UNC
KENAN-FLAGLER
BUSINESS SCHOOL

BUSI 410 MIDTERM EXAM – Fall 2023

This exam consists of three short cases:

Case	Pages	Points
BLACK PANTHER	2-6	40
PINEAPPLE	7-11	30
UNC SALARY EQUITY STUDY	12-16	30
	Total	100

Answer the questions after each case. Show all of your work on the exam paper. The cases/questions are **not** ordered in difficulty. If you are stuck, try the next question!

This exam is open book, note, and laptop. However, **no communication** of any kind is allowed—verbal, written, or electronic. **Disable** your laptop's Wi-Fi/Internet connectivity. **Turn off and put away** your cell phone.

Name_____
Please Print

Section (**circle start time**): 9:30 11:00 12:30 2:00

Honor Code Pledge: *I will neither give nor receive unauthorized aid during this exam.*

Signature_____

******** DO NOT TURN TO THE NEXT PAGE UNTIL YOU ARE ASKED TO ********

Black Panther

After stepping down from the throne, T'Challa has decided to open an eco-tourism lodge named after his father in a remote section of Wakanda. He advertises that guests who visit T'Chaka's Lodge can see exotic flora and fauna, perhaps even including a Black Panther, while enjoying 5-star accommodations that are powered primarily by renewable resources.

T'Challa installed a wind turbine at the rim of the vibranium mine a few years back. Even though he likes the wind, he is now considering adding solar panels to the roof of the lodge in the hopes of taking T'Chaka's Lodge off-grid. Please help him by answering the following questions.

	A	B	C	D
1	Day	T'Challa Wind (KW)	T'Chaka Usage (KW)	Nakia Output (KW)
2	1	5230	5540	6850
3	2	7140	5320	7210
39	38	3850	4850	4490
40	39	4140	4920	5720
41	40	6150	6240	6150
42	41	2710	5420	
75	74	4850	4640	
76	75	5710	4210	
77				
78	**Average**	5140	5270	6475
79	**Std. Deviation**	2140	1230	1780

a) [4 pts] In a random sample of 75 days, T'Challa's wind turbine has provided enough energy to power the whole lodge on 20 of the days. What would be the conservative, approximate 95% confidence interval estimating the true proportion of days in which the wind turbine provides enough energy for the whole lodge?

T'Challa's friend Nakia has a solar array on her roof and was kind enough to share data on the daily power output of her solar array for a random sample of 40 days. T'Challa typed the 40 numbers into Microsoft Excel and calculated their mean, 6475 kilowatts, and their standard deviation, 1780 kilowatts. He would like to use this data to perform a hypothesis test (at the 5% significance level) of whether the mean daily power output is significantly less than 7000 kilowatts.

b) [2 pts] What is the hypothesis test he would like to perform?

1. $H_a: \mu_{Nakia} \geq 7000$ versus; $H_0: \mu_{Nakia} < 7000$
2. $H_a: \mu_{Nakia} \neq 7000$ versus; $H_0: \mu_{Nakia} = 7000$
3. $H_a: \mu_{Nakia} > 7000$ versus; $H_0: \mu_{Nakia} \leq 7000$
4. $H_a: \mu_{Nakia} < 7000$ versus; $H_0: \mu_{Nakia} \geq 7000$

Base your answers to the next four questions on the hypothesis test described above:

c) [2 pts] What is the value of the standard error to use for the test?

d) [2 pts] What is the p-value for the test?

e) [1 pt] Given the test-statistic and p-value, should you reject or not reject the null hypothesis?
1. Reject
2. Not Reject

f) [3 pts] Provide a one-sentence interpretation of your conclusion in terms of what it implies about the power output of Nakia's solar array.

g) [5 pts] Using the same sample, what is the exact 90% confidence interval for the average daily power output from Nakia's solar array?

h) [2 pts] T'Challa has calculated the mean of his sample of power usage at T'Chaka's Lodge to be 5270 kilowatts. He would like to test whether the power output of Nakia's solar array would be enough to significantly exceed the power usage of T'Chaka's Lodge on average. Which of the following is the most accurate expression of what T'Challa would like to test?

