Final Examination

Econ 103, Statistics for Economists

May 14th, 2019

You will have two hours 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:	60	25	25	20	20	50	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, a point will be deducted for each page on which you do not write your name and student ID.

1. Answer each of the following. For full credit, explain your answers clearly and succinctly.

[10] (a) Three percent of *Tropicana* brand oranges are already rotten when they arrive at the supermarket. In contrast, six percent of *Sunkist* brand oranges arrive rotten. A local supermarket buys forty percent of its oranges from *Tropicana* and the rest from *Sunkist*. Suppose we randomly choose an orange from the supermarket and see that it is rotten. What is the probability that it is a *Tropicana*? In your answer, let R be the event that an orange is rotten and T be the event that it is a *Tropicana*.

Solution: By the law of total probability:

$$\begin{aligned} P(R) &= P(R|T)P(T) + P(R|T^c)P(T^c) \\ &= 0.03 \times 0.4 + 0.06 \times 0.6 \\ &= 0.012 + 0.036 \\ &= 0.048 \end{aligned}$$

and by Bayes' Rule:

$$\begin{aligned} P(T|R) &= \frac{P(R|T)P(T)}{P(R)} \\ &= \frac{0.012}{0.048} = 1/4 = 0.25 \end{aligned}$$

[7] (b) Let $X_1, X_2 \sim$ iid with mean μ and variance σ^2. Is $(0.1X_1 + 0.9X_2)$ is a more efficient estimator of μ than $(0.5X_1 + 0.5X_2)$?

Solution: No: both estimators are unbiased, so it makes sense to talk about "efficiency," but $\text{Var}(0.1X_1 + 0.9X_2) = 0.01\sigma^2 + 0.81\sigma^2 = 0.82\sigma^2$ which is much larger than $\text{Var}(0.5X_1 + 0.5X_2) = 0.5\sigma^2$.

[7] (c) Let X and Y be RVs with $Var(X) = 2$, $Var(Y) = 1$, and $Cov(X, Y) = 0$. Calculate $Var(X - Y)$.

Solution: $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y) = 3$

[7] (d) Suppose that Alice and Bob each draw independent random samples of size $n = 100$ from a normal population with unknown mean μ and known variance $\sigma^2 = 9$. They both construct 95% confidence intervals for μ. Will the widths of Alice and Bob's intervals be the same? Will Alice and Bob's intervals be identical?

Name: _____ Student ID #: _____

Solution: Both intervals will take the form $\bar{X} \pm 0.6$ since $\sigma/\sqrt{n} = 3/10 = 0.3$ in this example. But since Alice and Bob use different samples, they will not obtain the same realization for $\bar{X}$. Thus, their intervals will have the same width, 1.2, but will not coincide: they will be centered in different locations.

7 (e) In the "Pepsi Challenge" experiment from class there were four cups of Coke and four of Pepsi. In this question, consider a modified version of the experiment with *three* cups of each kind of soda. Everything else is unchanged. Calculate the probability that our test statistic, the number of cokes correctly identified, will equal two *under the null hypothesis*.

Solution:

$$\frac{\binom{3}{2} \times \binom{3}{1}}{\binom{6}{3}} = \frac{3 \times 3}{20} = 9/20 = 0.45$$

7 (f) Alice constructs a 95% CI for μ: $[-0.5, 0.3]$. Bob tests $H_0 \colon \mu = 0$ vs. $H_1 \colon \mu \neq 0$ with $\alpha = 0.01$ using the same dataset as Alice. Will he reject H_0?

Solution: To answer this, introduce a third character: Cheryl. Suppose Cheryl constructed a 99% confidence interval using the same data as Alice. To determine the result of Bob's test, we could simply check whether zero is contained in Cheryl's confidence interval. Unfortunately the problem statement doesn't give us Cheryl's interval, but from our study of confidence intervals, we know that Alice's interval would be a *subset* of Cheryl's interval. Since 0 is in Alice's interval, this implies that it will also be in Cheryl's interval. Hence, Bob will fail to reject H_0.

