Part I. (OLS) Regression, Sampling Distributions, and Hypothesis Testing (20 points)

There are 30,000 students enrolled in King University (KU). On average, KU students take 5.5 years to graduate (get a bachelor's degree). The years in college until graduation are normally distributed with the standard deviation of 0.8.

Mean=5.5
SD=0.8
N=30000

We don't know the average years in college for KU students majoring in political science, but we have a random sample of years in college for 16 KU political science students. The sample mean is 5.8 years.

N=16
Mean= 5.8

Can we state that KU political science students take longer time to graduate than does an average KU student?

Given Information:

Part I. (OLS) Regression, Sampling Distributions, and Hypothesis Testing (20 points)

There are 30,000 students enrolled in King University (KU). On average, KU students take 5.5 years to graduate (get a bachelor's degree). The years in college until graduation are normally distributed with the standard deviation of 0.8.

We don't know the average years in college for KU students majoring in political science, but we have a random sample of years in college for 16 KU political science students. The sample mean is 5.8 years.

Can we state that KU political science students take longer time to graduate than does an

average KU student?

1. Write a set of hypotheses (null and alternative hypotheses) to answer the question.

Null Hypothesis

H0= μ>5.5

2. Test your hypotheses.

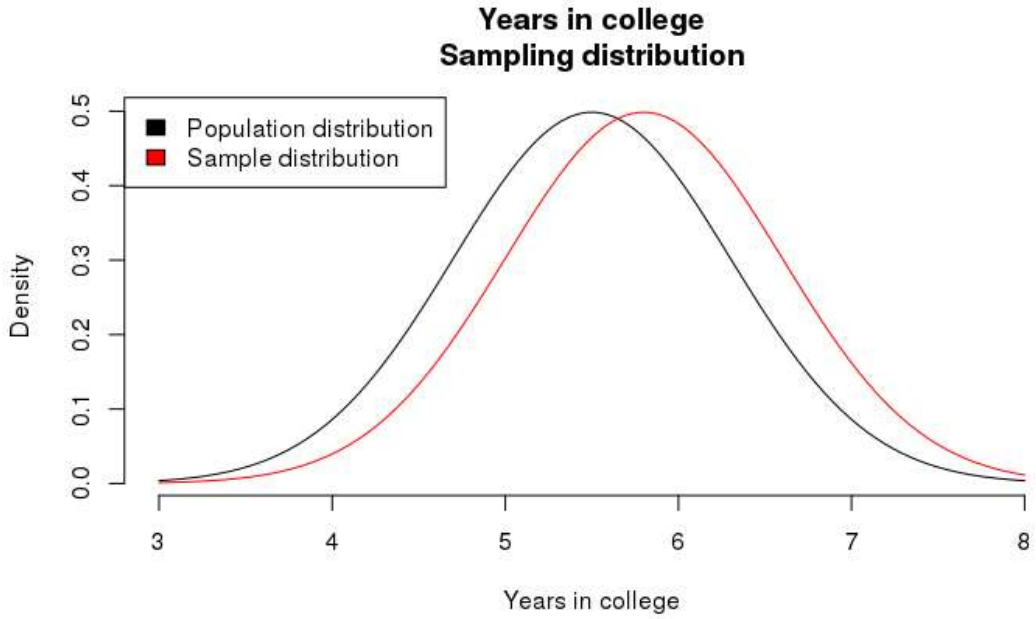
Solution:
Xbar= 5.5
.μ= 5.8
N=16
P(X>5.5)
Standard Normal Variate
Z= Xbar- μ/σ
Z= (5.5-5.8)/0.8
Z=-0.3/0.8
Z=-0.375
P(X>5.5)= 1-P(X<5.5)
P(Z>5.5)= 1-P(Z<5.5)
P(Z>5.5)= 1-(-0.375)= 1.375
P(Z>5.5)= 0.9154
Hence associated probability is 0.9154, hence there is no sufficient evidence that KU political science students take longer time to graduate than does an average KU student.