1. H_a: $\bar{X}_{\text{T'ChakaUsage}} = 5270$ versus H_0: $\bar{X}_{\text{NakiaOutput}} = 6475$
2. H_a: $\mu_{\text{NakiaOutput}} > 5270$ versus H_0: $\mu_{\text{NakiaOutput}} \leq 5270$
3. H_a: $\mu_{\text{NakiaOutput}} - \mu_{\text{T'ChakaUsage}} > 0$ versus H_0: $\mu_{\text{NakiaOutput}} - \mu_{\text{T'ChakaUsage}} \leq 0$
4. H_a: $\mu_{\text{NakiaOutput}} - \mu_{\text{T'ChakaUsage}} > 5270$ versus H_0: $\mu_{\text{NakiaOutput}} - \mu_{\text{T'ChakaUsage}} \leq 5270$
5. H_a: $\mu_{\text{NakiaOutput}} - \mu_{\text{T'ChakaUsage}} \neq 6475 - 5270$ versus H_0: $\mu_{\text{NakiaOutput}} - \mu_{\text{T'ChakaUsage}} = 6475 - 5270$

i) [2 pts] Which Excel formula should T'Challa use to calculate the p-value for this hypothesis test?

a. =T.Test(B2:B76,C2:C76,1,3)
b. =T.Test(B2:B76,C2:C76,1,1)
c. =T.Test(C2:C76,D2:D76,1,1)
d. =T.Test(C2:C76,D2:D41,1,1)
e. =T.Test(C2:C76,D2:D41,1,3)

j) [5 pts] T'Challa would like to create an approximate 95% confidence interval for the difference between the average power output of Nakia's solar array and the average power usage of T'Chaka's Lodge. Please indicate where this confidence interval is centered and its margin of error (i.e., $X \pm Y$).

k) [2 pts] T'Challa plans to take a new sample to determine a confidence interval for the true average daily power generated by Nakia's solar array. The confidence interval will be (choose one):

1. Shorter for 98% confidence than for 99% confidence.
2. Shorter for a sample of size 40 than for a sample of size 80
3. Shorter when the sample standard deviation s is large than when s is small.
4. None of the other options

l) [5 pts] What is the "99% confidence upper bound" for the true average power usage of T'Chaka's Lodge? In other words, can you find a threshold that you are 99% sure is higher than the true average daily power usage at T'Chaka's Lodge? Hint: this topic has not been formally covered in class, but with good intuition you should be able to calculate it using concepts covered in class.

m) [3 pts] What level of daily power usage at T'Chaka's Lodge would be considered an outlier?

n) [2 pts] You do a hypothesis test of H_a: $\mu_{T'Challa\ Wind} > 5{,}000$ versus H_0: $\mu_{T'Challa\ Wind} <= 5{,}000$. You do the test correctly, and (correctly) conclude that you <u>cannot</u> reject the null hypothesis at the 1% significance level. Which is the most accurate interpretation of this conclusion?

a. The sample mean is greater than or equal to 5,000
b. You have compelling evidence that the true population mean is greater than 5,000
c. You have not found compelling evidence that $\mu_{T'Challa\ Wind}$ is greater than 5,000
d. You are 1% confident that the null hypothesis is true

Pineapple

Pineapple is a fashion retailer with presence in the US, Canada, and Europe. Every season, they introduce new designs (called "new arrivals") to complement their assortment of different styles, the majority of which are designated classic designs (called "classics"). In order to understand what factors drive sales at their stores, they have sampled a few stores and identified their average weekly revenues, as well as the **following information** from the past fiscal year:

Traffic – The weekly average number of customers visiting that store
Product_categories – The number of different product categories offered in each store
Income – Median household income (in $K) in a 30-mile-radius of the store

Part of the data is presented below, along with part of the output from the regression model that was run with Revenues as the dependent variable and the **three potential drivers** as independent variables.