7 (g) Suppose that $X_1, \ldots, X_5 \sim$ iid N$(1, 4)$ independently of $Y_1, \ldots, Y_{20} \sim$ iid N$(-1, 24)$. Write a line of R code to calculate $P(\bar{X} - \bar{Y} > 0)$.

Solution: We have $\bar{X} \sim N(1, 4/5)$ independently of $\bar{Y}(-1, 6/5)$. Thus, it follows that $\bar{X} - \bar{Y} \sim N(2, 2)$ and accordingly

$$P(\bar{X} - \bar{Y} > 0) = P\left(\frac{\bar{X} - \bar{Y} - 2}{\sqrt{2}} > \frac{-2}{\sqrt{2}}\right) = P(Z > -\sqrt{2})$$

where $Z \sim N(0, 1)$. Therefore the desired probability is `1 - pnorm(-sqrt(2))`.

8 (h) The Fibonacci sequence is defined as follows: $F_1 = 1, F_2 = 1$, and $F_i = F_{i-1} + F_{i-2}$ for $i \geq 3$. In other words: $1, 1, 2, 3, 5, 8, 13, 21, 34, 55 \ldots$ and so on. Write R code to calculate the first 20 terms of the Fibonacci sequence $(F_1, F_2, \ldots, F_{20})$ and store them in a vector called `fib`.

Solution:

```
fib <- rep(NA, 20)
fib[1] <- 1
fib[2] <- 1
for(i in 3:20) {
  fib[i] <- fib[i - 1] + fib[i - 2]
}
```

2. Let X, Y, Z be iid discrete RVs with support set $\{-1, 1\}$ and probability mass function $p(-1) = 1 - p$, $p(1) = p$. Define $S = X + Y + Z$.

5 (a) Calculate $E[X]$.

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

5 (b) Calculate the variance of Z. How does it compare to that of a Bernoulli(p) RV?

Solution: We have $E[Z^2] = (-1)^2 \times (1-p) + 1^2 \times p = 1$, and from the preceding part it follows that $E[Z] = 2p - 1$ since X and Z are identically distributed. Hence, by the shortcut rule for variance:

$$Var(Z) = E[Z^2] - (E[Z])^2 = 1 - (2p - 1)^2 = 4p(1 - p).$$

We know that the variance of a Bernoulli(p) RV is $p(1-p)$. Hence, the variance of Z is always *larger* than that of the corresponding Bernoulli.

5 (c) Calculate $Var(S)$.

Solution: Since (X, Y, Z) are iid,

$$Var(S) = Var(X) + Var(Y) + Var(Z) = 3 \times 4p(1 - p) = 12p(1 - p)$$

10 (d) Calculate $P(S = 1)$.

Solution: This calculation effectively identical is to the reasoning we used to work out the pmf of a Binomial RV. The only way to obtain $S = 1$ is if *exactly one* of the RVs (X, Y, Z) takes on the value -1 while the rest take on the value 1. There are three mutually exclusive ways that this can occur:

(i) $X = -1$, $Y = Z = 1$

(ii) $Y = -1$, $X = Z = 1$

(iii) $Z = -1$, $X = Y = 1$

Since (X, Y, Z) are independent, $P(X = -1 \cap Y = 1 \cap Z = 1) = (1-p)p^2$. But by the same reasoning, $P(X = 1 \cap Y = -1 \cap Z = 1) = (1-p)p^2$ and $P(X = 1 \cap Y = 1 \cap Z = -1) = (1-p)p^2$. Hence, $P(S = 1) = 3(1-p)p^2$.

3. Dr. Evil gives twenty quizzes in his Henchman Studies 103 course. Each quiz has a single question, drawn from a list of ten review questions. Each list of review questions contains seven *Easy* questions and three *Hard* questions. Dr. Evil claims to select quiz questions completely at random with no regard to their difficulty. He claims, for example, that the first quiz will contain one question drawn at random from the ten review questions for Lecture #1. Yvonne suspects that Dr. Evil is *lying* about choosing questions completely at random. Because 9 out of the 20 quiz questions during the semester were *Hard*, she thinks Dr. Evil took question difficulty into account when creating his quizzes. Let H be the total number of *Hard* questions that appear on quizzes during the semester.

