First Midterm Examination

Econ 103, Statistics for Economists

February 14th, 2017

You will have 70 minutes to complete this exam. Graphing calculators, notes, and textbooks are not permitted.

I pledge that, in taking and preparing for this exam, I have abided by the University of Pennsylvania's Code of Academic Integrity. I am aware that any violations of the code will result in a failing grade for this course.

Name: ____________________

Signature: ____________________

Student ID #: ____________________ Recitation #: ____________________

Question:	1	2	3	4	5	6	Total
Points:	20	20	20	20	20	40	140
Score:							

Instructions: Answer all questions in the space provided, continuing on the back of the page if you run out of space. Show your work for full credit but be aware that writing down irrelevant information will not gain you points. Be sure to sign the academic integrity statement above and to write your name and student ID number on *each page* in the space provided. Make sure that you have all pages of the exam before starting.

Warning: If you continue writing after we call time, even if this is only to fill in your name, twenty-five points will be deducted from your final score. In addition, a point will be deducted for each page on which you do not write your name and student ID.

The following question appeared on the homework assignment to accompany Lecture #4:

$\boxed{20}$ 1. What value of a minimizes $\sum_{i=1}^{n}(y_i - a)^2$? Prove your answer.

Solution: Differentiating with respect to a, we find that the first-order condition is

$$-2\sum_{i=1}^{n}(y_i - a) = 0$$

Rearranging and splitting up the sum, this is equivalent to

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a$$

The sum on the right-hand side equals na. Dividing through by n gives $a = \bar{y}$.

$\boxed{20}$ 2. I have three boxes. The first contains two red balls, the second contains two blue balls, and the third contains one red ball and one blue ball. I choose a box at random so that each box is equally likely to be selected. I then reach inside and pull out a ball: it's red. Calculate the probability that the other ball in the box is also red. To help us grade your answer more easily, let B_1 be the event that I chose the first box, B_2 the second box, and B_3 the third box. Let R be the event that I drew a red ball.

Solution: The problem statement specifies $P(B_1) = P(B_2) = P(B_3) = 1/3$ along with $P(R|B_1) = 1$, $P(R|B_2) = 0$, and $P(R|B_3) = 1/2$. We are asked to calculate the probability that the other ball is red given that the ball I drew is red. This is equivalent to asking for the probability of B_1 given R. By Bayes' Rule

$$P(B_1|R) = \frac{P(R|B_1)P(B_1)}{P(R)}$$

For the denominator,

$$\begin{aligned} P(R) &= P(R|B_1)P(B_1) + P(R|B_2)P(B_2) + P(R|B_3)P(B_3) \\ &= 1 \times 1/3 + 0 \times 1/3 + 1/2 \times 1/3 \\ &= 1/2 \end{aligned}$$

Therefore, $P(B_1|R) = (1/3)/(1/2) = 2/3$.

3. Suppose we have a dataset $(x_1, y_1), \ldots, (x_n, y_n)$ of scores for n students on two exams: x denotes exam #1 and y denotes exam #2. Let s_{xy} denote the sample covariance between

exam scores, r denote the sample correlation, s_x and s_y the respective sample standard deviations, and $\bar{x}$ and $\bar{y}$ the respective sample means. Let $\widehat{y} = a + bx$ where a is the intercept and b the slope of a linear regression calculated from this dataset.

$\boxed{2}$ (a) Write down the formula for a in terms of b, $\bar{x}$, and $\bar{y}$.

Solution: $a = \bar{y} - b\bar{x}$

$\boxed{2}$ (b) Write down the formula for b in terms of s_{xy} and s_x.

Solution: $b = s_{xy}/s_x^2$

$\boxed{2}$ (c) Write down the formula for r in terms of s_{xy}, s_x and s_y

Solution: $r = s_{xy}/(s_x s_y)$

$\boxed{6}$ (d) Suppose that $b > r$. Then which is larger: s_x or s_y? Explain.

Solution: Taking the ratio of the formulas from (b) and (c) we have $b/r = s_y/s_x$ so if $b > r$ then $s_y > s_x$.

