Business Statistics Final Exam

Winter 2018

This is a closed-book, closed-notes exam. You may use a calculator.

Please answer all problems in the space provided on the exam.

Read each question carefully and clearly present your answers.

Here are some useful formulas:

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2ab \times Cov(X, Y)$
- The standard error for the difference in the averages between groups a and b is defined as:

$$s_{(\bar{X}_a - \bar{X}_b)} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

where s_a^2 denotes the sample variance of group a and n_a the number of observations in group a.

Good Luck!

Honor Code Pledge: "I pledge my honor that I have not violated the Honor Code during this examination."

Signed: _____

Name: _____

Problem 1: Who's to blame? (10 points)

In manufacturing its iPhone, Apple buys a particular kind of microchip from 3 suppliers: 30% from Freescale, 20% from Texas Instruments and 50% from Samsung.

Apple has extensive histories on the reliability of the chips and knows that 3% of the chips from Freescale are defective; 5% from Texas Instruments are defective and 4% from Samsung are defective.

In testing a newly assembled iPhone, Apple found the microchip to be defective. What provider is the likely culprit?

Problem 2: Breaking Bad... (10 points each)

Two chemists working for a chicken fast food company, have been producing a very popular sauce. Let's call then Jesse and Mr. White. Gus, their boss, is tired of Mr. White's negative attitude and is thinking about "firing" him and keeping only Jesse on payroll. The problem, however, is that Mr. White seems to produce a higher quality sauce whenever he is in charge of production if compared to Jesse. Before making a final decision, Gus collected some data measuring the quality of different batches of sauce produced by Mr. White and Jesse. The results, measured on a quality scale, are listed below:

	average	std. deviation	sample size
Mr. White	97	1	7
Jesse	94	3	10

Two questions:

1. Based in this data, can we tell for sure which one is the better chemist?

2. Gus wants to keep the mean quality score for the sauce above 90. In this case, can he can rid of Mr. White, i.e., is Jesse good enough to run the sauce production?

Problem 3: Portfolios (5 points each)

We're considering building a portfolio from three investments: a fund tracking the SP500, a bond fund, and a fund of large cap stocks. The portfolios under consideration are:

- Portfolio A: 50% SP500, 50% bonds
- Portfolio B: 50% SP500, 50% large-cap

Returns on the large cap fund and the bond fund have the same expected value and standard deviation. Historically, there is a small negative correlation between the bond and SP500 funds, and a small positive correlation between the large cap and SP500 funds. The returns on each investment have normal distributions.

Using only the information given above, choose the **single** correct response to each question below:

(a) (4 points) What is the relationship between the expected returns for each portfolio?

- ◯ Portfolio A has higher expected returns
- ◯ Portfolio B has higher expected returns
- ◯ Both portfolios have the same expected returns
- ◯ Impossible to say without more information

(b) (4 points) If we want the portfolio with the largest Sharpe ratio, which portfolio should we choose?

- ◯ Portfolio A
- ◯ Portfolio B
- ◯ Either one; their Sharpe ratios are the same
- ◯ Impossible to say without more information

(c) (4 points) If we want the portfolio with the most potential for growth (say, the portfolio that is most likely to generate returns greater than its average plus 2%), which portfolio should we choose?

- ◯ Portfolio A
- ◯ Portfolio B
- ◯ Either one; they are equally likely to generate returns greater than their average plus 2%
- ◯ Impossible to say without more information

Problem 4 (2 points each)

Assume the model: $Y = 5 + 2X_1 + 3X_2 + \varepsilon, \quad \varepsilon \sim \ N(0, 81)$

1. What is $E[Y|X_1 = 1, X_2 = 0]$?
   (a) 5
   (b) 9
   (c) 7
   (d) 8
2. What is the $Var[Y|X_1 = 0, X_2 = 4]$?
   (a) 9
   (b) 81
   (c) 3
   (d) 6
3. What is the $Pr(Y > 5)$, given $X_1 = 0.5$ and $X_2 = 3$?
   (a) 15%
   (b) 68%
   (c) 98%
   (d) 87%
4. What is the $Pr(28 < Y < 35)$, given $X_1 = 4$ and $X_2 = 4$?
   (a) 5%
   (b) 23%
   (c) 2.5%
   (d) 34%

Problem 5 (5 points each)

"ProShares UltraShort S&P500 (SDS) seeks daily investment results, before fees and expenses, that correspond to two times the inverse ($-2\times$) of the daily performance of the S&P 500"

The above quote is from ProShares' website, the manager of SDS.