2. Test your hypotheses.

2.1. Show a sampling distribution of the years in college for KU students using R. Submit your R commands, as well.

```
x <- seq(3, 8, length.out = 100)
y1 <- dnorm(x, 5.5, 0.8)
y2 <- dnorm(x, 5.8, 0.8)
plot(y1 ~ x, type = "l", ylab = "Density",
     xlab = "Years in college",
     main = "Years in college\nSampling distribution",
     axes = FALSE)
lines(y2 ~ x, col = 2)
legend("topleft", fill = 1:2, legend = c("Population distribution", "Sample
```

```
distribution"))
axis(1)
axis(2)
```



2.2. What is the probability of a KU student taking longer than 5.8 years to graduate?

```
Prob <- 1- pnorm(5.8, 5.5, 0.8)
```

the probability of a KU student taking longer than 5.8 years to graduate is 0.3538302

2.3. Report the result of your hypothesis testing.

For test my hypotesis I use the formula:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

where z has a normal strandar distribution.

```
z = (5.8 - 5.5)/(0.8/sqrt(16))
p_value = 1 - pnorm(z, 15)
```

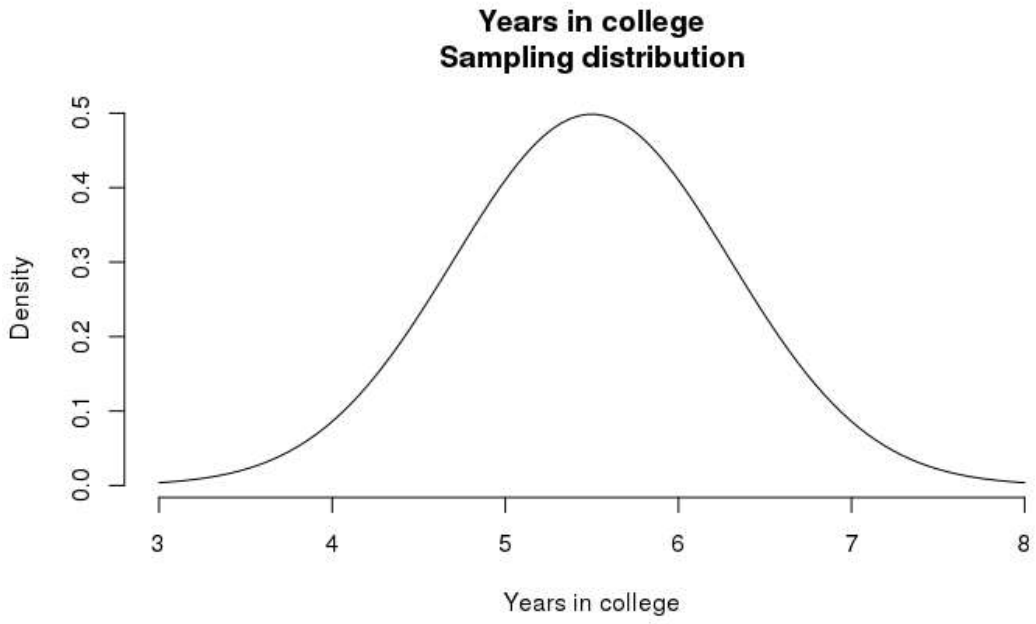
the p value of my hypotesis testing is 1. So, whith a significance level $\alpha = 0.05$ I cant find evidence of a statistically signican diference betwen the average years in college of political science students and the average years in college of all students.

The sample size for the years in college for KU political science students is increased from 16 to 25 (the sample mean remains the same).