$\boxed{8}$ (e) Suppose the sample mean score on the first exam, $\bar{x}$, was 75% while the sample mean score on the second exam, $\bar{y}$, was 85%. Moreover, our regression model predicts that someone who got 50 on the first exam will get 70 on the second. Calculate b.

Solution: When $x = 50$ we have $\widehat{y} = 70$. Thus $70 = a + 50b$. Using part (a),

$$a = \bar{y} - b\bar{x} = 85 - 75b$$

and combining these:

$$70 = (85 - 75b) + 50b = 85 - 25b$$

Solving, $b = -15/(-25) = 3/5 = 0.6$.

$\boxed{20}$ 4. Veronica wants to know the fraction of undergraduates who drink underage. Because she worries that her subjects may be unwilling to admit to breaking the law, Veronica conducts her interviews as follows:

> *Hi, I'm carrying out a survey on underage drinking. In a moment I'm going to ask if you drink underage, but before I do I'd like you to flip this fair coin in secret. If you get heads, answer my question truthfully. If you get tails,*

answer YES regardless of whether you drink underage. Because I don't know the outcome of your coin flip, if you answer YES I'll never know whether you really drink underage of just flipped a tails.

Veronica interviews a large number of subjects who all follow her instructions exactly: 85% answer YES. Calculate the fraction of undergraduates who drink underage.

Solution: Let Y be the event that someone answers YES, D be the event that he drinks underage, and T be the event that he flips a tails. Since $Y = D \cup T$, by the addition rule we have

$$P(Y) = P(D) + P(T) - P(D \cap T)$$

Since the coin is fair, $P(T) = 0.5$. Since the outcome of a coin toss doesn't depend on whether someone does or doesn't drink underage, D and T are independent which implies $P(D \cap T) = P(D)P(T)$. Substituting this information along with $P(Y) = 0.85$,

$$0.85 = P(D) + 0.5 - 0.5 \times P(D)$$

and solving we see that $P(D) = 0.7$.

Alternatively, you could solve this using the Law of Total Probability:

$$\begin{aligned} P(Y) &= P(Y|T)P(T) + P(Y|T^c)P(T^c) \\ 0.85 &= 1 \times 1/2 + P(Y|T^c) \times 1/2 \\ 0.7 &= P(Y|T^c) \end{aligned}$$

using the fact that $P(Y|T^c) = P(D)$ since anyone who flips heads answers truthfully, and the outcome of the coin flip has no bearing on whether someone drinks underage.

5. Let X be a random variable with support set $\{-1, 1\}$, $p(1) = q$, and $p(-1) = 1 - q$.

$\boxed{5}$ (a) Write down the CDF of X.

Solution:

$$F(x_0) = \begin{cases} 0, & x_0 < -1 \\ 1-q, & -1 \leq x_0 < 1 \\ 1, & x_0 \geq 1 \end{cases}$$

$\boxed{3}$ (b) Calculate $E[X]$.

Solution: $E[X] = -1 \times (1-q) + 1 \times q = 2q - 1$

$\boxed{3}$ (c) Calculate $E[X^2]$.

Solution: $E[X^2] = (1-q) \times (-1)^2 + q \times 1^2 = (1-q) + q = 1$

$\boxed{3}$ (d) Calculate $Var(X)$.

Solution: By the shortcut rule:

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 = 1 - (2q-1)^2 \\ &= 1 - (4q^2 - 4q + 1) \\ &= 4q(1-q) \end{aligned}$$

Another way to solve this is to notice that we can write X as a linear transformation of a Bernoulli RV. In particular if $Z \sim \text{Bernoulli}(q)$ then $X = 2Z - 1$ and hence $Var(X) = 4Var(Z) = 4q(1-q)$ as we calculated directly.

$\boxed{6}$ (e) Let X_1 and X_2 be independent RVs both of which have the same pmf as X, defined in the problem statement. Write out the support set and pmf of $Y = X_1 + X_2$.

