Final Examination

Econ 103, Statistics for Economists

December 12, 2012

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	7	8	9	10	11	12	Total
Points:	10	10	15	25	15	15	25	20	25	30	15	35	240
Score:													

Instructions: Answer all questions in the space provided. Should you run out of space, continue on the back of the page. 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.

1. (10 points) In an effort to reduce traffic fatalities, local authorities installed traffic cameras in particularly accident-prone locations in west London in 1992. In the preceding three years, there were 62 fatal accidents in these locations. In the three years following the introduction of the cameras, there were only 19 fatal accidents. Based on this evidence, local authorities concluded that traffic cameras reduce fatalities. Traffic cameras have since been installed in the whole of London. Using the information provided in this question *only*, do we have good evidence that traffic cameras reduce the frequency of fatal accidents? Explain why or why not.

Solution: No: this is an example of regression to the mean (the regression fallacy). The sites in question were chosen because they had unusually high fatalities. In any given period, the intersections with the greatest number of fatalities are likely to be intrinsically dangerous *and* to have suffered a spell of "bad luck." This is because fatalities are not perfectly correlated over time. Even if we had *not* installed cameras, we would expect fatalities to be lower on average in the second period at those intersections that had unusually high numbers of fatalities in the first period. Students may also mention that we didn't discuss normalizing for volume of traffic.

2. (10 points) Of women who undergo regular mammograms, two percent have breast cancer. If a woman has breast cancer, there is a 90% chance that her mammogram will come back positive. If she does *not* have breast cancer there is a 10% chance that her mammogram will come back positive. Given that a woman's mammogram has come back positive, what is the probability that she has breast cancer?

Solution: Let B be the event that a given woman has breast cancer and M be the event that her mammogram comes back positive. By Bayes' Rule,

$$\mathbb{P}(B|M) = \frac{\mathbb{P}(M|B)\mathbb{P}(B)}{\mathbb{P}(M)}$$

By the law of total probability,

$$\begin{aligned}\mathbb{P}(M) &= \mathbb{P}(M|B)\mathbb{P}(B) + \mathbb{P}(M|B^c)\mathbb{P}(B^c)\\ &= 0.9 \times 0.02 + 0.1 \times 0.98 = 0.018 + 0.098 = 0.116\end{aligned}$$

Hence,

$$\mathbb{P}(B|M) = \frac{0.9 \times 0.02}{0.116} = \frac{0.018}{0.116} \approx 0.16$$

3. (15 points) Explain the difference between μ, $\bar{X}_n$ and $\bar{x}$ using the concepts and terminology we have studied in this course.

Solution: We use the symbol μ to denote the population mean. This a *parameter*, in other words an unknown constant. Both of the remaining quantities are called the sample mean, but there is an important distinction between then. Whereas $\bar{X}_n$ is an *estimator*, $\bar{x}$ is an *estimate*. Under random sampling, $\bar{X}_n$ is a random variable: since the sample is random, the estimator is random. In contrast, $\bar{x}$ is a constant number, obtained by calculating the sample mean of the data (realizations) that we observed for in a *particular sample*. In other words, $\bar{X}_n$ denotes *process* of taking a random sample and computing the mean of that sample, whereas $\bar{x}$ denotes the *result* of this process.

4. Let X be a Uniform$(0,1)$ random variable.

(a) (5 points) Write down the probability density function (pdf) and support of X.

Solution: $f(x) = 1$ for $x \in (0,1)$, zero otherwise.

(b) (5 points) Calculate the cumulative distribution function (cdf) of X.

Solution:

$$F(x_0) = \int_{-\infty}^{x_0} f(x)\, dx = \int_0^{x_0} 1\, dx = x|_0^{x_0} = x_0$$

(c) (5 points) Calculate $\mathbb{E}[X]$

Solution:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\, dx = \int_0^1 x\, dx = \left.\frac{x^2}{2}\right|_0^1 = 1/2$$

(d) (5 points) Calculate $\mathbb{E}[X^2]$

Solution:

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_0^1 x^2\, dx = \left.\frac{x^3}{3}\right|_0^1 = 1/3$$

(e) (5 points) Using the shortcut formula, calculate $Var(X)$.