In trying to validate their claim and make sure that SDS is a good fund that appropriately tracks its target, I decided to collect data on monthly returns (in percentage terms) of SDS and the S&P500 Index since 2009 and run the following regression:

$$SDS = \beta_0 + \beta_1 SP500 + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

Regression Statistics	
Multiple R	0.994
R Square	0.989
Adjusted R Square	0.988
Standard Error	0.760
Observations	62.000

ANOVA

	df	*SS*	*MS*	*F*
Regression	1.000	3024.488	3024.488	5242.184
Residual	60.000	34.617	0.577	
Total	61.000	3059.106		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.437	0.103	-4.252	0.000
SP500	-1.867	0.026	-72.403	0.000

Answer the following questions:

1. In trying to evaluate the claim made by ProShares, test the appropriate hypotheses about β_0. What is your conclusion?

2. In trying to evaluate the claim made by ProShares, test the appropriate hypotheses about β_1. What is your conclusion?

3. What is your final evaluation? Is SDS a good ETF? Justify your answer (and don't forget to address the estimate of σ^2).

Problem 6: Crime data from our homework (5 points each)

Let's recall the "Crime" vs. "Police" example from our homework. There, we were trying to understand the effect of more police on crime and we couldn't just get data from a few different cities and run the regression of "Crime" on "Police". The problem here is that data on police and crime cannot tell the difference between more police leading to crime or more crime leading to more police... in fact I would expect to see a potential positive correlation between police and crime if looking across different cities as mayors probably react to increases in crime by hiring more cops. Again, it would be nice to run an experiment and randomly place cops in the streets of a city in different days and see what happens to crime. Obviously we can't do that!

The researchers from UPENN mentioned in the homework were able to estimate this effect by using what we call a natural experiment. They were able to collect data on crime in DC and also relate that to days in which there was a higher alert for potential terrorist attacks. Why is this a natural experiment? Well, by law the DC mayor has to put more cops in the streets during the days in which there is a high alert. That decision has nothing to do with crime so it works essentially as a experiment.

Here's is the main table displaying the results from the analysis:

EFFECT OF POLICE ON CRIME

TABLE 2

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

Figure 1: The dependent variable is the daily total number of crimes in D.C. This table present the estimated coefficients and their standard errors in parenthesis. The first column refers to a model where the only variable used in the High Alert dummy whereas the model in column (2) controls form the METRO ridership. * refers to a significant coefficient at the 5% level, ** at the 1% level.

Answer the following questions:

1. Why it was not enough to present the results from column (1) in the table? Why did they have to include the METRO ridership variable?

2. Can you explain why the estimates of the impact of police on crime from the columns are different?

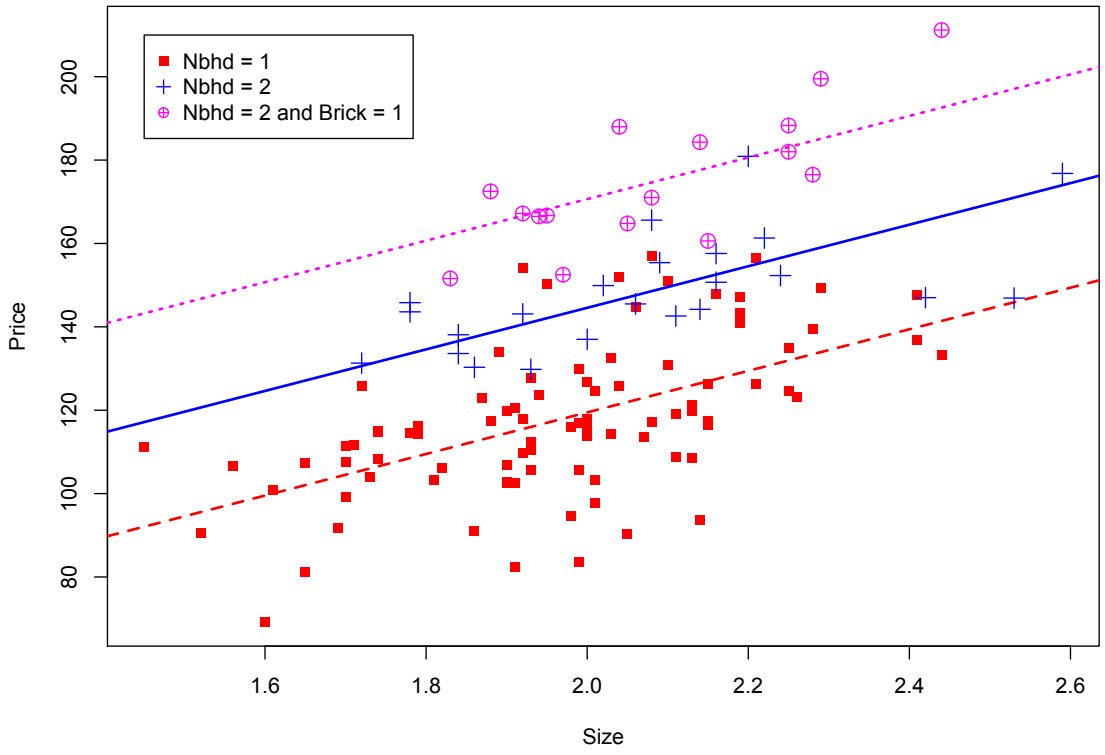
Problem 7: House Prices (2 points each)

