Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Syllabus

Office: BA2 302M

Telephone: (407) 823-2078

E-Mail: Harry.Paarsch@UCF.edu

Office Hours: Fridays, 2:00-3:15 p.m.

Prerequisites: You must understand basic concepts of probability and statistics as well as regression. I shall review some of those concepts, but if you are unfamiliar with that material (or forgotten it), then you must make time to learn (or relearn) it on your own.

Textbooks: *The Foundations of Empirical Intuition*, by Harry J. Paarsch. Unpublished manuscript, 2017.

A Primer in Econometric Theory, by John Stachurski. Cambridge, MA: MIT Press, 2016.

An Introduction to Generalized Linear Models, 3rd Ed., by Annette J. Dobson and Adrian J. Barnett. Boca Raton: Chapman & Hall, 2008.

In All Likelihood: Statistical Modelling and Inference Using Likelihood, by Yudi Pawitan. Oxford: Oxford University Press, 2013

R in Action: Data Analysis and Graphics with R, Second Edition, by Robert I. Kabacoff. Shelter Island, New York: Manning, 2015.

R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, by Hadley Wickam and Garrett Grolemond. Sebastopol, CA: O'Reilly, 2017.

Requirements: Nine problem sets, a midterm examination, and a final examination.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Syllabus (continued)

Grading: Together, the problem sets will account for 25 percent of your grade, the midterm examination for 25 percent, and the final for the remaining 50 percent. If, however, you do better on the final than the midterm, then your final will count for 75 percent of your grade.

Lateness Policy: Problem sets must be submitted at the beginning of the class for which they are due. Because you can always e-mail me a problem set, no late ones will be accepted. In other words, late problem sets will be assessed the score zero. E-mailed problems sets must be in PDF format.

Group Work: You are encouraged to work on the problem sets with other students, but you must write up your answers by yourself. The examinations will be closed-book. Obviously, you must complete those alone.

Important: Students are expected to be familiar with the University's standards regarding academic integrity and academic misconduct, as well as the course of action that will be taken if a violation occurs; these links

`http://goldenrule.sdes.ucf.edu/`

and

`http://osc.sdes.ucf.edu`

provide such information. Information on accommodations for those with disabilities may be found at

`http://sds.sdes.ucf.edu/`

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Additional Information

This course is the first in a two-course sequence designed to introduce graduate students in business fields (such as accounting, economics, finance, and marketing) to methods of data science in general, and those of econometrics in particular.

That said, recently, the fields of data mining and machine learning have garnered considerable attention; some statisticians refer to the implementation of methods native to their field but used in data mining and machine learning as statistical learning. Thus, in addition, we shall also investigate the relationship between methods that econometricians employ and those used in data mining, machine learning, and statistical learning.

Many of the methods employed in data mining, machine learning, and statistical learning are closely related to those used regularly by researchers in business fields; others, however, are new to researchers in business. In addition, data miners and machine learners as well as statisticians approach model validation and testing in a much more principled way than social scientists or researchers in business. We shall borrow from these strengths.

One important by-product of completing this course successfully is you will then be able read on your own a standard, introductory book to the field of statistical learning, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, which was published by Springer of New York in 2013.

When I lecture, I presume that my students have read the work assigned for that day. To help you prepare for lectures, I have attached a list of the topics I intend to cover over the next fifteen weeks. I urge you to read avidly prior to attending lectures because we have a lot of ground to cover; you cannot afford to get behind in your reading.

Neither the lectures nor the textbooks, however, will be self-sufficient for an adequate understanding of the course. Parts of the course will involve material not in the textbooks and parts of the textbooks will not be discussed in class. Thus, you are expected to read widely.

Also, the textbooks contain many examples; you should attempt as many of these as you can. Only through practice will you master this material.

Irregular attendance or inattentiveness during class will almost surely result in a poor grade in the course; that is, success will require attention and participation. To this end,

please turn off and put away cell phones as well as laptops, tablets, and other electronic devices or toys during class, unless asked to do otherwise. Also, please notify me if you plan to be absent.

Nine problem sets are assigned for grading; one will be due every week or two during the semester; the due dates are noted on the problem sets as well as the attached timetable. The solutions to these problem sets must be submitted at the beginning of the class that they are due. Because you can always e-mail me a problem set, no problem sets will be accepted. In other words, late problem sets will be assessed the score zero.