Store	Revenues ($)	Traffic	Product_categories	Income ($K)
1	31240	1820	22	100.8
2	15845	1452	31	41.3
3	16548	1227	38	32.6
4	28957	1365	25	92.5
5	14623	1371	40	51.5
6	17547	1475	48	64.7
7	22912	1650	51	85.6

Regression Statistics	
Multiple R	0.8056214
R Square	0.64902584
Adjusted R Square	0.61142147
Standard Error	3644.94
Observations	
ANOVA	
	df
Regression	
Residual	28
Total	

1. What are the number of observations for this model? (**2 pts**)

To further explain the variation in the Revenues, the analytics team at Pineapple decide to add a few more independent variables to their regression model. The regression outputs for the **original** and the **new model** are presented below on the **left-** and **right-hand** sides, respectively:

Regression Statistics	
Multiple R	0.8056214
R Square	0.64902584
Adjusted R Square	0.61142147
Standard Error	3644.94
Observations	
ANOVA	
	df
Regression	
Residual	28
Total	

Regression Statistics	
Multiple R	0.88015
R Square	0.77466402
Adjusted R Square	0.72058339
Standard Error	3001.1522
Observations	
ANOVA	
	df
Regression	
Residual	25
Total	

2. Use a partial F-test to compare the explanatory power of these two models at a **5% significance level**. Clearly state the conclusion you draw after performing the test. (**6 pts**)

After some further analysis, the analytics team decided to add the following independent variables to the original model with three independent variables:

Classics% - Percent of classic designs in the store assortment
New_arrivals% - Percent of new designs in the store assortment
Outlet – 1 if the store is located inside an outlet mall; 0 otherwise

Part of the data as well as the regression output are presented below for the new model. **Use this model for the rest of the questions**.

Store	Revenues ($)	Traffic	Product_categories	Income ($K)	Classics%	New_arrivals%	Outlet
1	31240	1820	22	100.8	61.5%	3.2%	1
2	15845	1452	31	41.3	33.4%	2.4%	0
3	16548	1227	38	32.6	45.6%	2.7%	0
4	28957	1365	25	24.8	52.7%	1.4%	1
5	14623	1371	40	71.5	33.1%	3.8%	0
6	17547	1475	48	54.7	47.4%	2.4%	1
7	22912	1650	51	55.7	56.2%	2.7%	1

Regression Statistics	
Multiple R	0.88015
R Square	0.77466402
Adjusted R Square	0.72058339
Standard Error	3001.1522
Observations	

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression		3.77314E+11	4.7164E+10	14.5341415	5.38241E-12
Residual	25	2.14174E+11	3245067278		
Total		5.91489E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	14717.63	10453	1.407981	0.171451	-6810.73	36246
Traffic		0.02167	3.00553761	0.00595835	0.02051	0.10975
Product_categories	69.0287	32.60489	2.11713	0.04437275		
Income	73.06619	39.7694	1.837247	0.078089	-8.84041	154.9728
Classics%	24080.47	11159.25	2.157892	0.04074	1097.551	47063.38
New_arrivals%	-142067	111037.3	-1.27945	0.212485	-370753	86618.67
Outlet	3922.598	1558.885	2.516284	0.018648	712.0136	7133.183

3. Calculate the regression coefficient for the Traffic variable (which is hidden in the table above). (**3 pts**)

4. Present the model above in standard format. (**6 pts**)

5. Construct an **approximate** 95% confidence interval for the slope of "Product_categories" variable. (**3 pts**)

6. Interpret the 95% confidence interval for the Outlets variable using **problem-specific** (rather than generic) language. (**3 pts**)

7. What are the null and alternative hypotheses of the t-test for which t-stat "1.837247" is calculated (in the row corresponding to the variable "Income" in the regression coefficient table). (**2 pts**) [MULTIPLE CHOICE]

   a) $H_0: \beta_{Income} \neq 0\ ;\ H_1: \beta_{Income} = 0$
   b) $H_0: \beta_{Income} > 0\ ;\ H_1: \beta_{Income} \leq 0$
   c) $H_0: \beta_{Income} > 0\ ;\ H_1: \beta_{Income} \neq 0$
   d) $H_0: \beta_{Income} = 0\ ;\ H_1: \beta_{Income} \neq 0$
   e) $H_0: \beta_{Income} < 0\ ;\ H_1: \beta_{Income} \geq 0$

8. What is the expected difference in the Revenues between a store which is located inside an outlet, has 40 product categories, 50% classics, and 20% new arrivals, and another store with the same Traffic and Income variables, not located in an outlet, and with 30 product categories, 40% classics, and 50% other styles (i.e., styles other than classics and new arrivals). (**5 pts**)