Solution: The support set is $\{-2, 0, 2\}$ and the pmf is

$$\begin{aligned} p(-2) &= P(X_1 = -1 \cap X_2 = -1) = (1-q)^2 \\ p(0) &= P(X_1 = -1 \cap X_2 = 1) + P(X_1 = 1 \cap X_2 = -1) = 2q(1-q) \\ p(2) &= P(X_1 = 1 \cap X_2 = 1) = q^2 \end{aligned}$$

6. This question concerns a `data.table` called `star`. Here are its first and last five rows:

```
      race classtype yearssmall hsgrad g4math g4reading
   1:    1         3          0     NA     NA        NA
   2:    2         3          0     NA    706       661
   3:    1         3          0      1    711       750
   4:    2         1          4     NA    672       659
   5:    1         2          0     NA     NA        NA
  ---
6321:    2         2          0     NA     NA        NA
6322:    1         2          0     NA     NA        NA
6323:    2         2          0      0     NA        NA
6324:    1         1          1     NA     NA        NA
6325:    1         1          1     NA     NA        NA
```

The data come from a randomized controlled experiment carried out in Tennessee in the 1980s. Students were randomly assigned to one of three class types, as indicated by the column `classtype`: 1 = small class, 2 = regular-sized class, 3 = regular-sized class with a teacher's aid. In the years that followed, researchers measured several outcomes: `hsgrad` is an indicator that takes the value 1 if a student graduated high school and 0 if not, `g4math` and `g4reading` are the student's scores on standardized 4th grade reading and math tests. In this question we will not work with the columns `race` or `yearssmall`.

5 (a) Say that the data are stored at `http://data.com/star.csv`. Give the full set of commands needed to read `star.csv` into R as a `data.table` called `star`.

Solution:

```
library(data.table)
star <- fread("http://data.com/star.csv")
```

5 (b) Give the R command needed to create a new `data.table` called `small` that only contains students who were randomly assigned to a small class.

Solution: There are many possible answers, such as

```
small <- star[classtype == 1]
```

or

```
small <- subset(star, classtype == 1)
```

5 (c) Write R code to carry out a linear regression that predicts a student's reading score from her math score. Use the data for all students in the experiment.

Solution:

```
lm(g4reading ~ g4math, data = star)
```

or

```
star[ , lm(g4reading ~ g4math)]
```

5 (d) Here are the results of running the command from the previous part:

```
Coefficients:
(Intercept)       g4math
   317.5765       0.5696
```

If Beatrice scored 10 points higher in math than Alejandro, how many points higher would we predict that Beatrice will score in reading?

Solution: 10 times the coefficient on `g4math`, or about 5.7 points higher.

5 (e) Write R code to calculate the average high school graduation rate separately for students assigned to each different kind of class. Be sure to properly account for missing values.

Solution: There are many possible answers. The one used to generate the output in the following section was:

```
star[, .(grad_rate = mean(hsgrad, na.rm = TRUE)), keyby = classtype]
```

5 (f) The results of the preceding part were as follows:

```
   classtype grad_rate
1:         1 0.8359202
2:         2 0.8251619
3:         3 0.8392857
```

Based on these results, do smaller class sizes improve high school graduation rates relative to regular-sized classes? Explain in no more than three sentences.

Solution: Students randomly assigned to the small classes (`classtype 1`) have a slightly higher graduation rate than those assigned to the regular classes (`classtype 2`) namely 83.6% versus 82.5%. On the other hand, students assigned to regular classes with a teacher's aid (`classtype 3`) have an even higher graduation rate: 83.9%. So it may not be the class size *per se* that matters. On the whole the differences are small.

10 (g) Write an R function called mycor that takes two input arguments, a numeric vector x and another numeric vector y, and returns the sample correlation between then. In your answer you may use any R commands you like *except* the built-in correlation command cor. For simplicity you may assume that neither x nor y contains any missing values.

Solution: There are many possible answers. Here's one possibility:

```
mycor <- function(x, y){
  s_xy <- cov(x, y)
  s_x <- sd(x)
  s_y <- sd(y)
  return(s_xy / (s_x * s_y))
}
```

Name: ____________________ Student ID #: ____________________