[5] (a) If Dr. Evil is telling the truth, what is $E[H]$?

Solution: If Dr. Evil is telling the truth, then $H \sim \text{Binomial}(20, 0.3)$. Hence, $E[H] = 20 \times 0.3 = 6$.

[5] (b) If Dr. Evil is telling the truth, what is $Var(H)$?

Solution: Continuing from the solution to the previous part, $Var(H) = 20 \times 0.3 \times 0.7 = 4.2$.

[10] (c) Yvonne decides to test the null hypothesis that Dr. Evil is telling the truth against the alternative that Hard questions are disproportionately *likely* to appear on quizzes, using the approximation based on the CLT. Calculate her test statistic.

Solution: If Dr. Evil is telling the truth then each Hard question has probability $p_0 = 0.3$ of appearing on a quiz. Yvonne's estimate, on the other hand, is $\hat{p} = 9/20 = 0.45$. Hence, the test statistic is

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.45 - 0.3}{\sqrt{(0.3 \times 0.7)/20}} \approx 1.46$$

5 (d) Yvonne enters the R commands

```
x <- 0.9 + 0:10 / 200
y <- qnorm(x)
cbind(x,y)
```

and obtains the following output from the console:

```
             x        y
 [1,] 0.900 1.281552
 [2,] 0.905 1.310579
 [3,] 0.910 1.340755
 [4,] 0.915 1.372204
 [5,] 0.920 1.405072
 [6,] 0.925 1.439531
 [7,] 0.930 1.475791
 [8,] 0.935 1.514102
 [9,] 0.940 1.554774
[10,] 0.945 1.598193
[11,] 0.950 1.644854
```

Continuing from the preceding part, approximately what is the p-value for her test? Interpret her results.

Solution: Her p-value is around 0.07. With $\alpha = 0.1$ we would reject the null, but with $\alpha = 0.05$ we would not. To put this into context, 0.07 is approximately 1/14. So if Dr. Evil teaches Henchman Studies 103 fourteen times, we would expect to see one semester in which the fraction of hard questions on quizzes exceeded 0.45 even if he's choosing the questions completely at random. We have found some evidence that Dr. Evil may be lying, but it's not overwhelming.

20 4. Write an R function called `myreg` to estimate β_0 and β_1 in the simple linear regression $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Your function should take two input arguments: a vector `y` of observed outcomes and a corresponding vector `x` of observed values for the predictor variable. You may assume that there are no missing values and that the lengths of

`x` and `y` are the same. Your function should return a vector with two elements: the estimate of β_0 and the estimate of β_1 (in that order). In your answer you may use any R functions that you like *except* for `lm`.

Solution: There are many possible correct answers. Here is one:

```
myreg <- function(x, y) {
  b1 <- cov(x, y) / var(x)
  b0 <- mean(y) - b1 * mean(x)
  return(c(b0, b1))
}
```

20 5. This problem is taken from the extensions. It has been re-worded slightly for clarity, but the solution is unchanged. Let Y and X be RVs. In this problem you will find the the constants β_0 and β_1 that solve

$$\min_{\beta_0,\beta_1} E[(Y - \beta_0 - \beta_1 X)^2].$$

For the purposes of this question you may assume that expectation and differentiation can be interchanged, i.e. that $\frac{\partial}{\partial\theta}E[f(Z,\theta)] = E[\frac{\partial}{\partial\theta}f(Z,\theta)]$. You do not have to check the second order condition.

(a) Show that $\beta_0 = E[Y] - \beta_1 E[X]$.

Solution: Differentiating with respect to β_0 gives the first order condition

$$-2E[Y - \beta_0 - \beta_1 X] = 0$$

Re-arranging using the linearity of expectation, $\beta_0 = E[Y] - \beta_1 E[X]$

(b) Using the preceding part, find β_1.