Solution:

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1/3 - (1/2)^2 = 1/3 - 1/4 = (4-3)/12 = 1/12$$

5. (15 points) About as many boys as girls are born in hospitals. Many babies are born every week at City General. In Cornwall, a country town, there is a small hospital where only a few babies are born every week. A *normal week* is one where between 45% and 55% of the babies are female. An *unusual week* is one where more than 55% are girls, or more than 55% are boys. Do unusual weeks occur more often at City General or Cornwall? Explain your answer using what you learned about sampling distributions in this course.

Solution: Unusual weeks are more common at Cornwall because the variance of the sampling distribution of a sample proportion is lower the larger the sample size: $Var(\widehat{p}) = p(1-p)/n$. The idea here is that p is the same at both hospitals, but n is larger at City General. Thus, there will be more sampling variability, and consequently more unusual weeks, at Cornwall. Another way to answer this would be to appeal to the law of large numbers.

6. This question concerns the quantity $\tilde{p}$ used to construct the "refined" confidence interval for a population proportion that I introduced in class. Recall that $\tilde{p} = (n\widehat{p} + 2)/(n+4)$, where $\widehat{p} = \left(\sum_{i=1}^{n} X_i\right)/n$ and $X_1, \ldots, X_n \sim$ iid Bernoulli(p).

   (a) (5 points) Is $\tilde{p}$ an unbiased estimator of p? If so, prove it. If not, calculate its bias.

Solution:

$$\mathbb{E}[\tilde{p}] - p \quad = \quad \mathbb{E}\left[\frac{n\widehat{p}+2}{n+4}\right] - p = \frac{n\mathbb{E}[\widehat{p}]+2}{n+4} - p = \frac{np+2-p(n+4)}{n+4} = \frac{2-4p}{n+4}$$

Hence the bias depends on $2-4p$. If $p = 1/2$, then $2 - 4p = 0$ so $\tilde{p}$ is unbiased. For all other values of p, it is biased. When $p > 1/2$, we have $2 - 4p < 0$, so the bias is negative. When $p < 1/2$, we have $2 - 4p > 0$ so the bias is positive.

   (b) (5 points) Which has a higher variance: $\widehat{p}$ or $\tilde{p}$? Prove your answer.

Solution: To begin, we know that $Var(\widehat{p}) = p(1-p)/n$. Now,

$$\begin{aligned} Var(\tilde{p}) &= Var\left(\frac{n\widehat{p}+2}{n+4}\right) = \left(\frac{n}{n+4}\right)^2 Var(\widehat{p}) = \left(\frac{n}{n+4}\right)^2\left(\frac{p(1-p)}{n}\right) \\ &= p(1-p)\left[\frac{n}{(n+4)^2}\right] \end{aligned}$$

So it suffices to compare $1/n$ against $n/[(n+4)^2]$. We have:

$$\frac{n}{(n+4)^2} = \frac{n}{n^2+8n+16} = \frac{1}{n+8+16/n}$$

Since $n < (n+8+16/n)$, it follows that $1/(n+8+16/n) < 1/n$. Therefore, $\tilde{p}$ has a lower variance than $\widehat{p}$. The difference is when n is relatively small.

(c) (5 points) Using your answers to (a) and (b), calculate the mean-squared error of $\tilde{p}$. Is $\tilde{p}$ a consistent estimator of p? Prove your answer.

Solution: Combining (a) and (b),

$$MSE(\tilde{p}) = \left(\frac{2-4p}{n+4}\right)^2 + \left[\frac{np(1-p)}{(n+4)^2}\right] \to 0$$

Hence, $\tilde{p}$ is a consistent estimator of p.

7. Suppose that $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. Let $\bar{X}_n$ be the sample mean and S^2 be the sample variance. Except in part (d), you do not need to provide any justification for your answers.

(a) (5 points) What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$?

Solution: $N(0,1)$

(b) (5 points) What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$?

Solution: $t(n-1)$

(c) (5 points) What is the sampling distribution of $(n-1)S^2/\sigma^2$?

Solution: $\chi^2(n-1)$

(d) (10 points) Using your answer to part (c), derive a $100 \times (1-\alpha)$ confidence interval for the variance of a normal population based on a random sample of n observations. Express the interval in terms of the appropriate R commands.

