Final Examination

Econ 103, Statistics for Economists

December 16th, 2014

You will have 120 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: ______________________________

Student ID #: ______________________________

Signature: ______________________________

Question:	1	2	3	4	5	6	Total
Points:	20	20	20	20	50	70	200
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, two points will be deducted for each page on which you do not write your name and student ID.

1. Consider the following simple dataset with nine observations of two variables:

x	y
1	1
2	1
3	1
1	2
2	2
3	2
1	3
2	3
3	3

(a) (4 points) Calculate $\bar{x}$ and $\bar{y}$.

(b) (4 points) Calculate s_x^2 and s_y^2.

(c) (6 points) Calculate s_{xy}.

(d) (6 points) Calculate the slope and intercept of a linear regression model that uses this dataset to predict y from x.

2. (20 points) Let $Y \sim$ Bernoulli(1/3) and define X *conditional* on Y as follows: if $Y = 0$ then $X \sim$ Bernoulli(3/4), otherwise $X \sim$ Bernoulli(4/5). Write the joint pmf of X and Y in a 2×2 table. Put the X-values in the *rows* and the Y-values in the *columns*.

3. Suppose I take a meter stick and break it into two pieces. The exact point at which I break it, S, is random and follows a Uniform distribution. Thus, the length of the first piece is simply S while the length of the second piece is $1 - S$.

   (a) (10 points) Let A be the *area* of a rectangle with sides S and $1 - S$. Calculate the expected value of A.

   (b) (10 points) The R command `runif(n)` draws `n` independent, Uniform$(0, 1)$ random variables. Using this command, write R code to verify your solution to the preceding part via Monte Carlo simulation using 1000 draws.

4. To pay off his gambling debts to Rodrigo, Rossa has taken a part-time job as a plumber and needs to measure the length of two pipes. When he uses his measuring tape, Rossa makes normally distributed errors with variance σ^2 and mean zero: if an object's true length is ℓ, his measurement is $L \sim N(\ell, \sigma^2)$. Suppose that each measurement error is independent of the others and let ℓ_A, ℓ_B denote the true lengths of pipes A and B.

   (a) (4 points) Rossa decides to start with pipe A. Following the adage "measure twice, cut once," his instinct is to make *two* measurements of the pipe and use the *average* to estimate ℓ_A. Calculate the bias and variance of this estimator.

   (b) (4 points) Rossa notices that pipe A is clearly longer than pipe B and comes up with an idea: rather than measuring each pipe twice, he'll lay the pipes end to end and measure the sum and difference of their lengths. Let D be Rossa's measurement of the *difference* of lengths, and S be his measurement of the *sum* of lengths. Assume that Rossa lines up the pipes perfectly: when he measures any length (of a single pipe, a sum or a difference) his measurement equals the true length plus a $N(0, \sigma^2)$ error, as above. What is the distribution of D? What is the distribution of S?

(c) (6 points) Rossa decides to estimate ℓ_A using $(S + D)/2$ and ℓ_B using $(S - D)/2$. Are these estimators unbiased? If so, prove it. If not, calculate the bias of each.

(d) (6 points) Calculate the variance of the two estimators from the preceding part.

5. Petra has a dataframe called `reaction` containing measurements of the reaction times of 19 students given in seconds. Although 19 observations is a relatively small sample size, you may assume for the purposes of this question that the approximation based on the Central Limit Theorem applies. Each row of `reaction` corresponds to an individual: the value in the column `dom` gives that individual's reaction time using her *dominant* hand while the value in the column `nondom` gives her reaction time using her *non-dominant* hand. For example, I am left-handed so my value for `dom` would be my reaction time with my *left hand*. Here are the first six rows of the dataframe and some summary statistics:

```
    dom nondom
1 0.159  0.188
2 0.176  0.194
3 0.180  0.171
4 0.130  0.195
5 0.180  0.199
6 0.121  0.179
```

	dom	nondom
Sample Mean	0.180	0.202
Sample S.D.	0.045	0.048
Correlation	0.83	

(a) (5 points) Give the units of each of the summary statistics from the above table.

(b) (5 points) All of the measurements in `reaction` are smaller than a second so Petra runs the R command `reaction <- 1000 * reaction` to convert the dataset to *milliseconds*. Give the updated values for each of the above summary statistics.

(c) (5 points) Petra wants to use the data contained in `reaction` to determine whether people's reaction times differ when they use their dominant versus non-dominant hand. Is this a problem based on two independent samples or matched pairs? Explain briefly.