The problem sets will sometimes require you to use statistical software. I recommend the freely available R system, a programming language and software environment for statistical computing and graphics. No specific knowledge of R is required; you can learn "on-the-job" so to speak. R is available for the three major operating systems—Windows, OS X, and Linux; you can install it on your computer easily. To download R, go to `https://www.r-project.org/` and follow the instructions. You may also find the integrated development environment (IDE) RStudio easier to use than R; to download it, go to `http://www.rstudio.org/`.

To help you learn R, I recommend that you purchase and read the second edition of Robert I. Kabacoff's book *R in Action: Data Analysis and Graphics with R*; this book is a wonderful introduction to the R system, written for novices. I expect you to work through the book on your own, but you are welcome to ask me questions about R during my office hours.

One way to help you reinforce what you have learned by reading Kabacoff's book is to install the R package `swirl` and then to work through the interactive tutorials; to learn more about `swirl` go to `http://http://swirlstats.com/`.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Timetable

Date	Day	Topic. DB:=Dobson & Barnett; P:=Paarsch; S:=Stachurski. PS due?
01/12	F.1	Introduction. **P**: Ch. 1; **DB**: Chs. 1–2; **S**: Chs. 1–3.
01/12	F.2	Probability and Statistics. **P**: Ch. 2–4; **S**: Chs. 4–5.
01/19	F.1	Least-Squares Regression. **DB**: Ch. 6; **P**: Ch. 4; **S**: Ch. 11. PS #1 due.
01/19	F.2	Properties of the LS Estimator. **P**: Ch. 4; **S**: Chs. 8, 9, & 12.
01/26	F.1	Inference using the LSE. **P**: Ch. 4; **S**: Ch. 10. PS #2 due.
01/26	F.2	ML Estimation with Gaussian Errors, **DB**: Ch. 6; **S**: Ch. 8, 9, & 13
02/02	F.1	Inference with Gaussian Errors, **DB**: Ch. 6; **S**: Chs. 9–10. PS #3 due.
02/02	F.2	Numerical Implementation of LSE/MLE. **S**: Ch. 3; **P**: Appendices A & B.
02/09	F.1	Exponential Family of Distributions. **DB**: Ch. 3.
02/09	F.2	General ML Estimation. **P**: Ch. 4; **S**: Ch. 6; **DB**: Ch. 4.
02/16	F.1	Inference with General ML Estimators. **P**: Ch. 4; **DB**: Ch. 5. PS #4 due.
02/16	F.2	Inference with General ML Estimators. **DB**: Ch. 5; **S**: Chs. 6 & 10.
02/23	F.1	Numerical Implementation of General MLEs. **P**: Appendix B.
02/23	F.2	Examples of the Exponential Family. Class notes.
03/02	F.1	Examples of the Exponential Family. Class notes. PS #5 due.
03/02	F.2	Resampling methods. **P**: Ch. 4 and Appendix C; **S**: Ch. 9.
03/09	F.1	Midterm examination, in class.
03/09	F.2	Midterm examination, in class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Timetable (continued)

Date	Day	Topic. DB:=Dobson & Barnett; P:=Paarsch; S:=Stachurski. PS due?
03/23	F.1	Resampling methods. **P**: Ch. 4 and Appendix C; **S**: Ch. 9.
03/23	F.2	Binary Logistic Regression. **DB**: Ch. 7.
03/30	F.1	Prediction. Class handout. PS #6 due.
03/30	F.2	Dependence. **P**: Ch. 2; **S**: Ch. 7
04/06	F.1	Poisson Regression. **DB**: Ch. 9. PS #7 due.
04/06	F.2	Poisson Regression. **DB**: Ch. 9.
04/13	F.1	Survival Analysis. **DB**: Ch. 10. PS #8 due.
04/13	F.2	Survival Analysis. **DB**: Ch. 10.
04/20	F.1	Survival Analysis. **DB**: Ch. 10. PS #9 due.
04/20	F.2	Course review.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #1

To prepare for this problem set, you should read the chapters entitled "Introduction" and "Language" as well as the one entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*.

