STAT 352: Data Modeling and Inference

Homework 3. Due Thursday, February 10th. I encourage you to type all of your solutions, though this is not necessary. However, you must scan (or photograph) any handwritten portions and upload the files to Canvas. For questions that require `R` code, you must turn in your `R` code on Canvas. Your code must in a .Rmd file.

Question 1 (Computation): *Preamble:* We have seen a number of loss functions in this course that provide measures of center as their optimal action. Most of the ones that we have seen involve two parameters, the location of the center and a scaling parameter (only not squared error loss and absolute error loss, which are scale-free). Many of these can also be modified to include a tail-shape parameter. In this question, you are going to investigate the inclusion of such a parameter into the logistic loss function. The (shifted and scaled) logistic density is given by

$$f(y;\mu,\delta) = \frac{1}{\delta}\left(\exp\left(\frac{y-\mu}{2\delta}\right) + \exp\left(-\frac{y-\mu}{2\delta}\right)\right)^{-2},$$

where μ is a location parameter and $\delta > 0$ is a scale parameter. The (shifted and scaled) Type III logistic density is given by

$$f(y;\mu,s,\alpha) = \frac{1}{\text{Beta}(\alpha,\alpha)}\frac{1}{\delta}\left(\exp\left(\frac{y-\mu}{2\delta}\right) + \exp\left(-\frac{y-\mu}{2\delta}\right)\right)^{-2\alpha},$$

where $\alpha > 0$ and $\text{Beta}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function. The parameter α is a shape parameter that controls the tails of the distribution. The logistic distribution is a special case of the Type III logistic distribution with $\alpha = 1$.

Useful functions: There are built-in functions for the logistic distribution (`?dlogis`) that we can use when trying to fit a logistic distribution to some data. Unfortunately, there are no built in functions for the Type III logistic distribution. In order to not make this HW too difficult, I am going to provide the standard suite of functions for densities, probabilities, quantiles, and random number generators for the Type III logistic distribution.

```
dlogisIII=function(x,location=0,scale=1,shape=1,log=FALSE){
  x = (x-location)/scale
  log_y = ifelse(x<0,x-log(1+exp(x)),-log(1+exp(-x)))
  log_1my = ifelse(x<0,-log(1+exp(x)),-x-log(1+exp(-x)))
  out = -log(scale)-lbeta(shape,shape) + shape*(log_y+log_1my)
  if(log){return(out)}else{return(exp(out))}
}
plogisIII=function(q,location=0,scale=1,shape=1,lower.tail=TRUE,log.p=FALSE){
  q = (q-location)/scale
  Q = ifelse(q<0,exp(q)/(1+exp(q)),1/(1+exp(-q)))
  return(pbeta(Q,shape,shape,0,lower.tail,log.p))
}
qlogisIII=function(p,location=0,scale=1,shape=1,lower.tail=TRUE,log.p=FALSE){
  Q = qbeta(p,shape,shape,0,lower.tail,log.p)
  q = ifelse(Q<0.5,log(Q)-log(1-Q),-log(1/Q-1))
```

```
  return(q*scale+location)
}
rlogisIII=function(n,location=0,scale=1,shape=1){
  b = rbeta(n,shape,shape,0)
  x = ifelse(b<0.5,log(b)-log(1-b),-log(1/b-1))
  return(x*scale+location)
}
```

Notice that these functions are taking advantage of a transformation to a Beta distribution (which we will talk about later in the class). Also, they are DUMB functions; they do not do any error catching (if you give bad input, they are not going to try to stop you).

I will also provide an MLE function that uses the built in `dlogis` function and uses transformations of the parameters for using `optim`.

```
mle_logis = function(y){
  #function to optimize using optim
  f = function(theta){
    mu = theta[1]
    delta = exp(theta[2]) #note the transformation
    sum(dlogis(y,location=mu,scale=delta,log=TRUE))
  }
  #using optim - starting at a guess of mu= and log(delta)=0 -- par=c(0,0)
  opt_out = optim(par=c(0,0),f,control=list(fnscale=-1,maxit=1000,reltol=1e-8))
  par = opt_out$par
  par[2] = exp(par[2]) #transforming to delta
  names(par) = c("mu","delta") #NAMES ARE GOOD
  out = list(theta=par,loglik=opt_out$value)
  return(out)
}
```

These functions are all in the `HW3_352Fa2020_helper_code.R` on Canvas.

Do the following: We are going to analyze a dataset on melanoma tumor thickness. To load the dataset, use the code

```
library(boot); data(melanoma); y = log(melanoma$thickness)
```

a) Write an MLE function similar to those from the `likelihoods.pdf` from class that computes the MLE for the Type III logistic model. Make sure to use transformations to deal with the fact that the scaling δ and the shape α are positive.

b) Compute the MLE under for the data `y`. Plot the density of the fitted model (use `dlogisIII`, the fitted parameter values, and some sequence of values you want the density at) with a density plot of the data. Does the model appear to capture the data well?

c) Do a goodness of fit test for the fitted Type III logistic distribution using `ks.test`. Your line of code should look something like

```
ks.test(y, plogisIII, location = mu, scale = delta, shape = alpha)
```

where `mu, delta, alpha` are the appropriate values from your MLE fit for the Type III Logistic distribution. Do you think that this model is adequate for modeling the data?

d) Perform tasks b) and c) using the `mle_logis` function that I provided and the built-in `dlogis` and `plogis` functions in `R` .

e) Compare the log-likelihood evaluated at the MLE for the Logistic and Type III logistic models. Which model has higher log-likelihood? Which model is better at capturing the data? Do you think it was worth adding the shape parameter α to the Logistic distribution for modeling this dataset?