Solution: Define:

$$\begin{aligned} a &= \texttt{qchisq}(\alpha/2, \texttt{df} = \texttt{n - 1}) \\ b &= \texttt{qchisq}(1-\alpha/2, \texttt{df} = \texttt{n - 1}) \end{aligned}$$

Then, by (c)

$$\begin{aligned} \mathbb{P}\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) &= 1-\alpha \\ \mathbb{P}\left(\frac{a}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)S^2}\right) &= 1-\alpha \\ \mathbb{P}\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) &= 1-\alpha \end{aligned}$$

Therefore, the desired confidence interval is

$$\left[\frac{(n-1)S^2}{\texttt{qchisq}(1-\alpha/2, \texttt{df = } n-1)}, \frac{(n-1)S^2}{\texttt{qchisq}(\alpha/2, \texttt{df = } n-1)}\right]$$

8. Suppose that $X_1, \ldots, X_n \sim$ iid Bernoulli(p) and define $\widehat{p} = \sum_{i=1}^n X_i/n$.

(a) (5 points) Calculate the standard deviation of the sampling distribution of $\widehat{p}$.

Solution:

$$Var(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = p(1-p)/n$$

Hence, $SD(\widehat{p}) = \sqrt{p(1-p)/n}$.

(b) (5 points) There is a technical term for the standard deviation of a sampling distribution. What is it?

Solution: Standard error (SE).

(c) (5 points) What value of p maximizes the standard deviation of the sampling distribution of $\widehat{p}$? What is the maximized standard deviation? Be sure to check the second order condition.

Name: ______________________ Student ID #: ______________________

Solution: It suffices to maximize the function $p(1-p) = p - p^2$. The FOC is $1 - 2p = 0$ so that $p^* = 1/2$. The second derivative of this function is -2 for all p, so we have a global maximum. The maximized standard deviation is $\sqrt{(1/2)^2/n} = (1/2)/\sqrt{n} = 1/(2\sqrt{n})$.

(d) (5 points) Sarah is a graduate student in the political science department and is preparing to carry out a poll designed to estimate the proportion of college students who support gay marriage. Sarah plans to construct a 95% confidence interval based on her results and wants to gather enough data so that the margin of error of her poll is no more than 4%. Based on your answer to part (c) and the *textbook confidence interval* for a population proportion, what is the smallest sample size that Sarah should consider using in her poll?

Solution: The margin of error for a textbook 95% confidence interval for a population proportion is approximately $2\sqrt{\widehat{p}(1-\widehat{p})/n}$. We need to find the minimum n such that this quantity is no greater than 0.04. Rearranging, and using the fact that both sides are non-negative

$$\begin{aligned} 2\sqrt{\widehat{p}(1-\widehat{p})/n} &\leq 0.04 \\ \sqrt{\widehat{p}(1-\widehat{p})} &\leq 0.02\sqrt{n} \\ 50\sqrt{\widehat{p}(1-\widehat{p})} &\leq \sqrt{n} \\ 2500 \times \widehat{p}(1-\widehat{p}) &\leq n \end{aligned}$$

Now, as a side-effect of our answer to part (c), the *largest possible* value that $\widehat{p}(1-\widehat{p})$ can take is when $\widehat{p} = 1/2$, in which case $\widehat{p}(1-\widehat{p}) = 1/4$. Hence, $n \geq 625$ guarantees a margin of error of no more than 4% for a 95% CI.

9. This question concerns the following line of R code:

```
sims <- replicate(10000, quantile(rnorm(10), 0.5))
```

(a) (5 points) Explain what this command does. What is the point of running it?

Solution: This code replicates the following simulation experiment 10000 times and stores the result in a vector called `sims`: "take a random sample of size 10 from a standard normal distribution and calculate the sample median." By examining `sims` we can study the sampling distribution of the `sample median` calculated for a normal population.

(b) (5 points) After running this command I entered `mean(sims)` in the R console and got the result `-0.006093938`. What does this suggest about the sampling distribution of the sample median?

> **Solution:** This is extremely close to zero, which we know is the true median of a normal distribution. This suggests that the sample median is an unbiased estimator of the population median for a normal population.