1. Answer the following questions:

   a) Suppose U_1 and U_2 are independent. What is

   $$\text{cov}(U_1, U_2)?$$

   Be sure to demonstrate why this is true, not just list the answer.

   b) Two random variables Z_1 and Z_2 are uncorrelated. Are these two random variables independent? Why or why not? What does this mean $\text{cov}(Z_1, Z_2)$ equals? Provide an example.

   c) Find the minimum of the following function with respect to μ:

   $$S(\mu) = \sum_{n=1}^{N} (Y_n - \mu)^2.$$

   d) Suppose $\{Y_n\}_{n=1}^N$ are distributed jointly normal where a representative random variable Y_n has mean μ_n and variance σ^2, where the covariance between any pair (m, n) is $0.5\sigma^2$. Find the mean, variance, and distribution of

   $$B = \sum_{n=1}^{N} k_n Y_n$$

   where the $\{k_n\}_{n=1}^N$ are constants.

   e) Consider a random sample $\{Y_n\}_{n=1}^N$ from a normal distribution having mean μ and variance σ^2. What is the distribution of

   $$\frac{\bar{Y}_N - \mu}{\sigma}?$$

Be sure to provide all of the steps necessary to support your claims.

What is the distribution of

$$\frac{\sum_{n=1}^{N}(Y_n - \bar{Y}_N)^2}{\sigma^2}?$$

Be sure to provide all of the steps necessary to support your claims.

What is the distribution of

$$\frac{\sqrt{N}(\bar{Y}_N - \mu)}{\sqrt{\frac{\sum_{n=1}^{N}(Y_n - \bar{Y}_N)^2}{(N-1)}}}?$$

Be sure to provide all of the steps necessary to support your claims.

f) Consider a random sample $\{Y_n\}_{n=1}^{N}$ for a Bernoulli random variable having mean θ. Consider an estimator of θ based on the analogy principle—namely, the sample proportion

$$\bar{Y}_N = \frac{\sum_{n=1}^{N} Y_n}{N}.$$

Contrast the difference between what happens to the random variable $\bar{Y}_n$ and a linear transformation of that random variable

$$Z_N = \frac{(\bar{Y}_N - \theta)}{\sqrt{\frac{\theta(1-\theta)}{N}}}$$

as $N \to \infty$?

Due on 19 January 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #2

To prepare for this problem set, you should read the chapters entitled "Introduction" and "Language" as well as the one entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*, as well as the relevant chapters of *R in Action: Data Analysis and Graphics with R* and *R for Data Science: Import, Tidy, Transform, Visualize, and Model*. Also read the first part of Chapter 14 of *A Primer in Econometric Theory*.

1. On the course website, you will find a file `Galton.dat`. The data in this file were obtained from photographs of the pages from Galton's original notebooks taken by Colin Hanley and Louise Koo under the direction of Professor James A. Hanley of McGill University.[1] The `jpg`s of these photographs live at the following URL:

   `http://www.med.mcgill.ca/epidemiology/hanley/galton/`

   The information was then key-punched by hand into the file `Galton.dat` which has five columns and 934 rows. The first column is a family identification number, while the second column contains the height of the father in inches minus 60 inches (so a 9.0 means the father was five foot nine inches tall), the third column contains the height of the mother in inches, again minus 60 inches, while the fourth column is an indicator variable that equals one if this is a son and zero if this is a daughter. The fifth, and final, column contains the height of the off-spring, again in inches minus 60.

   a) Load these data into R and calculate the sample descriptive statistics concerning: i) the height of sons and ii) the height of daughters. That is, present the sample mean, median, minimum, maximum, lower quartile, upper quartile, range, inter-quartile range, and standard deviation for each group. When the data set is in this format, explain why you cannot do the same for the heights of fathers and mothers.

   b) Using R, plot graphs of the heights of fathers on the abscissa (x-axis) and the heights of sons on the ordinate (y-axis). Export this graph to a file in Portable Document Format (`PDF`) and then print it.

[1] See James A. Hanley's " 'Transmuting' women into men: Galton's family data on human stature," *The American Statistician*, 58 (2004), 237–243.