Let's go back to the Midcity housing prices dataset from our homework... For simplicity I have combined the two cheap neighborhoods into one group so we are left with only two neighborhoods.

Let's start by looking at the following model:

$$\text{Model 1:} \quad Prices = \beta_0 + \beta_1 Size + \beta_2 NBH + \beta_3 BRICK \times NBH + \epsilon$$

where NBH is a dummy variable that takes the value 1 if the house is in neighborhood 2 and $BRICK$ is a dummy variable that equals 1 if the house is made out of brick. The figure below displays the results from the regression. This is a graphical representation of of the estimates of all coefficients in this regression.



Based on the figure, answer the following questions:

1. What is the estimated value for the effect of $Size$ on $Prices$ for houses in neighborhood 1?

   (a) 65.32

   (b) 30.45

   (c) 17.98

   (d) 49.85

2. What is the estimated value for the effect of *Size* on *Prices* for houses in neighborhood 2?

   (a) 65.32

   (b) 49.85

   (c) 20.31

   (d) 12.67

3. What is the estimated premium for brick houses is neighborhood 2?

   (a) 15.76

   (b) 38.61

   (c) 26.08

   (d) 52.10

4. What is the estimated average difference between a 1,800 sqft wood house in neighborhood 2 and neighborhood 1?

   (a) 25.09

   (b) 39.78

   (c) 48.90

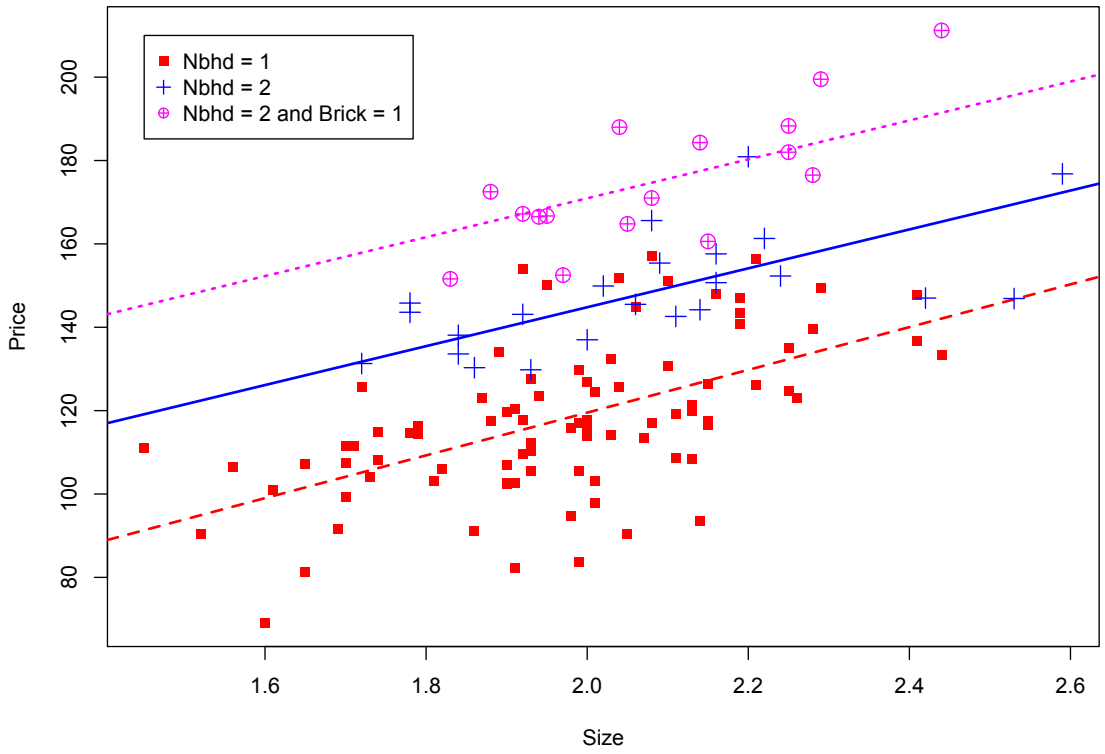
   (d) 13.94

Problem 8: House Prices again! (2 points each)

