Econ 21410 - Problem Set V

Selection Part II*

John Eric Humphries†

May 14, 2015

This homework should be done in LaTeX The homework will be graded on correctness, but will also heavily weight clarity and professionalism. Its better to not do some of the harder parts than to turn in an incomprehensible document. Your R script as well as a log-file should be submitted. Alternatively, use knitr to print your code inline in your latex document.

SUBMISSION: The homework must be emailed to Oliver and myself by 2p.m. Thursday, May 7th. The email must include a pdf with the filename lastname_pset5.pdf and R code called lastname_pset5_code.R where "lastname" should be replaced with your last name. The subject of your email should be [ECON 21410: pset5 submission]

If you are struggling, please use the github page to ask for help!* Remember that asking and answering questions on our github page, coming to office hours to ask questions, and contributing to the class wiki are all worth participation credit, which is 10% of your grade in this class.

In this homework we will do a short MLE exercise, a couple integration problems, then a short research proposal.

*Please email johneric@uchicago.edu and obrowne@uchicago.edu if you have questions.
†Please email johneric@uchicago.edu and obrowne@uchicago.edu if you have questions.

1 MLE: probit model

In this section, you will write up your own maximum likelihood estimation for a probit model.

Here is some code to generate results from a probit model:

```
GenerateDataProbit = function(n){
    # Generates data with from a simple linear model
    # args: n - number of observations to generate
    # output: A list containing a Y vector and an X matrix
    x1     <- rnorm(n,-2,.5)
    x2     <- rnorm(n,2.5,0.2)
    X      <- cbind(1,x1,x2)
    theta2 <- 5
    eps    <- rnorm(n,0,1)
    beta   <- matrix(cbind(0.5,3,2),3,1)
    Ystar   <-  X %*% beta + eps
    Y       <- Ystar>0
    colnames(X) <- c("const", "x1","x2")
    colnames(Y) <- c("Y")
    return(data.frame(Y,Ystar,X))
}
```

Note I could solve this problem using Rs built in "glm()" command with:

```
data = GenerateDataProbit(100)
glm(Y ~ x1 + x2, family=binomial(link="probit"), data=data)


##
## Call:  glm(formula = Y ~ x1 + x2, family = binomial(link = "probit"),
##     data = data)
##
## Coefficients:
## (Intercept)           x1           x2
##      6.2971       4.5621       0.8373
##
## Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
## Null Deviance:     136.7
## Residual Deviance: 46.51  AIC: 52.51
```

which may be helpful for checking your results.

In case you feel unfamiliar with the Probit model, please review these additional resources:

- https://en.wikipedia.org/wiki/Probit_model
- https://www.youtube.com/watch?v=WflqTUOvdik
- http://faculty.smu.edu/tfomby/eco6352/Notes/Logit%20and%20Probit%20Notes.pdf

Please complete the following:

1. Write the equation for the log-likelihood. How many variables are you solving for?
2. $\mathcal{L}_n(\beta) = \sum_{i=1...n}(Y_i \log(\Phi(X_i\beta)) + (1 - Y_i) \log(1 - \Phi(X_i\beta))$
3. Write a function that takes two inputs (1) a list of parameters and (2) a data frame and returns the log-likelihood.
4. Write the mathematical equations for the gradient (i.e. the first derivatives of the log-likelihood function).
5. $\frac{\mathcal{L}_n(\beta)}{\beta} = \sum_{i=1...n}\left(Y_i \frac{\phi(X_i\beta)}{\Phi(X_i\beta)} + (1 - Y_i)\frac{\phi(X_i\beta)}{1-\Phi(X_i\beta)}\right) X_i$
6. Write a function that takes the same 2 inputs as part 2 and returns the gradient.
7. Use the "optim()" function to optimize your log likelihood using the default "Nelder-Mead" method.
8. Do this again but use the "BFGS" method.
9. Do this again, but use the "BFGS" method and use your gradient function as well.
10. Use the microbenchmark package to compare how long the three methods above take.