2. Consider a continuous random variable Y which has probability density function $f_Y^0(y)$ and cumulative distribution function $F_Y^0(y)$ defined by

$$F_Y^0(y) = \int_{-\infty}^{y} f_Y^0(u)\ \mathrm{d}u.$$

For $\{y_n\}_{n=1}^N$, a random sample of size N concerning Y, perhaps the most important summary statistic to calculate is the empirical distribution function (EDF). The EDF is the sample analogue of the cumulative distribution function (cdf). It is defined by

$$\hat{F}_Y(y) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(y_n \leq y)$$

where $\mathbb{I}(\mathsf{A})$ denotes the indicator function, equalling one when A obtains and zero otherwise.

   a) Using R, calculate the EDF of heights of sons and daughters. Then plot the two EDFs on the same graph, making sure you identify the EDF for sons that for daughters clearly. Export this graph to a `PDF` file and then print it.

   b) Using R, calculate the EDF of heights for fathers and sons and plot the two EDFs on the same graph, making sure you identify the EDF for fathers that for sons clearly. Export this graph to a `PDF` file and print it.

3. The virtue of high-level programming environments, like R, is that they allow a researcher to create functions that are custom-made for a particular purpose. For example, consider a continuous random variable Y that has population probability density function $f_Y^0(y)$ that one would like to estimate consistently at the point y. The sample relative frequency distribution (histogram) is an estimator of $f_Y^0(y)$, but it is nondifferentiable. Under suitable regularity conditions, for some sample $\{y_n\}_{n=1}^N$, the following estimator is both consistent and differentiable:

$$\hat{f}_Y(y) = \frac{1}{Nh}\sum_{n=1}^{N} \kappa\left(\frac{y - y_n}{h}\right)$$

where $\kappa(\cdot)$ is a *kernel* density function, having the following properties:

$$\kappa(u) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \kappa(u)\ \mathrm{d}u = h < \infty,$$

while h is a bandwidth parameter that must be strictly positive. An example of $\kappa(u)$ would be the *normal kernel*

$$\kappa(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right).$$

a) Write an R function that provides a kernel-smoothed estimate of the probability density function for a continuous random variable at a point. (Hint: first do this for a particular kernel function, but then try to make the choice of kernel function an option in your code.)

b) Use your kernel smoother to provide an estimate of the probability density functions of heights for fathers and sons over the entire range, and then depict these on the same graph. Export this to a `PDF` file and print it.

Due on 26 January 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #3

To prepare for this problem set, you should read the chapter entitled "Introduction" as well as the one entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*. Also read Chapter 6 of *An Introduction to Generalized Linear Models*, and Chapters 11 and 12 of *A Primer in Econometric Theory*.

1. Consider a variant on the regression that Galton first considered—namely,

$$Y_n = \beta_1 + \beta_2 x_n + U_n$$

where Y_n is the height of a son and x_n is his father's height. Suppose that the U_ns have the following properties:

$$\begin{aligned} \mathbb{E}(U_n|x_n) &= 0 \\ \operatorname{cov}(U_n, x_n) &= 0 \\ U_n &\sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

   a) Within this model, formulate the null hypothesis that sons are, on average, the same height as their fathers.

   b) Using the data set `Galton.dat`, test this hypothesis.

2. Now consider one of the regressions that Galton entertained—namely,

$$Y_n = \beta_1 + \beta_2 x_n + U_n$$

where Y_n is the height of a son, while x_n is the simple average of his mother's and his father's height. Suppose, now, that the U_ns have the following properties:

$$\begin{aligned} \mathbb{E}(U_n|x_n) &= 0 \\ \operatorname{cov}(U_n, x_n) &= 0 \\ U_n &\sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

a) Within this model, formulate the alternative hypothesis that, on average, son's are affected differently by their mother's height than by their father's height.

b) Using the data set `Galton.dat`, test the null hypothesis that a mother's height and a father's height have an equal partial effect.

Due on 2 February 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #4

To prepare for this problem set, you should read the chapters entitled "Introduction" and "Language" as well as the one entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*. Also read the entry in *The New Palgrave Dictionary*, by Arthur Lewbel, which is located on the course website.

1.) One of the oldest empirical problems in economics involves the estimation of Engel curves. This began in the nineteenth century with the research of Ernst Engel, a German economist. In preparation for this problem, read the entry in *The New Palgrave Dictionary* written by Arthur Lewbel, which is located on the course website.