(c) (5 points) Next, I entered `hist(sims)` into the console. The resulting graph appears in Figure 1. Based on this histogram, what would you conjecture about the sampling distribution of the sample median?

> **Solution:** The histogram looks bell-shaped. This suggests that the sample distribution of the sample median is approximately normal provided that the population is normal.

(d) (5 points) Now suppose I were to run the following command

```
sims2 <- replicate(10000, mean(rnorm(10)))
```

If I entered `mean(sims2)` into the console, approximately what value would you expect R to return? What if entered `var(sims2)`?

> **Solution:** This command examines the sampling distribution of the sample *mean* rather than median, again for a sample size of ten drawn from a standard normal population. We know that the sample mean is an unbiased estimator of the population mean under random sampling, so we should get approximately zero for `mean(sims2)`. We also know that the variance of the sample mean is σ^2/n, so we would expect to get a result of approximately 0.1 for `var(sims2)`.

(e) (5 points) Finally, I entered `var(sims)` at the console and got a result of `0.141154`. Based on this result and your answers to parts (b) and (d), which would you recommend as an estimator of the median of a normal population: the sample median or the sample mean?

> **Solution:** The mean of a normal distribution equals the median so the first thing to notice is that these are both estimating *the same population quantity*, namely μ. We know that the sample mean is an unbiased estimator of the population mean under random sampling, hence it is also an unbiased estimator of the population median. Our results above suggest that the sample median is also an unbiased estimator of the median of a normal population. However, the

Name: ______________________ Student ID #: ______________________

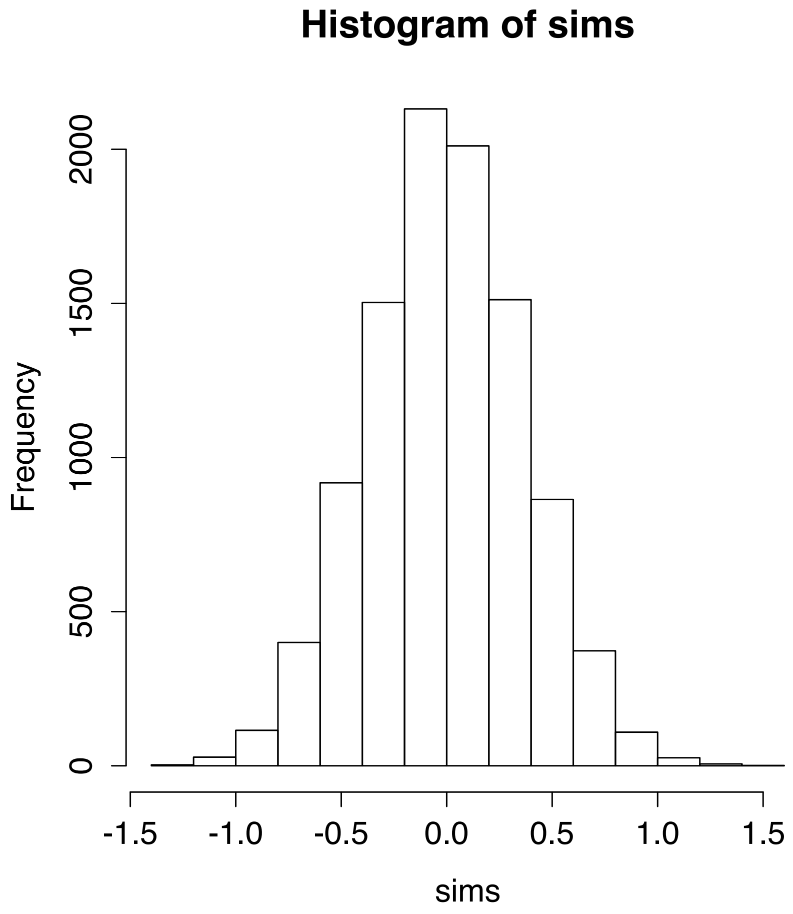

Figure 1: Histogram for Question 9

sample median seems to have a substantially higher variance than the sample mean: 0.14 versus 0.1. Therefore, we should prefer the sample mean.

b

10. This question concerns the "Pepsi Challenge" experiment. The setup is identical to the one we used in class except with five cups of each soda rather than four. The experimental procedure is as follows. We first fill five cups with Pepsi and five with Coke. Then we randomize the order of the cups and allow our expert to taste each. Finally, we ask the expert to identify which five cups contain Coke and record the number of cups that she correctly identifies. The idea is to determine whether our expert can really tell the difference between Coke and Pepsi based on the results of the experiment. This question asks to you formalize the experiment as a statistical hypothesis test. Our test statistic T is the number of Cokes that our expert correctly identified.