Continuing in analyzing the MidCity data (same as the previous question), I now decided to investigate whether or not the effect of *Size* on *Prices* changes in the different neighborhoods. To this end, I worked with the following model:

Model 2: $Prices = \beta_0 + \beta_1 Size + \beta_2 NBH + \beta_3 BRICK \times NBH + \beta_4 Size \times NBH + \epsilon$

The results are summarized in the figure below:



Based on the figures, answer the following questions:

1. In model 2, what is the estimated value for the effect of *Size* on *Prices* for houses in neighborhood 1?

   (a) 71.30

   (b) 30.45

   (c) 17.98

   (d) 51.27

2. In model 2, what is the estimated value for the effect of $Size$ on $Prices$ for houses in neighborhood 2?

   (a) 75.23

   (b) 46.67

   (c) 20.31

   (d) 51.27

3. In model 2, what is the estimate for β_4?

   (a) 46.67

   (b) 51.27

   (c) 13.15

   (d) -4.60

4. What is the t-stat for the difference between the slope for $Size$ in the two neighborhoods?

   (a) 2.15

   (b) -4.44

   (c) -0.35

   (d) 5.63

Problem 9: Medal Count
(3 points each)

Using data from Beijing 2008 and London 2012 I run a regression trying to understand the impact of **GDP** (gross domestic product measured in billions of US$) and **Population** (in millions of people) on the **total number of medals** won by a country in the summer Olympics. The results are

Regression Statistics	
Multiple R	0.82488
R Square	0.68043
Adjusted R	0.67660
Standard E	10.83097
Observatic	170.00000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00000	41712.86080	20856.43040	177.78909	0.00000
Residual	167.00000	19590.76273	117.30996		
Total	169.00000	61303.62353			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.77423	0.90407	5.28082	0.00000	2.98935	6.55911
Population	0.01267	0.00467	2.71239	0.00738	0.00345	0.02189
GDP	0.00778	0.00050	15.67150	0.00000	0.00680	0.00876

(a) Is the intercept interpretable in this regression? Why?

(b) Provide an interpretation for the coefficients associated with **Population** and **GDP**?

(c) What is the t-stat for **Population** telling you? Clearly explain the hypothesis being tested and your conclusion.

(d) From the results, give a 95% prediction interval for the total number of medals for the U.S. in the Rio 2016 Olympics, given that the U.S. current GDP is of 18.5 trillion of dollars and population is 300 million?

The following table shows the total medal count for a few countries in Rio 2016 Olympics along with their current GDP and Population:

Country	Total Medals	GDP (in US$ billions)	Population (in millions)
U.S.	121	18,500	300
Great Britain	67	2,800	64
China	70	11,300	1,357
Brazil	19	1,600	200
India	2	1,877	1,250
Holland	19	853	16.8
Fiji	1	3.8	0.881

(e) Using the results from the regression, which of these countries performance in the Rio 2016 is not surprising? Why?

(f) Based on the regression results, rank the performance of these countries in the Rio Olympics. Explain your ranking methodology.

I proceeded to add a dummy variable for the **host** country into the regression... I also ran a regression with only **GDP** and **Host**. The results are below:

Regression Statistics	
Multiple R	0.8639
R Square	0.7462
Adjusted R	0.7417
Standard E	9.6805
Observatic	170.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3.0000	45747.2827	15249.0942	162.7214	0.0000
Residual	166.0000	15556.3409	93.7129		
Total	169.0000	61303.6235			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.8246	0.8081	5.9705	0.0000	3.2292	6.4200
Population	0.0034	0.0044	0.7763	0.4387	-0.0053	0.0121
GDP	0.0077	0.0004	17.4626	0.0000	0.0069	0.0086
Host	48.3225	7.3648	6.5613	0.0000	33.7819	62.8632

Regression Statistics	
Multiple R	0.86332
R Square	0.74532
Adjusted R	0.74227
Standard E	9.66902
Observatic	170.00000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00000	45690.80714	22845.40357	244.36222	0.00000
Residual	167.00000	15612.81639	93.48992		
Total	169.00000	61303.62353			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.92223	0.79728	6.17374	0.00000	3.34817	6.49628
GDP	0.00789	0.00041	19.46333	0.00000	0.00709	0.00869
Host	50.15148	6.96945	7.19590	0.00000	36.39190	63.91107

(h) Of the 3 models presented, which one is the best in your opinion? Carefully explain why?

(i) In the last model presented, provide an interpretation for the coefficient associated with **Host**.

(j) Using your chosen model, evaluate Brazil's performance in the Rio Olympics. Compare and explain the difference in the results if you were to talk about Brazil's performance based on the first regression.