In the file `Engel.dat`, which is located on the course website, you will find two series of data, the first being x_n, the total expenditures of unit i, while the second being y_n, the expenditures on food of unit n. Consider the following empirical specification:

$$Y_n = \mathbb{E}(Y_n|x_n) + U_n \quad n = 1, \ldots, N$$

where the U_ns are independently distributed error terms, having $\mathbb{E}(U|x_n)$ zero with finite variances, where $\text{cov}(U_n, X_n)$ is zero. In various parts below, a distributional assumption may be invoked concerning the U_ns, but that shall be explicitly noted at the time.

a) Estimate the following empirical specification by the method of least squares:

$$Y_n = \alpha + \beta x_n + U_n$$

and report both the least-squares standard errors and the Eicker-White standard errors.

b) Assuming that the $U_{i,n}$s are normally distributed, having constant variances, estimate the following four empirical specifications:

$$\begin{aligned} Y_n &= \alpha_1 + \beta_1 x_n + U_{1,n} \\ Y_n &= \alpha_2 + \beta_2 \log x_n + U_{2,n} \\ \log Y_n &= \alpha_3 + \beta_3 x_n + U_{3,n} \\ \log Y_n &= \alpha_4 + \beta_4 \log x_n + U_{4,n} \end{aligned}$$

by the method of maximum likelihood. Compare the fits of the four estimated regression functions on the same graph as well as with the actual data.

c) Again, assuming that the $U_{A,n}$s are normally distributed, having constant variances, consider the following empirical specification:

$$\frac{Y_n^\lambda - 1}{\lambda} = \theta_0 + \theta_1 \frac{x_n^\psi - 1}{\psi} + U_{A,n}.$$

Explain how this empirical specification nests the four in part b). Using each of the empirical specifications from part b) as the null, describe how you would test it against the above alternative using a likelihood-ratio test. What is "wrong" with using the likelihood-ratio test for the empirical specification listed in this part? How might one get around this problem?

d) Using the Gaussian kernel, calculate the kernel-smoothed estimate of the population regression function $\mathbb{E}(Y|x_n)$ at each point in the data set. Also, for each point in the data set, calculate

$$\tilde{v}_n = \mathbb{E}(\widehat{\log Y}|x_n) - \hat{\alpha} - \hat{\beta} \log x_n.$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the least-squares estimates from part b). Based on the $\{\tilde{v}_n\}_{n=1}^N$, test whether the parametric specification is rejected by the nonparametric one.

Due on 16 February 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #5

To prepare for this problem set, you should read the chapter entitled "Learning" as well as Appendix B in the manuscript *The Foundations of Empirical Intuition*. Also read Chapters 4 and 5 of *An Introduction to Generalized Linear Models*.

1. Consider a discrete random variable Y having probability mass function:

$$p_Y(y;\theta) = \frac{-\theta^y}{y \log(1-\theta)} \qquad y = 1, 2, \ldots \tag{1}$$

where the unknown parameter θ lives in the open unit interval $(0, 1)$.

a) Prove that

$$\sum_{y=1}^{\infty} p_Y(y;\theta) = 1.$$

[Hint: consider the Maclaurin series expansion of the function $\log(1+x)$ and substitute in $x = -\theta$.]

b) Find $\mathbb{E}(Y;\theta)$. [Hint: $\sum_{i=1}^{\infty} \rho^i = \frac{\rho}{1-\rho}$.]

c) Find $\mathbb{V}(Y;\theta)$. [Hint: recall the proof of the moments of the geometric distribution.]

d) Demonstrate that this distribution is a member of the exponential family of distributions. Find the sufficient statistic for an independent and identically-distributed sample of size N, namely, $\{Y_1, Y_2, \ldots, Y_N\}$.

e) Define $\hat{\theta}$, the maximum likelihood estimator of θ^0, the true value of the parameter θ, in terms of the sufficient statistic.

f) Show that the condition that defines $\hat{\theta}$ is a monotonic function of θ. [Hint: draw a graph.]

g) Set up the recursion you would use in order to employ the method of Newton-Raphson to solve for $\hat{\theta}$.

h) Demonstrate that $\hat{\theta}$ is a consistent estimator of θ^0.

i) Find an approximation to the variance of $\hat{\theta}$.

j) Characterize the asymptotic distribution of $\hat{\theta}$, explaining why this is so.

After considerable effort, a researcher has obtained a random sample of one thousand measurements on Y. These data are summarized in Table 1.