(d) (15 points) Write R code that computes a 90% confidence interval for the difference of population mean reaction times: *non-dominant* minus *dominant*.

(e) (15 points) Now suppose that, instead of calculating a confidence interval, Petra wanted to test the null hypothesis that reaction times are *the same* regardless of whether one uses one's dominant or non-dominant hand against the two-sided alternative. Calculate the value of the appropriate test statistic.

(f) (5 points) Approximately what is the p-value for Petra's test from the preceding part? What should she conclude?

6. This question concerns a dataframe called `birthdata` containing observational data on 1000 mothers and their first-born children: `birthweight` is a given child's birth weight in grams, `weeksgest` is the number of weeks between that child's conception and his or her birth (i.e. weeks of gestation), and `smoker` is a dummy variable that takes on the value one if that child's mother smoked during pregnancy. Here are the first few rows:

```
  birthweight weeksgest smoker
1        4252        38      1
2        4229        42      0
3        4338        41      0
4        3850        39      0
5        3430        41      0
6        3260        39      0
```

To answer this question, refer to the regression results on final page of the exam.

(a) (6 points) What is the sample mean birth weight for children whose mother smoked during pregnancy? How does this compare to the sample mean birth weight for children whose mothers did *not* smoke during pregnancy?

(b) (6 points) Construct an approximate 95% confidence interval for the population mean difference of birth weights between children whose mothers smoked during pregnancy and those whose mothers did not.

(c) (6 points) Suppose you wanted to carry out a two-sided test of the null hypothesis that the children of smokers and non-smokers weigh the same, on average, at birth. What is the value of your test statistic? Write out the full R command needed to calculate the p-value for this test. Approximately what would be your result?

(d) (6 points) Interpret your results from the preceding two parts. Do they provide evidence of a causal relationship between smoking and birth weight?

(e) (5 points) What is the sample correlation between `birthweight` and `weeksgest`?

Name: ______________________ Student ID #: ______________________

(f) (6 points) Suppose we wanted to use `weeksgest` *alone* to predict `birthweight`. For two newborns who differ by one week in gestation time, by how much would we predict that their birth weights differ?

(g) (5 points) What are the units of the slope in Regression #2?

(h) (6 points) What is the meaning of the intercept in Regression #2?.

(i) (6 points) If you were given the task of predicting birthweight as accurately as possible *either* using `smoker` *or* using `weeksgest` but not both, which would you use? How much more accurate is your preferred model? Explain briefly.

(j) (6 points) Suppose you wanted to predict `birthweight` using *both* `smoker` and `weeksgest`. Two of the four regressions are relevant for this task, although they differ in the *way* in which they use the information from the two variables. Which models are they, and how do they differ in the relationship they fit between `birthweight` and `weeksgest` depending on the value of `smoker`? In your answer, discuss only the regression *models*, not the *results* of fitting these models to `birthdata`.

(k) (12 points) For each of the models you listed in your answer to the preceding part, use the appropriate regression results to write out the *rule* we would use to predict `birthweight` from `weeksgest` for a child whose mother smoked during pregnancy. Repeat for a child whose mother did *not* smoke during pregnancy.

Regression #1

```
lm(formula = birthweight ~ smoker, data = birthdata)
            coef.est coef.se
(Intercept) 3472.48    18.30
smoker      -292.91    50.96
---
n = 1000, k = 2
residual sd = 540.20, R-Squared = 0.03
```

Regression #2

```
lm(formula = birthweight ~ weeksgest, data = birthdata)
            coef.est coef.se
(Intercept) -1009.00   281.10
weeksgest     112.82     7.13
---
n = 1000, k = 2
residual sd = 490.87, R-Squared = 0.20
```

Regression #3

```
lm(formula = birthweight ~ smoker + weeksgest, data = birthdata)
            coef.est coef.se
(Intercept) -940.49   276.31
smoker      -278.90    45.49
weeksgest    111.99     7.00
---
n = 1000, k = 3
residual sd = 482.11, R-Squared = 0.23
```

Regression #4

```
lm(formula = birthweight ~ smoker + weeksgest + smoker:weeksgest,
    data = birthdata)
                 coef.est coef.se
(Intercept)      -1069.20   303.79
smoker             461.93   728.26
weeksgest          115.26     7.70
smoker:weeksgest   -18.85    18.49
---
n = 1000, k = 4
residual sd = 482.10, R-Squared = 0.23
```