Solution: Substituting the expression for β_0 back into the objective function,

$$E[(Y - \mu_Y - \beta_1\mu_X - \beta_1 X)^2] = E\left[\{(Y - \mu_Y) - \beta_1(X - \mu_X)\}^2\right]$$

using the shorthand $E[Y] = \mu_Y$ and $E[X] = \mu_X$. Now, differentiating with respect to β_1 gives the first order condition

$$-2E\left[\{(Y - \mu_Y) - \beta_1(X - \mu_X)\}(X - \mu_X)\right] = 0$$

Finally, rearranging and solving for β_1 using the linearity of expectation,

$$\begin{aligned} E[(Y-\mu_Y)(X-\mu_X)] &= E[\beta_1(X-\mu_X)^2] \\ Cov(X,Y) &= \beta_1 Var(X) \\ \beta_1 &= \frac{Cov(X,Y)}{Var(X)} \end{aligned}$$

6. This question relies on an R dataframe called `kaiser` with data for 1174 babies born at the Kaiser Foundation hospital in Oakland California. Here are the first few rows:

```
  bwt gestation smoke
1 120       284     0
2 113       282     0
3 128       279     1
4 108       282     1
5 136       286     0
6 138       244     0
```

Each row in `kaiser` is a newborn baby: `bwt` gives the baby's birthweight in ounces, `gestation` gives the length of the pregnancy in days, and `smoke` is a dummy variable taking the value one if the baby's mother smoked during pregnancy. The last page of this exam contains results for five regression models estimated using `kaiser`. You may find it helpful to tear out the page of regression results for ease of reference.

[5] (a) What is the sample mean of `bwt`?

Solution: Regression #1 contains only an intercept: $Y = \beta_0 + \varepsilon$. Hence, the estimated intercept equals the sample mean: 119.46 ounces.

[5] (b) Approximately what is the sample variance of `bwt`?

Solution: The standard error of the intercept in Regression #1 is 0.53. But we know that the intercept in this case is simply the sample mean. Hence, 0.53 is the standard error of the sample mean: $S/\sqrt{n} = 0.53$. From the regression results, we see that $n = 1174$. Hence, re-arranging and solving for S, we obtain $S \approx 18$ ounces, in other words approximately one pound.

[5] (c) Explain why the R-squared of Regression #1 is exactly zero.

Solution: By definition, $R^2 = 1 - \sum_{i=1}^n \widehat{\varepsilon}_i^2 / \sum_{i=1}^n (y_i - \bar{y})^2$. Since Regression #1 has only an intercept, $\widehat{y}_i = \widehat{\beta}_0 = \bar{y}$ and hence $\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \bar{y}$. Substituting this into the definition of R-squared, the numerator and denominator in the fraction are identical so the fraction itself equals one. Hence,

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - 1 = 0.$$

5 (d) Which mothers have heavier babies: those who smoke or those who do not? How large is the difference?

Solution: From Regression #2, we see babies born to mothers who *do not smoke* are, on average, approximately 9 ounces heavier that those born to mothers who smoke.

5 (e) Continuing from the preceding part, is there convincing evidence of a difference in the population, or could our estimate by explained by sampling variation?

Solution: From the output of regression #2, the standard error of the difference of means is approximately 1. Hence, an approximate 95% confidence interval for the difference of means (smokers − non-smokers) is $-9 \pm 2 = (-11, -7)$ while a 99.7% confidence interval is $-9 \pm 3 = (-12, -6)$. We have found extremely strong evidence of a difference in the population. Moreover, this difference is large. The smallest difference from the 99.7% CI is -6 ounces. From part (b), this corresponds to 1/3 of a standard deviation in birthweight. This is substantial, and the point estimate is even larger: 9 ounces 1/2 of a standard deviation of birthweight.

5 (f) Continuing from the preceding part, does the `kaiser` dataset provide evidence that smoking during pregnancy has a *causal effect* on birthweight? Why or why not?

Solution: Putting to one side any outside information about the health effects of smoking, the `kaiser` dataset *alone* does not provide evidence of a causal effect. Mothers who smoke during pregnancy are likely different from those who do not in myriad ways: smokers tend to be poorer, for example. The comparison from Regression #2 does not hold these other factors constant, so the estimated effect is not necessarily causal.