Table 1

Observed Frequency Distribution for a Logarithmic Series Model

Y	1	2	3	4	5	6	7	8	9+
Observed Frequency	710	175	74	23	10	4	2	1	1

k) Calculate the maximum likelihood estimate of θ^0 using the above data.

l) At size 0.05, test the hypothesis

$$\begin{aligned} \mathrm{H}_0 &: \theta = 0.50 \\ \mathrm{H}_1 &: \theta \neq 0.50. \end{aligned}$$

m) At size 0.10, test the hypothesis

$$\begin{aligned} \mathrm{H}_0 &: \log\theta = -0.70 \\ \mathrm{H}_1 &: \log\theta \neq -0.70. \end{aligned}$$

n) At size 0.05, ignoring that $\hat{\theta}$ is estimated, but acknowledging that only counts nine and above are recorded, test whether the empirical frequency is consistent with equation (1), which is referred to as the logarithmic series distribution.

Due on 2 March 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #6

To prepare for this problem set, you should read the chapter entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*. Also read Chapters 4, 6, 8, and 9 of *A Primer in Econometric Theory*.

1.) Consider an exponential random variable having hazard-rate parameter θ, so the probability density function is

$$f_Y(y|\theta) = \theta \exp(-\theta y) \quad y > 0, \theta > 0.$$

Now, the population first moment of Y is

$$\mathbb{E}(Y) = \int_0^\infty y f_Y(y|\theta^0) = \int_0^\infty y\theta^0 \exp(-\theta^y)\ dy = \mu_1(\theta^0) = \frac{1}{\theta^0}.$$

In this case, the maximum-likelihood estimator of θ^0 is

$$\hat{\theta} = \frac{1}{M_1} = \frac{N}{\sum_{n=1}^N Y_n}$$

where M_1 is the sample first moment of the data—that is,

$$M_1 = \frac{\sum_{n=1}^N Y_n}{N}$$

Now, M_1 converges in probabilty to $\mu_1(\theta^0)$ by a law of large numbers, so $\hat{\theta}$ converges in probability to θ^0 by the continuous mapping theorem. In small samples, however, $\hat{\theta}$ is biased.

a) Using the δ method, find an expression for the order of the bias of $\hat{\theta}$. Express the bias in terms of a percentage as a function of the sample size N.

b) Use the jackknife to estimate the bias of $\hat{\theta}$.

c) For a sample of $N = 25$, when $\theta^0 = 1$, simulate the jackknife estimator and compare its behavior to the answer in part a).

2.) Consider the following empirical specification:

$$Y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 y_{t-1} + U_t \tag{S}$$

where the x_t's are strictly exogenous and where U_ts are distributed independently and normally, having mean zero and variance σ^2 for $t = 0, 1, \ldots, T$. The *common factor hypothesis* is the following:

$$h(\boldsymbol{\beta}) = \beta_3 + \beta_2 \beta_4 = 0 \tag{C}$$

for $\boldsymbol{\beta}$ equal $(\beta_1, \beta_2, \beta_3, \beta_4)^\top$. Denote the least-squares estimator of $\boldsymbol{\beta}$ from a sample of size T by $\hat{\boldsymbol{\beta}}$. Assume the conditions necessary to obtain:

$$\operatorname*{plim}_{T \to \infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^0$$

where $\boldsymbol{\beta}^0$ is the true value of $\boldsymbol{\beta}$ and

$$\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{d}} \mathcal{N}\left[\boldsymbol{\beta}^0, \mathbb{V}(\hat{\boldsymbol{\beta}})\right].$$

a) Propose a consistent estimator of $h(\boldsymbol{\beta}^0)$ and derive its asymptotic distribution.

b) Describe the Wald statistic for testing (C) and state its asymptotic distribution when the common factor hypothesis is true.

c) For this testing problem, is there any advantage in using the Wald statistic over either the Likelihood Ratio or the Lagrange Multiplier statistics? Are there any limitations to the Wald statistic?

Due on 30 March 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #7

To prepare for this problem set, you should read the chapter entitled "Language" as well as the one entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*. Also read the Chapter 7 of *An Introduction to Generalized Linear Models*.

1. An oldster is trying to estimate the distance of his daily walk, but the odometer in his car only measures to the nearest kilometre. For example, the starting value could be

   ```
   12345
   ```

   while the ending value could be

   ```
   12351
   ```

   implying that his walk could have started at 12345.00 and ended at 12351.99 or started at 12345.99 and ended at 12351.00, a difference of 1.98 kilometres.

   a) Describe an estimation strategy that the oldster can use to obtain any level of accuracy he might want—for example, one decimal place, two, three, and so forth.

   b) Assuming that the measurement error is distributed uniformly, calculate the mean squared-error of the rule proposed in part a), for different sampling scenarios.

