Sample final questions

Wenda Zhou

June 23, 2017

The final will have 6 questions. All questions will be marked and your grade will be computed from the 5 best questions. You will have access to an A4 (letter) sized note sheet, front and back. Each question is worth 20 marks. You may use a calculator.

1. (20 points) Estimation

   The negative binomial distribution with parameters r and p has p.m.f. given by:

   $$\mathrm{P}(X = k) = \binom{k + r - 1}{k}(1 - p)^r p^k \tag{1}$$

   it has mean $pr/(1-p)$ and variance $pr/(1-p)^2$.

   The Beta distribution with parameters α, β has p.d.f. given by (for $0 < x < 1$):

   $$f_X(x) = C(\alpha, \beta)x^{\alpha-1}(1 - x)^{\beta-1} \tag{2}$$

   It has mean $\alpha/(\alpha + \beta)$.

   (a) Suppose that $x_1, \dots, x_n$ are i.i.d. random variables following a negative binomial distribution with parameter p (unknown) and r (known).

   Compute the m.l.e. of p.

   (b) Suppose that both p and r are unknown. What is the method of moments estimator for p and r?

   (c) Suppose instead that we wish to obtain a Bayesian estimator of p. Let p follow a prior $\mathrm{Beta}(\alpha, \beta)$ distribution. Compute the posterior distribution of p and the Bayes estimator.

2. (20 points) Classification

   Suppose that we have collected the following data, and wish to classify the plotted points.

   (a) We first fit a logistic regression to predict the class of the observations. The fit is included below.

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(), data = data.mix)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.28489  -0.86579   0.05965   0.90614   1.88232

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9780     0.2945  -3.321 0.000897 ***
x1           -0.1344     0.1372  -0.980 0.327272
x2            1.3981     0.2316   6.035 1.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

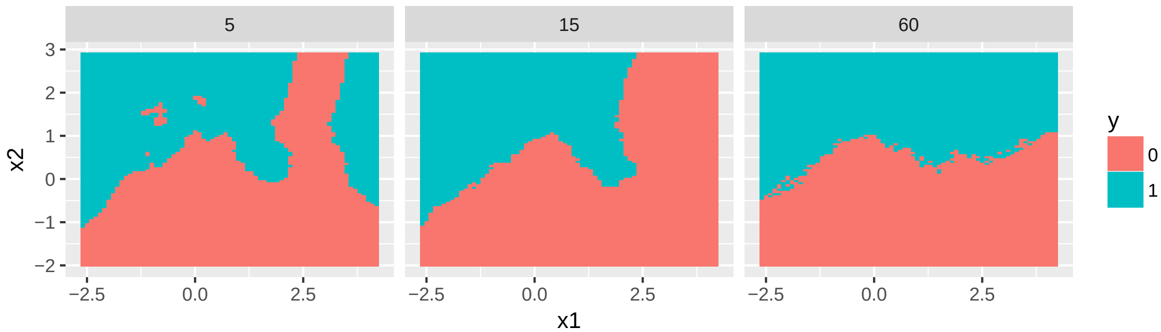
H

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 209.54  on 197  degrees of freedom
AIC: 215.54

Number of Fisher Scoring iterations: 4
```

Suppose that we wish to classify the points using this glm. We classify a point to be 1 whenever the predicted probability is greater than 0.5. Write the condition in x_1 and x_2 corresponding to this classification rule.

(b) On the graph, plot the region that we would classify as of class 0 and that we would classif as of class 1.

(c) We have also fit three k-nn classifier, with 5, 15 and 60 neighbours. Their classification regions are plotted below. Which one do you think is best? Justify your choice.



3. (20 points) Poisson regression

In a particular species of horseshoe crabs, female crabs, in addition to their main partner, may also additional main partners called satellites. We wish to understand how that relates to some characteristics of the female crab: the width in centimeter, whether the crab is dark coloured (yes/no), and whether the crab has a good spine (yes/no).

The output of the regression is included below:

```
Call:
glm(formula = Satellites ~ Width + Dark + GoodSpine, family = poisson(),
    data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9343  -1.9988  -0.4123   1.0239   4.6961

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.820088   0.570859  -4.940 7.81e-07 ***
Width         0.149196   0.020753   7.189 6.52e-13 ***
Darkyes      -0.265665   0.104972  -2.531   0.0114 *
GoodSpineyes -0.002041   0.097990  -0.021   0.9834
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.96  on 169  degrees of freedom
AIC: 924.25

Number of Fisher Scoring iterations: 6
```

(a) For each of the variables, describe whether and what effect they have on the average number of satellites for a given female crab.

(b) What is the average number of satellites for a female crab with a width of 26.3, with dark colouring and a good spine?

(c) For the same crab as above, what is the probability of it having at least one satellite?

(d) Compute a 50% confidence interval for the change in log number of satellites per centimeter of width. Deduce a confidence interval for the multiplicative change in average number of satellites per centimeter of width.

You may be interested in the following normal quantiles:

$$z_{0.975} = 1.96$$
$$z_{0.75} = 0.67$$
$$z_{0.6} = 0.25$$

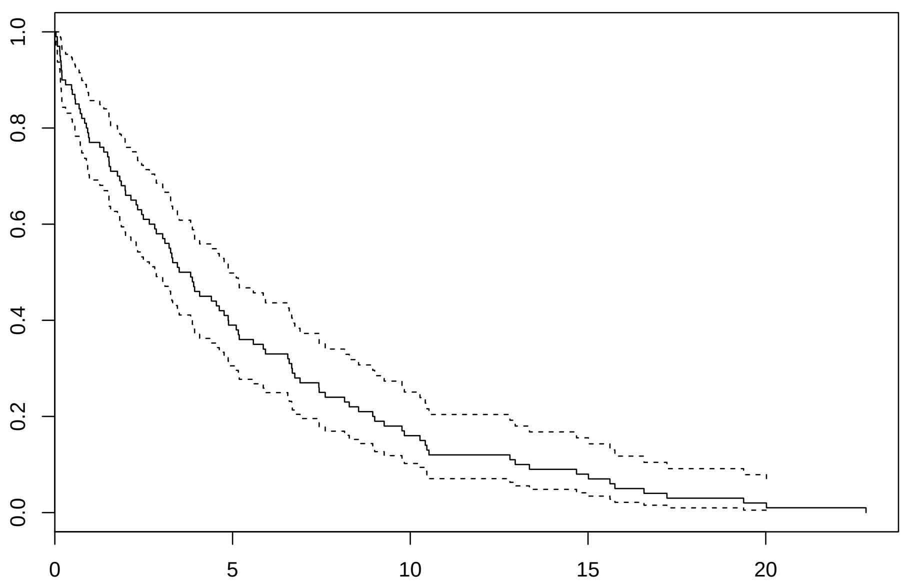
4. (20 points) Survival analysis

In this question, we consider a parametric model for survival.

Let T denote the time to event, and suppose that T follows an exponential distribution with rate λ.

(a) Compute the survival function and the hazard function for this model.

(b) Suppose that we collect some data and fit the Kaplan-Meier estimate. The fit is plotted below. What is (approximately) the probability of $T > 5$ according to the estimate? Can you use this to estimate the average survival time?



(c) Suppose that we have two independent observation, one with $T_1 = t_1$, and the second one that is censored so that we only know $T_2 \geq t_2$.

Write down the joint likelihood (supposing that T_1, T_2 are independent and follow an exponential distribution with rate λ), and find the m.l.e. estimator of λ.