UNC SALARY EQUITY STUDY

UNC-Chapel Hill School of Medicine conducted a study of the salaries of the professors from different ranks to see if there exists any statistical evidence of discrimination in pay. They collected a sample of assistant (without tenure) professors as well as tenured associate and full professors. They measured the experience of the faculty using the following variables:

- "Yr rank - Yr hire": Years between initial hire at UNC-Chapel Hill and date of current rank
- "Yr Hire - Yr degree": Years between highest degree and hire date at UNC

They split the data into training and validation datasets. A sample of the training data is shown below:

OBS	TOTSAL	GENDER	RANK	Yr Hire - Yr degree	Yr rank - Yr hire
1	$154,690	M	TENURED PROFESSOR	2	12
4	$218,000	M	TENURED ASSOCIATE	12	7
5	$87,475	F	TENURED ASSOCIATE	2	7
7	$225,000	M	TENURED ASSOCIATE	15	5
8	$212,610	F	TENURED PROFESSOR	27	0
9	$137,873	M	TENURED ASSOCIATE	8	3
10	$135,350	M	ASSISTANT	9	0
11	$160,700	M	TENURED ASSOCIATE	7	7
12	$152,260	F	TENURED PROFESSOR	10	15
15	$183,814	M	ASSISTANT	16	0
18	$133,557	F	TENURED PROFESSOR	4	15
19	$129,101	M	TENURED ASSOCIATE	3	10

They first run a regression with the two variables defined above. A summary of the regression output is shown below:

Regression Statistics	
Multiple R	0.6659194
R Square	0.44344864
Adjusted R Square	0.41336479
Standard Error	38874.1647
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4.4551E+10	2.2276E+10	14.7404185	1.9582E-05
Residual	37	5.5914E+10	1511200681		
Total	39	1.0047E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	88250.7595	15390.0592	5.73427029	1.4382E-06	57067.5375	119433.982
Yr rank - Yr hire	4120.82021	1235.37882	3.33567334	0.00194541	1617.70495	6623.93547
Yr Hire - Yr degree	6127.93943	1144.36325	5.3548901	4.6915E-06	3809.23924	8446.63962

1. Which of the two independent variables are statistically significant (at the 1% significance level)? [MULTIPLE CHOICE] (**2 pts**)

   A) Yr rank - Yr hire
   B) Yr Hire - Yr degree
   C) Both
   D) Neither

In order to study possible discrimination in pay, we would like to incorporate the gender and rank of the faculty in our regression analysis. We define an indicator variable for "Male" and another indicator for "Tenured" faculty – the tenured indicator would distinguish the assistant professors from the tenured associate and full professors.

2. How would you define the "Tenured" indicator using an IF statement in Excel? In other words, if you were creating the "Tenured" indicator in column G, what Excel formula would you put in cell G3? Please write the IF statement clearly. (**3 pts**)

The following is a summary the regression output for the new regression model that also incorporates the two indicators defined above:

Regression Statistics					
Multiple R	0.73578846				
R Square	0.541384657				
Adjusted R Square	0.488971475				
Standard Error	36282.6983				
Observations	40				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	54390689397	13597672349	10.32916981	1.24587E-05
Residual	35	46075196862	1316434196		
Total	39	1.00466E+11			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	79867.32809	17681.36184	4.517034877	6.83437E-05	43972.25524
Male	16221.36835	12843.36509	1.263015435	0.214933199	-9852.048947
Tenured	51575.51259	20168.1759	2.557272053	0.01504069	10631.9388
Yr Hire - Yr degree	4626.407768	1264.491984	3.658708658	0.00082733	2059.352567
Yr rank - Yr hire	57.50929861	1958.842299	0.02935882	0.976745189	-3919.151984

3. We can see that the p-value of the "Yr rank - Yr hire" variable in the second model has significantly changed compared to the first model. How do explain this? **Please explain your answer in a few sentences only.** (**2 pts**)

4. Based on the second regression model, is there any significant evidence of discrimination in terms of gender? [MULTIPLE CHOICE] (**2 pts**)

   A) Yes, because p-value of Male indicator is more than 10%.
   B) No, because p-value of Male indicator is more than 10%.
   C) Yes, because the coefficient of Male indicator has a positive sign.
   D) No, because the coefficient of Male indicator has a positive sign.