   Now consider recording values of the odometer at the start and the finish on a random sample of n days. Assume, again, that the measurement error for any given odometer reading is uniform.

   c) Derive the distribution of the difference of two readings. Calculate the mean and the variance of the measurement error associated with this difference.

   d) Demonstrate that the average of odometer readings over the n observations is an unbiased measure of the true distance walked and calculate its variance. Characterize the asymptotic distribution of the sample average and provide a justification for this result.

   e) Explain why the method of part a) is probabilistically more accurate than that of part d). Hint: characterize the order of the accuracy (as a function of n) of each estimator.

2. Perhaps the greatest success story in data science involves binary prediction, for example, predicting whether a student will pass an examination using information concerning the student's grade-point average, major, class year, and so forth.

   In general, a data set will be composed of labels, denoted by y. Often referred to as the dependent variable by statisticians as well as social scientists, y can take on either `True` (1) or `False` (0). In addition, some features created from the information known for each student, collected in a vector denoted by $\mathbf{x}$ are typically available, too.

   Among social scientists the variables in $\mathbf{x}$ are often referred to as *covariates*, whereas among data miners and machine learners the variables in $\mathbf{x}$ are referred to as *features*. A data set having N observations for social scientists, and examples for data miners and machine learners, is then $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$.

   a) Using the file `PassFail.dat`, which has $10,000$ observations in the following format:

   ```
   y x1 x2 x3 x4 x5 x6
   ```

   write an R script that does the following: (1) read the data into a data frame; (2) from the data frame, determine the number of observation; (3) use the random number generator in R to select a training data set that comprises sixty percent of the entire data set; reserve the complementary portion of the entire data set for the testing data set.

   b) Use the `glm()` function in R to train a logistic regression model using the features and labels provided in the training data set.

   c) Score the observations in the test data set using the strategy described in the "Prediction" handout.

   d) Using the `ROCR` package in R, create the receiver operating characteristic curve, plot that curve, and then print it.

Due on 6 April 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #8

To prepare for this problem set, you should read the chapter entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*, as well as the class handout. Also read Chapter 5, 7, and 13 of *A Primer in Econometric Theory*.

1. Put somewhat simply (and perhaps too crudely), the conventional wisdom among physicians, at least in the 1950s and 1960s, was that the tonsils are physiologically superfluous and that their presence leads to a higher risk of infections in the throat. To avoid these infections, it was routine to remove the tonsils using a simple surgical procedure. Recently, some physicians have pointed out that the tonsils are an intricate part of the immune system and to remove them can upset a delicate equilibrium. The empirical question, from a health economist's perspective, is: Do tonsilectomies affect the incidence and dynamics of throat infections? A second related question is: Are tonsilectomies a cost effective way of reducing throat infections?

   To develop an empirical framework within which to address these questions, consider a simple two-state *mover-stayer* Markov model. In this model, in any period t and for the k^{th} individual, assume two states exist—sickness ($X_t^k = 1$) and health ($X_t^k = 0$). Suppose that the transition probabilities from sickness in period t to sickness in period $t+1$ as well as from health in period t to health in period $t+1$ are constant, but that they may vary across agents depending on whether the individual has had a tonsilectomy. The transition matrix for those who have no tonsils (N) is

$$\mathbf{\Pi}^N = \begin{pmatrix} \pi_{00}^N & 1 - \pi_{00}^N \\ 1 - \pi_{11}^N & \pi_{11}^N \end{pmatrix},$$

   while for those with tonsils (W) it is

$$\mathbf{\Pi}^W = \begin{pmatrix} \pi_{00}^W & 1 - \pi_{00}^W \\ 1 - \pi_{11}^W & \pi_{11}^W \end{pmatrix}.$$

   Consider a random sample of n_N individuals who have no tonsils and n_W individuals with them. Assume that each of these individuals has been observed for $t = 0, 1, \ldots, T$ periods and that for each it is known whether that person had a throat infection ($X_t^k = 1$) or not ($X_t^k = 0$) in period t.

a) Calculate the equilibrium proportion of time that the type W and the type N people spend having throat infections. Denote these π_1^W and π_1^N.

b) Under what conditions are the two proportions π_1^W and π_1^N equal.

