LPS ECON-103-601, Fall 2014

Final Exam 12.17.14

Exam Instructions:

- This exam is due by 20.00 on 12.17.14. No late exams will be accepted.
- You may submit your exam either under my door or the course e-mail (econstats103@gmail.com).
- You must submit the accompanying R Code to econstats103@gmail.com.
- You may not consult anyone for this exam and doing so will constitute a violation of academic integrity and result in a 0 at least.
- There are 30 marks. Marks are indicated on each question.
- Clearly show all work and indicate your final answer, but marks will be awarded based on a correct solution. Answers without solutions will not receive full credit.
- Your solutions or proofs should justify the steps taken and clearly explain the logic used in each step.
- State all theorems, propositions or definitions you apply. Prove any propositions you use not presented in class.

Academic Integrity Statement

I certify that all work is my own, and I have not consulted with anyone or in any way compromised the academic integrity of this exam. I understand that violations of academic integrity as well as failure to report observed violations will result in a zero mark and potential disciplinary action.

Signature: ______________________________
Date: ______________________________

Notice: Failure to follow instructions will result in a loss of at least 4 marks. No exams written in pencil or erasable ink will be considered for regrade. This does not apply to exams submitted via e-mail.

Q1 Probability [6 marks]

(a) [1 mark]

A bowl contains M black and $N - M$ white chips. Dustin draws n chips one at a time at random without replacement. Let P_k^n denote the probability a sequence of n chips consists of k black chips and $n - k$ white chips. Write a formula for P_k^n, stating what values can k take on.

(b) [1.5 marks]

State and prove the Law (or Rule) of Total Probability.

(c) [1.5 marks]

Sean has two six-sided dice in his pocket: one fair die and one loaded die. The fair die equal the usual probabilities, but the probability of getting a 6 when rolling the loaded die is 1/2. Suppose he reaches into his pocket and draws one of the two dice at random (both are equally likely to be drawn). He rolls this randomly chosen die and gets a 6. What is the probability that he drew the loaded die?

(d) [2 marks]

Suppose Luke rolls a pair of fair dice over and over so that he forms an infinite sequence of independent trials. What is the probability that a sum of 5 appears before a sum of 7?

Q2 Random Variables [5 marks]

(a) [1 mark]

Suppose the random variable ξ takes on a value α with probability ρ and a value β with probability $1 - \rho$. Show that $Var[\xi] = \rho(1 - \rho)(\alpha - \beta)^2$.

(b) [1 mark]

Assume the random variable z has the exponential distribution with $\lambda = 1$, i.e., its density function is $f_z(z) = \exp(-z)$ for $z \geq 0$ and 0 for $z < 0$. Define $u = \sqrt{z}$. Find the density function of u.

(c) [1 mark]

Suppose τ is a geometric random variable with parameter ν, i.e.

$$p_\tau(k) = \nu(1 - \nu)^{k-1} \quad k = 1, 2, ...$$

The geometric random variable describes the number of times one must perform a Bernoulli trial with success probability ν to obtain the first success. Show for $m, n \in \mathbb{Z}_{++}$ and $m < n$ that $\mathbb{P}r[\tau = n | \tau > m] = \mathbb{P}r[\tau = n - m]$ **and** interpret this “memory-less” property of the geometric random variable.

(d) [1 mark]

Using the definition of expected value, find an expression for $\mathbb{E}[\tau]$ in terms of ν.

(e) [1 mark]

Now derive the variance of the random variable τ (Hint: You may want to consider $\mathbb{E}[\tau(\tau - 1)]$).

Q3 Estimators [7 marks]

(a) [1 mark]

Suppose $t_1, ..., t_n$ are *i.i.d.* exponential(λ). Find the maximum likelihood estimator for λ. Note that $t \sim$ exponential(λ) $\Leftrightarrow f(t;\lambda) = \lambda \exp(-\lambda t)$, $t > 0$.

(b) [0.5 mark]

Is $\hat{\lambda}_{MLE}$ in part (a) unbiased general? Why or why not?

(c) [0.5 mark]

Let y_1 and y_2 be two independent and normally distributed with mean μ and variance σ^2. Show that the estimator $\sum_{i=1}^{2}(y_i - \bar{y})^2 \sim \sigma^2\chi^2_1$.

(c) [2 marks]

Write a function in R to construct a $X\%$ confidence interval for the estimator $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ where $y_i \sim$ *i.i.d.* with some unknown distribution. Your function should take X and the data y as inputs, draw critical values from the appropriate sampling distribution, and produces two numbers (the bounds) as output.