5 (g) About how accurately does a regression that uses *only* `gestation` predict birth-

weight?

Solution: The residual standard deviation of regression #3 is 16.74 ounces, so `gestation` predicts birthweight to an accuracy of about 1 pound.

5 (h) What is the approximate value of the correlation between `bwt` and `gestation`?

Solution: The R-squared of Regression #3 is 0.17, so the correlation $\sqrt{0.17} \approx 0.4$.

5 (i) Consider two mothers, both of whose pregnancies lasted exactly d days: Xanthippe smoked during pregnancy while Yvonne did not. Based on this information, whose baby would we predict will be heavier at birth? Does your answer depend on whether we use Regression #4 or #5 to make our prediction? Explain briefly.

Solution: Based on Regression #4 we would predict that Yvonne's baby will be heavier: this regression only allows a difference of *intercepts*, so the predicted difference in birthweight for smokers versus non-smokers is the same for all values of `gestation`. In contrast, Regression #5 allows both a different slope and intercept. The intercept in Regression #5 is about 73 ounces lower for smokers, while the slope is about 0.23 ounces per day higher. This means that the regression line for smokers starts off *below* that for non-smokers, but eventually the two lines cross. The question is, *where* do they cross? For each additional day of gestation, the regression line for smokers "gains" on that for non-smokers by 0.23 ounces. Hence, making up for its initial deficit of 73 ounces will take $73/0.23 \approx 317$ days. This would correspond to a pregnancy of 45 weeks, which is so long as to be practically unheard of. For any "reasonable" value of d, Regression #5 will also predict that Yvonne's baby will be heavier.

5 (j) Is there convincing evidence of a different slope in relationship between `gestation` and `bwt` for smokers versus non-smokers? If so, what is the nature of the difference?

Solution: Regression #5 allows for a different slope and intercept in the relationship between `gestation` and `bwt` depending on whether or not a mother smoked during pregnancy. The interaction term `smoke:gestation` gives the difference of slopes for smokers versus non-smokers. The estimate is 0.23 with a standard error of 0.06. If we were to test the null hypothesis of no difference of slopes, our test statistic would be approximately 4. This gives a p-value below 0.003, so there is very strong evidence against the null. An approximate 95%

confidence interval for the difference of slopes is $0.23 \pm 0.12 = (0.11, 0.35)$, so we have strong evidence that the slope is *higher* for mothers who smoke. In other words, we would predict a *larger* difference in birthweight between babies with different lengths of gestation if their mothers smoked than if they did not smoke.

Regression #1

```
lm(formula = bwt ~ 1, data = kaiser)
            coef.est coef.se
(Intercept) 119.46     0.53
---
n = 1174, k = 1
residual sd = 18.33, R-Squared = 0.00
```

Regression #2

```
lm(formula = bwt ~ smoke, data = kaiser)
            coef.est coef.se
(Intercept) 123.09     0.66
smoke        -9.27     1.06
---
n = 1174, k = 2
residual sd = 17.77, R-Squared = 0.06
```

Regression #3

```
lm(formula = bwt ~ gestation, data = kaiser)
            coef.est coef.se
(Intercept) -10.75     8.54
gestation     0.47     0.03
---
n = 1174, k = 2
residual sd = 16.74, R-Squared = 0.17
```

Regression #4

```
lm(formula = bwt ~ smoke + gestation, data = kaiser)
            coef.est coef.se
(Intercept) -3.18     8.33
smoke       -8.37     0.97
gestation    0.45     0.03
---
n = 1174, k = 3
residual sd = 16.25, R-Squared = 0.22
```

Regression #5

```
lm(formula = bwt ~ smoke + gestation + smoke:gestation, data = kaiser)
                coef.est coef.se
(Intercept)      19.64    10.29
smoke           -72.69    17.23
gestation         0.37     0.04
smoke:gestation   0.23     0.06
---
n = 1174, k = 4
residual sd = 16.16, R-Squared = 0.22
```

Name: _____ Student ID #: _____