Name: ____________________ Student ID #: ____________________

(a) (5 points) What is our null hypothesis, H_0, in this experiment?

Solution: The null hypothesis is that our so-called "expert" is in fact just guessing: she is simply choosing five cups at random and calling them Cokes.

(b) (5 points) What is our alternative hypothesis, H_1, in this experiment?

Solution: Our alternative is that the expert has actual skill in discerning the difference between Coke and Pepsi. This is a one-sided alternative: we'll reject the null if she identifies a sufficient number of Cokes correctly.

(c) (5 points) Under H_0 what is the probability that $T = 5$?

Solution: Under the null, our expert is merely choosing randomly, and each possible allocation of five cups to "Coke" is equally likely. There are

$$\binom{10}{5} = \frac{10!}{5!5!} = 252$$

ways to choose the cups at random, of which only one is correct. Therefore, the desired probability is $\mathbb{P}(T = 5|H_0) = 1/252 \approx 0.004$.

(d) (5 points) Under H_0 what is the probability that $T = 4$?

Solution: Under the null, each of the 252 ways to choose five cups is equally likely. The case $T = 4$ corresponds to mis-identifying one Coke. There are 5 ways to choose which Coke is mis-identified as a Pepsi, and 5 ways to choose which Pepsi is mis-identified as a Coke. Hence $\mathbb{P}(T = 5|H_0) = 25/252 \approx 0.1$

(e) (5 points) Suppose I wanted to test H_0 at the 5% significance level. What is my decision rule given the alternative you specified?

Solution: Reject H_0 if $T > 4$ or equivalently if $T = 5$.

(f) (5 points) Suppose our expert identified four Cokes correctly. What is the p-value for our test against the alternative you specified?

Solution: $\mathbb{P}(T \geq 4|H_0) = 26/252 \approx 0.103$.

11. Suppose there is a logarithmic relationship between random variables X and Y, namely $\ln Y_i = \beta X_i + \epsilon_i$ where $\epsilon_i \sim \text{iid}(0, \sigma^2)$ is an unobserved error term.

Name: ____________________ Student ID #: ____________________

(a) (5 points) Suppose we want to estimate β by least squares regression using a random sample of observations $(y_1, x_1), \ldots, (y_n, x_n)$. What optimization problem should we solve?

Solution: In least squares regression, we minimize the sum of squared vertical deviations (residuals). In other words:

$$\min_{\beta} \sum_{i=1}^{n} (\ln y_i - \beta x_i)^2$$

(b) (10 points) Solve the problem from part (a) to derive an explicit formula for $\widehat{\beta}$. You do not need to check the second order condition.

Solution: We differentiate with respect to β, and manipulate the first order condition as follows:

$$\begin{aligned} -2\sum_{i=1}^{n} (\ln y_i - \beta x_i)\, x_i &= 0 \\ \sum_{i=1}^{n} (\ln y_i - \beta x_i)\, x_i &= 0 \\ \sum_{i=1}^{n} \ln{(y_i)} x_i &= \beta \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Hence,

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} \ln{(y_i)} x_i}{\sum_{i=1}^{n} x_i^2}$$

12. This question is based on an dataset containing observations on students in Econ 103: `male` takes the value 1 if a given student is male, zero otherwise; `midterm2` gives that student's score on the second midterm; and `midterm1` gives the student's score on the first midterm. Using this dataset, I estimated four regression models using R. The results appear in Table 1.

(a) (5 points) Suppose I wanted to test the null hypothesis that men and women do just as well, on average, on the second midterm of Econ 103 against the two-sided alternative using a 5% significance level. I can carry out this test directly from the results present above. Which results should I use and how should I carry out the test? In particular, what is the appropriate test statistic, what is the appropriate critical value, and what is the outcome of the test?

