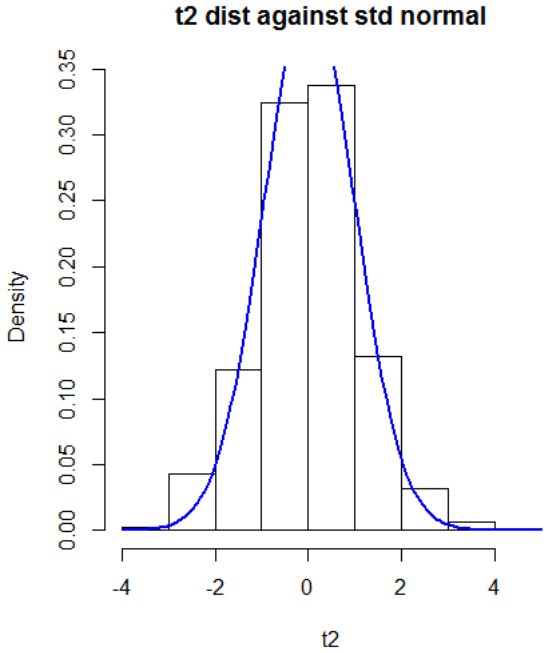
(e) n=10



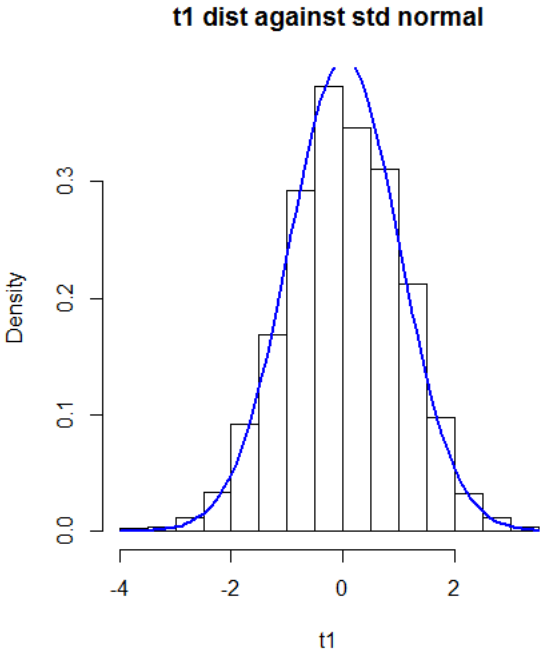

t2 dist against std normal
Density
0.00
0.05
0.10
0.15
0.20
-10
-5
0
5
10
15
t2

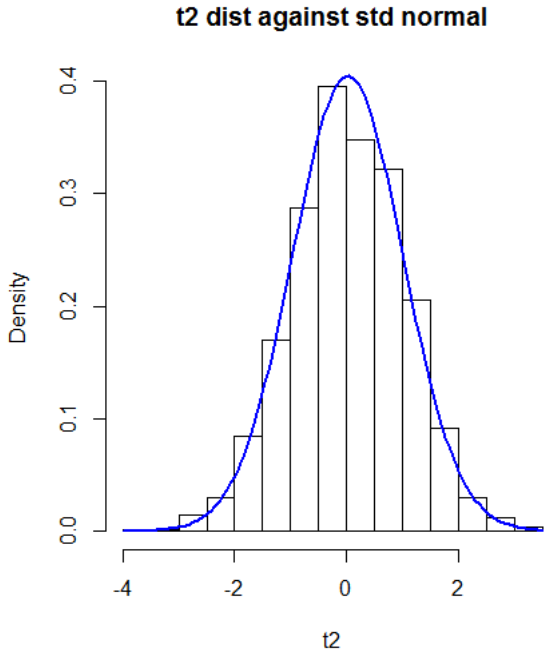
n=20


t1 dist against std normal
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-4
-2
0
2
4
t1