The example from class posted on my website can be used a a guide for many of these steps.

```
# Section 1: Probit Model
#========================

#Question 1.2
LogLik <- function(beta,data) {
    X = as.matrix(data[,c('const','x1','x2')])
    Y = data[,'Y']
    LogLik <- sum( Y * log( pnorm(X %*% beta) ) + (1-Y) * log( 1-pnorm(X %*% beta) ) )
    return(LogLik)
}

#Question 1.4
Grad <- function(beta,data) {
    X = as.matrix(data[,c('const','x1','x2')])
    Y = data[,'Y']
    grad = t(Y * dnorm(X %*% beta)/pnorm(X %*% beta) -
                 (1-Y) * dnorm(X %*% beta)/(1-pnorm(X %*% beta))) %*% X
    return(grad)
}

#Question 1.5
optim(c(0,0,0),LogLik,data=data, control=list(fnscale=-1))


## $par
## [1] 6.2968074 4.5623227 0.8375366
```

```
##
## $value
## [1] -23.25688
##
## $counts
## function gradient
##      172       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
#Question 1.6
optim(c(0,0,0),LogLik,data=data, method = 'BFGS', control=list(fnscale=-1))
```

```
## $par
## [1] 6.2976281 4.5621902 0.8370892
##
## $value
## [1] -23.25688
##
## $counts
## function gradient
##       29       15
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
#Question 1.7
optim(c(0,0,0),LogLik,data=data, gr=Grad, method = 'BFGS', control=list(fnscale=-1))
```

```
## $par
## [1] 6.297625 4.562185 0.837087
##
## $value
## [1] -23.25688
##
## $counts
## function gradient
##       29       15
##
## $convergence
## [1] 0
```

```
##
## $message
## NULL


#Question 1.8
microbenchmark(
    optim(c(0,0,0),LogLik,data=data, control=list(fnscale=-1)),
    optim(c(0,0,0),LogLik,data=data, method = 'BFGS', control=list(fnscale=-1)),
    optim(c(0,0,0),LogLik,data=data, gr=Grad, method = 'BFGS', control=list(fnscale=-1))
)


## Unit: milliseconds
##
##                                     optim(c(0, 0, 0), LogLik, data = data, control = list(
##                   optim(c(0, 0, 0), LogLik, data = data, method = "BFGS", control = list(
##  optim(c(0, 0, 0), LogLik, data = data, gr = Grad, method = "BFGS",      control = list(
##        min        lq      mean    median
##  24.905434 25.420832 26.120038 26.259059
##  17.265786 17.452905 18.116907 17.854928
##   6.818184  6.916419  7.541881  6.994542
##         uq      max neval
##  26.590348 30.18558   100
##  18.832925 19.95755   100
##   7.213095 39.60238   100


#=========================
```

2 Integration

Please complete the following exercises:

1. Numerically evaluate the integral of $3 * x^2 - 12 * x + sin(x)$ between -10 and 10 using R's "integrate()" command.

2. Numerically evaluate the integral of $\frac{e^x}{1+e^{x^2}} + 0.5 * log(x)$ for x between 1 and 100 using R's "integrate()" command.

```
# Section 2: Integration
#=========================

#Question 2.1
integrate(function(x) 3*x^2-12*x+sin(x), lower = -10, upper= 10)


## 2000 with absolute error < 2.3e-11
```

```r
#Question 2.2
integrate(function(x) exp(x)/(1+exp(x^2))+0.5*log(x), lower = 1, upper= 100)


## 181.2273 with absolute error < 0.017
```

3 Research Proposal

Write a 2-4 page research proposal that explains what you would like to do your project on, why it is interesting/relevant, what data you plan to use, and what type of analysis you plan to complete. Make sure that you answers each of the following questions.

1. Propose your research idea/question in 1-5 sentences. This should encompass what you want to look at and why. It should use minimal jargon.

2. In 1-5 sentences provide some motivation on why this is a good problem (plots from the news, prelimnary results showing large changes, etc work well, but mostly, make sure you can explain why your research idea may be interesting or matters).

3. In more detail lay out your research idea in 1/2 to 1 page.

4. What data will you use for your project? Be as specific as possible. Double check that it has the right time-range and variables available.

5. What analysis do you plan to use for your project? If you plan to run regressions, please explain what regression you plan to run in as much detail as you can at this point.

6. Are there related papers you know of? Provide a very brief summary of the one or two most relevant papers.

We realize some of these sections may be rough. If you remain stuck, please contact Oliver and I a few days before the homework is due.