Table 1: Regression Results

Regression 1:

```
lm(formula = midterm2 ~ male)
            coef.est coef.se
(Intercept) 81.82     2.12
male        -0.22     2.91
---
n = 70, k = 2
residual sd = 12.17, R-Squared = 0.00
```

Regression 2:

```
lm(formula = midterm2 ~ midterm1)
            coef.est coef.se
(Intercept) 34.63     9.32
midterm1     0.59     0.12
---
n = 70, k = 2
residual sd = 10.35, R-Squared = 0.28
```

Regression 3:

```
lm(formula = midterm2 ~ male + midterm1)
            coef.est coef.se
(Intercept) 34.79     9.47
male        -0.31     2.50
midterm1     0.59     0.12
---
n = 70, k = 3
residual sd = 10.43, R-Squared = 0.28
```

Regression 4:

```
lm(formula = midterm2 ~ male + male:midterm1 + midterm1)
              coef.est coef.se
(Intercept)   19.31    14.22
male          26.76    18.82
midterm1       0.78     0.18
male:midterm1 -0.34     0.23
---
n = 70, k = 4
residual sd = 10.34, R-Squared = 0.30
```

Solution: We should use the results of the first regression: the coefficient estimate for `male` is the difference of mean test scores between men and women (i.e. $\bar{x}_M - \bar{x}_W$). The test statistic is the absolute value of the ratio of this estimated coefficient divided by its estimated standard error: $|-0.22/2.91| \approx 0.08$. The approximate critical value for this test is 2, so we fail to reject the null hypothesis.

(b) (5 points) Explain the difference between the models used in Regression 2, versus 3 and 4. Do not comment on the results, only the *models themselves.*

Solution: Regression 2 estimates the same relationship between score on the first midterm and the second for men and women. Regression 3 allows the intercept of the regressions to vary across sex, and Regression 4 allows both the intercept and the slope to vary across sex.

(c) (5 points) Based on the above regression results, what is the sample correlation between students' scores on the two midterms?

Solution: For a simple linear regression, the R^2 is the square of the sample correlation between x and y. Hence, $\sqrt{0.28} \approx 0.53$ is the sample correlation between scores on the first and second midterm.

(d) (5 points) Explain the meaning of the coefficient estimate `midterm1` in Regression 2 and construct a 95% confidence interval for this parameter. What would you conclude on the basis of this interval?

Solution: This estimate tells us the difference in score on midterm two that we would predict between two groups of students who differed by one point in their scores on midterm one: people who did one point better on the first exam do about 0.6 points better on the second exam. An approximate 95% confidence for this parameter is 0.6 ± 0.24, in other words $(0.36, 0.84)$. This interval does not include zero and is bounded substantially away from it. Our data strongly suggest that people who did better on the first exam continue to do better on the second.

(e) (5 points) Instead of constructing a confidence interval, suppose I wanted to test the null hypothesis that the coefficient on `midterm1` in Regression 2 is zero against the two-sided alternative. What is my test statistic? Would I reject the null at the 5% level? What R command should I enter to find the p-value for this test?

Solution: We know immediately that we can reject at the 5% level since zero is not contained in the confidence interval from part (d). The estimate is about 0.6 and the standard error is 0.12 so the test statistic is approximately 5. To calculate the p-value for the two-sided alternative, we use `2 * (1 - pnorm(5))`.

(f) (10 points) Suppose we want to predict a student's score on the second midterm. Based on the results given above, do you think we should take into account that student's sex? If so, how should we use this information? Explain your reasoning.

Solution: Based on the above results, it does not seem that a student's sex is helpful in predicting his or her score on the second midterm. In part (a), for example, we found no evidence that men and women score differently on average on the second midterm, using the results of Regression 1. In Regression 3 a 95% confidence interval for `male` is approximately -0.3 ± 5, so we find no evidence that we should allow a different intercept for men and women. In Regression 4, a 95% confidence interval for `male` is about 27 ± 38 while a 95% confidence interval for `male:midterm1` is about -0.34 ± 46 so we find no strong evidence that we need to allow a different intercept *or* slope for men versus women. Viewed from a predictive perspective, although the residual standard deviation is lowest in Regression 4, the difference between this and the value for Regression 2, which does not even consider sex, is minuscule: 0.01 points.