3. Test your hypotheses again.

3.1. Show a sampling distribution of the years in college for KU students using R. Submit your R commands, as well. Compare this distribution to that of #2.1.

```
x <- seq(3, 8, length.out = 100)
y <- dnorm(x, 5.5, 0.8)
plot(y ~ x, type = "l", ylab = "Density",
     xlab = "Years in college",
     main = "Years in college\nSampling distribution",
     axes = FALSE)
axis(1)
axis(2)
```



3.2. What is the probability of a KU student taking longer than 5.8 years to graduate?

```
Prob <- 1- pnorm(5.8, 5.5, 0.8)
```

the probability of a KU student taking longer than 5.8 years to graduate is 0.3538302

3.3. Report the result of your hypothesis testing.

For test my hypotesis I use the formula:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

where z has a normal strandar distribution.

```
z = (5.8 - 5.5)/(0.8/sqrt(25))
p_value = 1 - pnorm(z, 1)
```

the p value of my hypotesis testing is 0.190787. So, whith a significance level $\alpha = 0.05$ I cant find evidence of a statistically signican diference betwen the average years in college of political science students and the average years in college of all students.

We wanted to know the impact of a student's first year college GPA on her years in college, and thus collected first year GPAs from 81 KU students randomly. When the first year GPAs are regressed on the years in college, the slope is –0.19 with the standard error of 0.09.

4. Interpret the slope coefficient.

Slope is -0.19 which is negative meaning by GPA and years in college is negatively related with each other for one unit increase in Year GPA fall by 0.19 points.

5. Can we state that a student's first year college GPA can be a good predictor for that student's years in college?

As there is fall in GPA score by 0.19 points hence we can not say that Students first year GPA can be good predictor for that students years in college.

5.1. Write a set of hypotheses (null and alternative hypotheses) to answer the question.

Null Hypothesis:
H0= β=0
Alternative Hypothesis:
H1:=β≠0

5.2. Test your hypotheses, and report the result.

Test Statistic is

T= β/SE(β)
T= -0.19/0.09 with (n-2) df
T=-1.49

At 5% level of significance probability is 1.990 hence cal value is less than tabulated value we fail to reject the null hypothesis and conclude that β(slope) =0.

Part II. Logit Models (10 points)

Investigate the relationship between trust in government and the likelihood of voting for a third party candidate using a data set in “trust and vote.xlsx” file.

1. Use as many relevant statistical models as possible, and compare the results carefully.

Logistic Regression Model:

Variable coding

Trust	Coding
Low	0
Medium	1
High	2
Vote Third Party	
Liberal	0
Third Party	1

Log(P/1-P) = β0+β1*(Vote)

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	194.530[a]	.040	.054
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.			

Classification Table[a]					
	Observed		Predicted		
			Vote Third Party		Percentage Correct
			.00	1.00	
Step 1	Vote Third Party	.00	99	0	100.0
		1.00	55	0	.0
	Overall Percentage				64.3
a. The cut value is .500					

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Trust			5.849	2	.054	
	Trust(1)	1.022	.426	5.746	1	.017	2.778
	Trust(2)	.727	.445	2.665	1	.103	2.069
	Constant	-1.204	.329	13.380	1	.000	.300
a. Variable(s) entered on step 1: Trust.							

$\text{Log}(P/1-P) = \beta 0+\beta 1*(\text{Vote})$

$\text{Log}(P/1-P)= -1.204+1.022(\text{Trust}_{(1)})+0.727\ (\text{Trust}_{(2)})$

Model identifies the 64.3% cases correctly. Nagelkerke R Square is 5.4%, With reference to low trust odds ration of medium is 2.778, meaning by as compared to low trust medium trust candidate has 2.77 times more chance of liberal. But high trust has 2.07 times more chance of liberal when compared with low trust.

2. Make a table of predicted probabilities of voting for a third party candidate, based on the best model that you choose. Explain what the table shows.

Predicted Probabilities:

Record	Trust	Predicted Probabilities
1-100 Record	0(Low)	0.45455
101-200 Record	1(Medium)	0.38298
201-300 Record	2(High)	0.23077

Predicted probability of voting for a third party candidate is decreasing with the increase in unit Trust.