c) Find the contribution to the likelihood function of the k^{th} person between period $t-1$ and t.

d) Explain how you would estimate the transition probabilities from the sample of type W people; *i.e.*, how would you estimate

$$(\pi_{00}^W, \pi_{11}^W).$$

e) Describe a method of estimating the transition probabilities for the entire sample (both the type W and type N people), introducing an indicator variable P_t^k where

$$P_t^k = \begin{cases} 0 & \text{if person } k \text{ has no tonsils in period } t \\ 1 & \text{if person } k \text{ has tonsils in period } t. \end{cases}$$

f) Explain how you would decide whether tonsilectomies have affected the incidence and dynamics of throat infections.

g) Suppose you find that tonsilectomies have actually reduced throat infections by a small, but statistically significant amount. Provide an economic rationale for why they still should not be performed routinely.

h) Suppose that in some country A (which is similar to the US culturally, demographically, geographically, and meteorologically) tonsilectomies are prohibited. In this country, an estimate of π_{11}^A is larger than estimates of either π_{11}^N or π_{11}^W and an estimate of π_{00}^A is smaller than estimates of either π_{00}^N or π_{00}^W. How would you interpret this evidence? Would this evidence change your answer to part g)?

i) On the course website, the file `Tonsils.dat` has five columns. The first four are indicator variables where

$$\texttt{Dij} = \begin{cases} 1 & \text{if } X_{t-1}^k = i \text{ and } X_t^k = j \\ 0 & \text{otherwise,} \end{cases}$$

while the fifth is a dummy variable where

$$\texttt{P} = \begin{cases} 0 & \text{if person } k \text{ has no tonsils in period } t-1 \\ 1 & \text{if person } k \text{ has tonsils in period } t-1. \end{cases}$$

Using the `glm()` command in R, estimate the parameters of the model you outlined in part e).

j) Test the hypothesis that the tonsils have no effect on the incidence and dynamics of throat infections using a likelihood-ratio test at size 0.05.

Due on 13 April 2018 at the beginning of class.

Spring 2018 **Harry J. Paarsch**

University of Central Florida
Department of Economics

ECO 6424
Econometrics I

Problem Set #9

To prepare for this problem set, you should read the chapter entitled "Learning" in the manuscript *The Foundations of Empirical Intuition*. Also read Chapter 9 of *An Introduction to Generalized Linear Models*.

1. In 1898, the Russian economist Ladislaus Bortkiewicz published a book, entitled *Das Gesetz der keinem Zahlen*, in which he included his famous example that illustrated the Poisson distribution: the annual deaths by horse kicks in the Prussian Army from 1875–1894, which included data concerning 14 different army corps as well as the Guard Corps. The Poisson distribution has probability mass function

$$p_Y(y;\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!} \quad y = 0, 1, 2, \ldots; \ \lambda > 0$$

where

$$\mathbb{E}(Y;\lambda) = \mathbb{V}(Y;\lambda) = \lambda.$$

On the course website, you will find `PrussianArmy.dat`, which contains three fields: first, the year; next, the corps; finally, the number of deaths. The corps identifiers include `G` for the Guard Corps, as well as the Roman numerals `I`, `II`, `III`, `IV`, `V`, `VI`, `VII`, `VIII`, `IX`, `X`, `XI`, `XIV`, and `XV` for the other Corps—so no `XII` or `XIII`.

Consider a random sample of $n = 1, \ldots, N$ observations $\{y_n\}_{n=1}^N$.

   a) Derive the maximum likelihood estimator of the parameter λ. Find the variance of this estimator. Demonstrate that this estimator is consistent and finds its asymptotic distribution. Using the data contained in the file `PrussianArmy.dat`, estimate the parameter λ and calculate its standard error.

   b) Examine the fitted residuals of based on the maximum-likelihood estimates—plotting them in different colors and symbols based on the values for the different corps in field two of the file `PrussianArmy.dat`.

Consider a vector of corps-specific dummy variables $\mathbf{x}$.

   c) Explain how you would introduce $\mathbf{x}$ into the Poisson model. Define the conditions for the maximum likelihood estimator in this model, given the way in which you have introduced the covariate vector.

d) Using all of the data in `PrussianArmy.dat` and the `glm()` command in R, calculate the maximum likelihood estimates and their standard errors. Test the hypothesis that none of the dummy variables matters.

Due on 20 April 2018 at the beginning of class.