(e) [1 mark]

Show that the $MSE(\hat{\theta}) = bias(\hat{\theta})^2 + Var(\hat{\theta})$. Recall $MSE(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)^2\right]$ where θ_0 is the true parameter value.

(f) [2 marks]

Suppose $Var(y_i) = \sigma^2$ and $n > 1$. Show that the MSE for the *Theil-Schweitzer* estimator

$$\frac{1}{n+1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1}$$

is smaller than that of the unbiased estimator for the variance of y_i (s^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2}$$

and then say what you can conclude from this problem about unbiased estimators.

Q4 Linear Regression I [6 marks]

Consider the simple linear regression model

$$y_i = \alpha + x_i\beta + \varepsilon_i$$

with data $\{x_i, y_i\}_{i=1}^n$.

(a) [1 mark]

State the assumptions for the simple normal linear regression regression model **and** what properties they ensure for the OLS estimators of α and β (e.g. identification, unbiasedness, normality, etc.).

(b) [2 marks]

State the OLS estimators for α and β and derive the variances for these estimators (shown below). Be sure to state the assumptions you use at each step and prove any claims you use.

$$Var(\hat{\alpha}) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

$$Var(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

(c) [1 mark]

Suppose we estimate the model assuming $\alpha = 0$. Derive an expression for this new OLS estimator ($\tilde{\beta}$) of the restricted model. Also, derive an expression for its potential bias and state the conditions under which $\tilde{\beta}$ is an unbiased estimator for β.

(d) [2 marks]

Show that the MSE of $\tilde{\beta}$ is smaller than that of $\hat{\beta}$ if and only if the true (unknown) parameters α and σ_ε^2 satisfy

$$\frac{\alpha^2}{\sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \right)} < 1,$$

and then say what you can conclude from this problem about biased estimators.

Q5 Linear Regression II [6 marks]

I recently came across some anonymous data from last fall's ECON-103 class. It contained scores from the diagnostic exam, midterm, and an indicator that takes on a value of 1 if the student was active in the class *before* the midterm and is 0 otherwise. I ran a series of regression in R to determine how well the diagnostic and being active in the class predict midterm scores. The results are shown on the next page.

(a) [1 mark]

Use the regression results to construct an approximate 99% confidence interval for the difference of mean scores on the midterm between students who were active in the class and those who were not. Interpret and explain your results.

(b) [1 mark]

Is there any evidence that students who do well on the diagnostic test tend to do better on the midterm? If so, about how much better? Your claims should be based on statistical inference − not point estimates.

(c) [1 mark]

Is there evidence that even after controlling for math diagnostic test results, students who are active did better on the midterm?

(d) [1 mark]

Do the regression results provide any evidence that the relationship between math diagnostic test results and midterm scores differ according to whether or not a student was active?

(e) [1 mark]

Compare the predictive accuracy of the four regression models. How accurate is the most accurate model compared to the least accurate model? Which model would you choose to predict midterm scores and why? What information provided gives you a measure of the prediction accuracy (which differs from the fit of the regression)?

(f) [1 mark]

The R^2 for each regression is fairly small ($< 20\%$). Does this *necessarily* indicate a problem? Why or why not? If this necessarily indicates a problem, then describe the problem and potential solutions. If this does not necessarily indicate a problem, then provide a simple example (i.e. simulation) using R showing that it is not necessarily a problem (you must send your code for your example to count).

Regression 1:

```
lm(formula = midterm1 ~ active)
            coef.est coef.se
(Intercept) 66.75     2.37
active       9.19     3.55
---
n = 79, k = 2
residual sd = 15.69, R-Squared = 0.08
```

Regression 2:

```
lm(formula = midterm1 ~ diagnostic)
            coef.est coef.se
(Intercept) 47.81     7.72
diagnostic   0.34     0.11
---
n = 79, k = 2
residual sd = 15.45, R-Squared = 0.11
```

Regression 3:

```
lm(formula = midterm1 ~ active + diagnostic)
            coef.est coef.se
(Intercept) 44.16     7.56
active       9.00     3.37
diagnostic   0.33     0.11
---
n = 79, k = 3
residual sd = 14.87, R-Squared = 0.18
```

Regression 4:

```
lm(formula = midterm1 ~ active + diagnostic + active:diagnostic)
                  coef.est coef.se
(Intercept)       45.04     9.41
active             6.62    15.52
diagnostic         0.32     0.13
active:diagnostic  0.04     0.22
---
n = 79, k = 4
residual sd = 14.96, R-Squared = 0.19
```