Question 2 (Theory): *Preamble:* We are going to think about the log-likelihood in the context of five different models for different kinds of data. For each question, I will give you the density (mass) function for a single data point. Your task is to compute the log-likelihood for an observed data vector. You will notice that these are all examples where the sample mean is sufficient for the parameter, as evidenced by the part of the log-likelihood function that I have sectioned off as $n \times [\cdots]$.

Useful formulas and facts:

$$\begin{aligned}
\exp(-a) &= 1/\exp(a) \\
\exp(a+b) &= \exp(a)\exp(b) \\
\exp(\textstyle\sum a_i) &= \prod \exp(a_i) \\
\log(1/a) &= -\log(a) \\
\log(ab) &= \log(a) + \log(b) \\
\log(\textstyle\prod a_i) &= \sum \log(a_i) \\
\log(a^b) &= b\log(a) \\
\exp(a)^b &= \exp(ab) \\
\log(\exp(b)) &= b \\
\exp(\log(a)) &= a \\
k! &= 1 \times 2 \times 3 \times \cdots \times k \\
\tbinom{n}{k} &= \tfrac{n!}{k!(n-k)!}
\end{aligned}$$

Do the following:

a) A random variable Y follows a Bernoulli(p) distribution if Y can take values only 0 or 1 and $p = P(Y = 1)$. This is the most basic model for $0-1$ events like disease diagnosis or whether a machine will work properly when you turn it on. The mass function for a Bernoulli(p)

$$f(y; p) = p^y(1-p)^{1-y}.$$

The parameter p must be between 0 and 1. Let $Y_1, \ldots, Y_n \overset{iid}{\sim}$ Bernoulli(p) with realizations $y_1, \ldots, y_n$. Show that the log-likelihood for p is

$$\ell(p; y_1, \ldots, y_n) = n \times [\bar{y}\log(p) + (1-\bar{y})\log(1-p)].$$

b) A random variable K follows a Negative-Binomial(r, p) distribution if K counts the number successes are observed before the r-th failure in independent and sequential Bernoulli(p) trials.

The random variable K can take any non-negative integer value and r is a fixed positive integer. This is often used to model the number of times a certain type of machine will function properly in sequential tasks before you might want to think about replacing or repairing it. The mass function for a Negative-Binomial(r, p) random variable is

$$f(k; p) = \binom{r+k-1}{k}(1-p)^r p^k.$$

The parameter p must be between 0 and 1. Let $K_1, \ldots, K_n \overset{iid}{\sim}$ Negative-Binomial(r, p) with realizations $k_1, \ldots, k_n$. Show that the log-likelihood for p is

$$\ell(p; k_1, \ldots, k_n) = \sum_{i=1}^{n} \log\left(\binom{r+k_i-1}{k_i}\right) + n \times \left[\bar{k}\log(p) + r\log(1-p)\right].$$

c) A random variable $X > 0$ follows an Exponential(λ) distribution if it is memoryless, which means that $P(X > t+s|X > t) = P(X > s)$. As a waiting time to an event, this means that if you have already waited for t minutes then the probably of waiting a further s minutes is the same as the probability of waiting s minutes if you are just starting to wait. It is often used as a simple first model for waiting times for things like internet queues. The density function for an Exponential(λ) random variable is

$$f(x; \lambda) = \lambda \exp(-\lambda x).$$

The parameter λ must be positive. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Exponential(λ) with realizations $x_1, \ldots, x_n$. Show that the log-likelihood for λ is

$$\ell(\lambda; x_1, \ldots, x_n) = n \times \left[\log(\lambda) - \lambda\bar{x}\right].$$

d) A random variable $G > 0$ follows a Gamma(m, λ) distribution if G is the total amount of waiting time for m independent and sequential events to occur where the waiting time for each event is distributed as Exponential(λ). An example would be something like the total amount of time for m customers to get through a queue. The density function for an Gamma(m, λ) random variable is

$$f(g; \lambda) = \frac{\lambda^m}{(m-1)!} g^{m-1} \exp(-\lambda g).$$

The parameter λ must be positive and m is a fixed positive integer. Let $G_1, \ldots, G_n \overset{iid}{\sim}$ Gamma(m, λ) with realizations $g_1, \ldots, g_n$. Show that the log-likelihood for λ is

$$\ell(\lambda; g_1, \ldots, g_n) = (m-1)\sum_{i=1}^{n} \log(g_i) - n\log((m-1)!) + n \times \left[m\log(\lambda) - \lambda\bar{g}\right].$$

e) A random variable C follows a Poisson(λ) distribution if C counts the number of sequential events that are observed in a fixed window of time if the waiting times between events are independent and follow an Exponential(λ) distribution. This is often used as a basic model for the number of customers that get through a queue in an hour. The random variable C can take any non-negative integer value. The mass function for a Poisson(λ) random variable is

$$f(c; \lambda) = \frac{\lambda^c}{c!} \exp(-\lambda).$$

The parameter λ must be positive. Let $C_1, \ldots, C_n \overset{iid}{\sim} \text{Poisson}(\lambda)$ with realizations $c_1, \ldots, c_n$. Show that the log-likelihood for λ is

$$\ell(\lambda; c_1, \ldots, c_n) = -\sum_{i=1}^{n} \log(c_i!) + n \times [\log(\lambda)\bar{c} - \lambda] .$$