5. The coefficient of the Tenured indicator is positive, and its p-value is 0.015. If we use a **5% significance level**, can we conclude that there is **inappropriate discrimination** between the pay of tenured and untenured faculty? Please justify your answer. (**3 pts**)

6. To further investigate the possible discrimination in gender, we would like to see if the impact of "Yr Hire - Yr degree" variable on pay might be different for men versus women. What independent variable could you add to your model to account for this? Please be as specific as you can. **(2 pts)**

We decide to finalize the regression model on the training dataset with 3 independent variables "Yr Hire - Yr degree", "Male" indicator, and "Tenured" indicator. A summary of the regression output is shown below:

Regression Statistics	
Multiple R	0.75673784
R Square	0.57265215
Adjusted R Square	0.5238124
Standard Error	35024.0245
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	5.7532E+10	1.4383E+10	11.7251236	3.8099E-06
Residual	35	4.2934E+10	1226682291		
Total	39	1.0047E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	87779.7579	16516.1675	5.31477765	6.1845E-06	54250.1553	121309.36
Male	-10590.7589	20820.9161	-0.5086596	0.61418243	-52859.4658	31677.9479
Tenured	58576.4015	12163.4851	4.81575808	2.7954E-05	33883.214	83269.589
Yr Hire - Yr degree	3395.75578	1192.61253	2.84732525	0.00732943	974.623616	5816.88793

7. Construct an **exact 90% prediction interval** for the salary of a male assistant professor with 6 years of experience prior to being hired at UNC-Chapel Hill. (**6 pts**)

We finally copy and paste the regression coefficients of the model above to the validation dataset to calculate the RMSE. The validation dataset is comprised of "Fixed-Term" Assistant, Associate, and (Full) Professors – **"Fixed-Term" means that such faculty are NOT on tenure track.** The screenshot below shows a sample of the validation dataset with the regression coeffecients on top:

	A	B	C	D	E	F	G	H	I	J
1	**RMSE**	$ 55,867.89		Intercept	Male	Tenured	Yr Hire - Yr degree			
2				87779.75791	-10590.7589	58576.4015	3395.755775			
3										
4	**OBS**	**TOTSAL**	**GENDER**	**RANK**	**Male**	**Tenured**	**Yr Hire - Yr degree**	**Predicted Salary**	**Error**	**Error^2**
5	2	$85,104	M	FIXED TERM ASSISTANT	1	0	2	$ 83,980.51		
6	6	$170,000	M	FIXED TERM ASSISTANT	1	0	4	$ 90,772.02	($79,228)	$6,277,072,487.30
7	13	$130,000	F	FIXED TERM ASSOCIATE	0	0	1	$ 91,175.51	($38,824)	$1,507,340,737.23
8	14	$125,000	F	FIXED TERM ASSISTANT	0	0	2	$ 94,571.27	($30,429)	$925,907,641.96
9	16	$95,000	F	FIXED TERM ASSOCIATE	0	0	0	$ 87,779.76	($7,220)	$52,131,895.77
10	20	$83,406	M	FIXED TERM ASSISTANT	1	0	1	$ 80,584.75	($2,821)	$7,959,424.81
11	22	$231,377	M	FIXED TERM PROFESSOR	1	0	7	$ 100,959.29	($130,418)	$17,008,779,240.09

8. What Excel function has been written in Cell H5 to calculate the predicted salary of $83,980.51? Write the Excel function clearly. (**3 pts**)

9. What should be the numerical value of the "Error" in Cell I5? This is the error that will be used in order to calculate the RMSE. (**2 pts**)

10. What Excel function should be written in Cell B1 to calculate the RMSE for the validation dataset. (**2 pts**)

11. Once we calculate the RMSE, we notice that it is 60% higher than the Standard Error (SE) from the training data – the SE is $35,024.02 while the RMSE turns out to be $55,867.89. Does such a high increase in RMSE (relative to SE) indicate an overfitting problem or might there be another issue with our validation process? Please justify your answer. (**3 pts**)