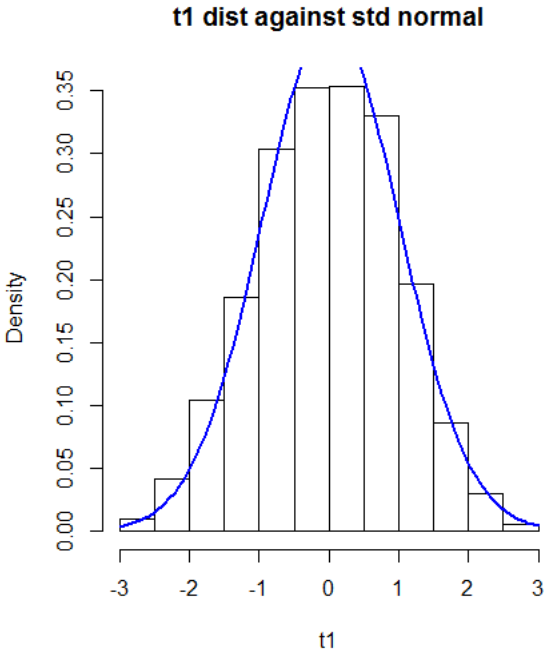

t2 dist against std normal
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
-4
-2
0
2
4
t2

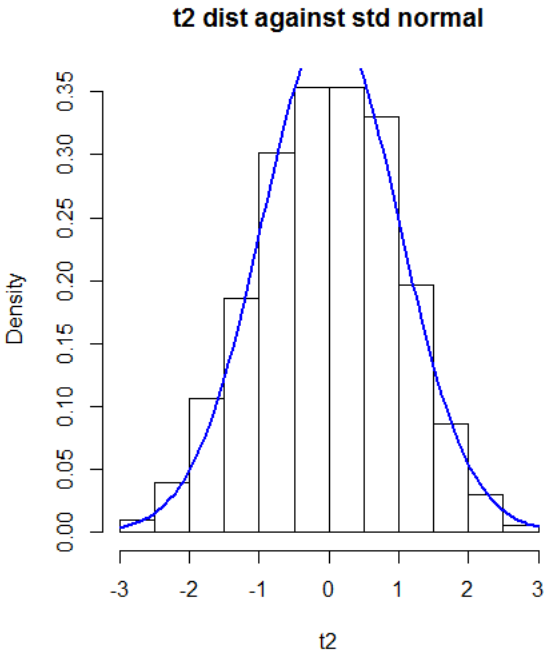
n=100





n=1000

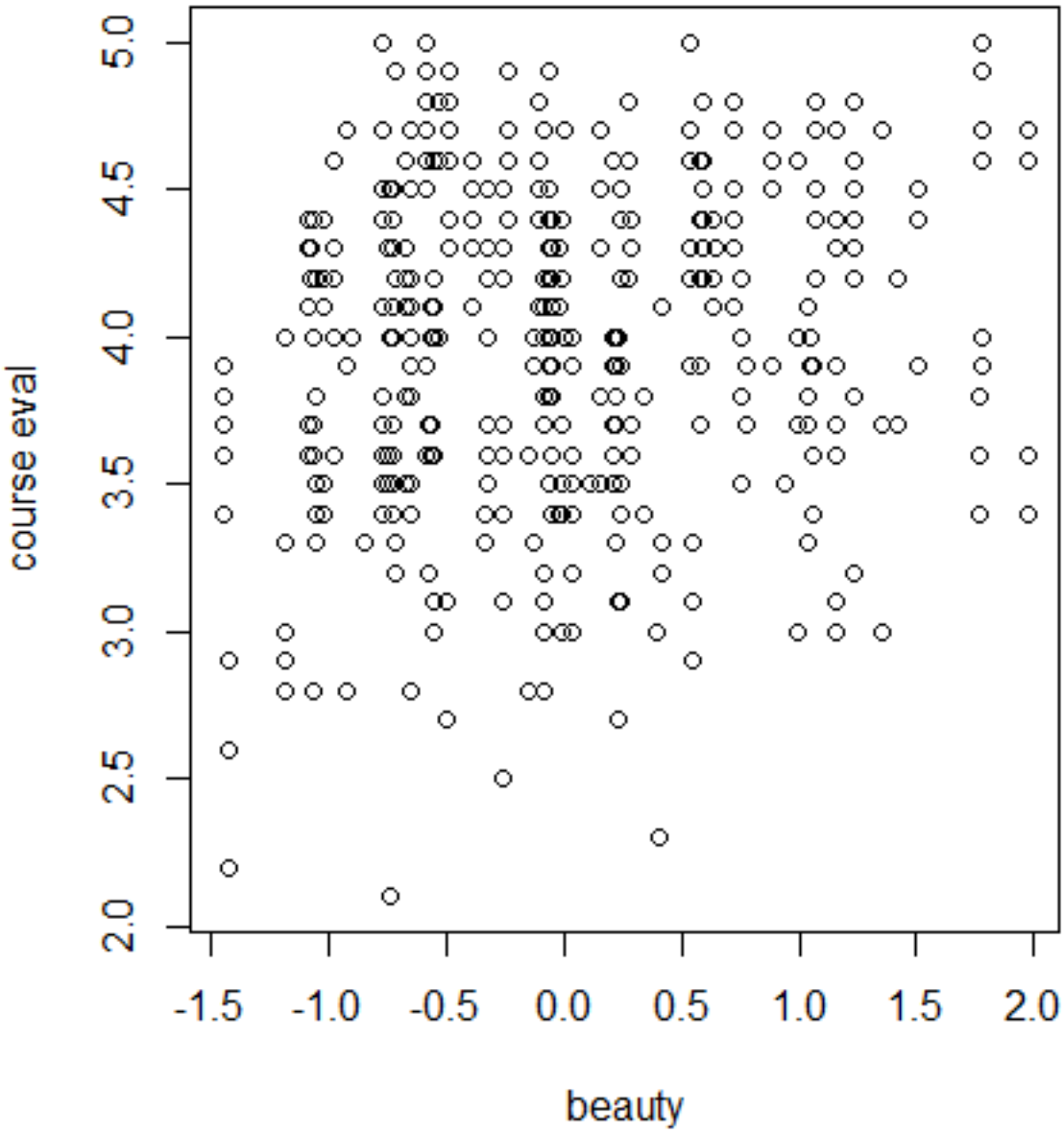

t1 dist against std normal
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
-3
-2
-1
0
1
2
3
t1


t2 dist against std normal
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
-3
-2
-1
0
1
2
3
t2

N=10 reject .389
N=20 reject .582
N=100 reject .997
N=1000 reject 1

As N increases the test rejects more and more which is what you would want as the null hypothesis that $\beta_0 = 0$ is false. I conclude that the test is consistent

2. (a) see code
   (b) beta_hat = [3.98,.134,.00029]
   mean(residual) = 5.127 e-15
   var(residual) = .297
   mean(residual_from_maker) = 3.68 e -15
   sd(residual_from_maker) = .544


course eval
beauty
2.0
2.5
3.0
3.5
4.0
4.5
5.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

(c) FW estimates = [.134,.00029]

The OLS estimates (reported in part ii) are the same as the FW estimates though the FW estimates do not include the intercept part of the vector.

(d) variance covariance matrix:
intercept beauty age
intercept 0.0178836140 -1.319840e-03 -3.564487e-04
beauty -0.0013198398 1.138662e-03 2.728914e-05
age -0.0003564487 2.728914e-05 7.369971e-06

CI beta1 = [.0679,.2002]
CI beta2 = [-.005,.0056]

These seem like reasonable confidence intervals and both are reasonably tight around the original value meaning the estimates are decently close to the true value

(e) uncentered_r_squared = .981
centered_r_squared = .0357

The uncentered r squared is very close to 1 suggesting that the regression we have is a very good fit, while the centered r squared is very close to 0 suggesting we have a bad fit. This difference makes it important which one we prefer

I prefer the centered R squared because it is based on the assumption that there is some intercept not equal to zero which looking at the data seems very reasonable. Additionally looking at the scatterplot it does not seem like the variation is explained very heavily by the regressors indicating that such a high r squared is not accurate

(f) mean(residuals) = 3.68 e-15
cor(residual, X[,'beauty']) = -2.92 e-15
cor(residual, X[, 'age']) = -6.42 e-15

These do make sense given what we have seen in class b/c our assumption is that E[u—x] = 0 which means that the mean of the residuals should be zero which we are very close to achieving here. Additionally the correlation being very close to zero also makes sense because if there was a correlation that implies that there is more information about that error (and thus y) that is contained in either of the X values that is not currently being used. This goes against our assumptions and the math so we want these values to be very very close to zero which they are.