Assignment 10, ST2304

Problem 1 Several hypotheses have been proposed for how blood types in humans are determined. We now know that the observable phenotypes A, B, AB and 0 corresponds to a number of underlying genotypes at a single triallelic locus listed in Table 1 (under hypotheis 1). In handout 2 we saw how the this hypothesis could be tested against data assuming that the population is in Hardy-Weinberg equilibrium by maximising the likelihood function for the observed counts (the rigthmost column) numerically using the methods in handout 5.

Until 1924, a competing hypothesis which we shall consider here (see Crow, 1993, for an historical account of this) was that blood types were determined by two diallelic loci, with a dominant allele A at the first locus determining the A antigen and a dominant allele B at the second locus determining the B antigen (hypothesis 2 in Table 1). The probabilities of observations in the four phenotypic categories are then functions of two unknown parameters, the allele frequency at each of the two loci. We choose to work with the allele frequencies p_a and p_b of the recessive alleles.

The proability that both loci are homozygeous for the recessive alleles a and b is then $p_a^2 p_b^2$. This assumes random association between loci, so called linkage equilibrium or gametic phase equilibrium and not only Hardy-Weinberg equilibrium at each respective loci. The probability that at least one copy of the dominant allele A is present at the first locus (the genotype A-, that is, either AA or Aa) is $(1-p_a)^2 + 2p_a(1-p_a) = 1 - p_a^2$. Thus, the probability of the genotypes A-bb is $(1-p_a^2)p_b^2$.

		Hypothesis 1		Hypothesis 2		Observed
i	Phenotype	Genotype	Probability p_i	Genotype	Probability p_i	counts X_i
1	A	AA, A0	$p_A^2 + 2p_A p_O$	A- bb	$(1-p_a^2)p_b^2$	44
2	B	BB, B0	$p_B^2 + 2p_B p_O$	aa B-	$p_a^2(1-p_b^2)$	27
3	AB	AB	$2p_A p_B$	A- B-	$(1-p_a^2)(1-p_b^2)$	4
4	0	00	p_O^2	aabb	$p_a^2 p_b^2$	88

Table 1: Underlying hypothetical genotypes and genotype frequncies corresponding to observable blood types in humans and observed counts from an African population.

1. First fit the model represented by hypothesis 1 to the observed counts in Table 1 using the code listed in handout 2. Also compute approximate standard errors of the estimates of the allele frequencies using the method given in section 2.1 in handout 5.

2. Next modify the function `lnL` and `multinomialprobs` so that the model corresponding to hypothesis 2 can be fitted to the data. What are the MLEs of the allele frequencies p_a and p_b? Also compute the standard errors of these estimates.

   Hint: You may need to supply "box-contraints" on the parameters when using `optim`, for example, `lower=c(.001,.001),upper=c(.999,.999)`, otherwise `optim` may try to evaluate the log likelihood outside the permitted parameter values which may cause errors. Alternatively, try specifing reasonable starting values for the parameters close to the maximum likelihood estimates.

3. What are the expected number of observations of each of the four observable bloodtypes based on this alternative model?

4. Compute the chi-square statistic for the goodness-of-fit test of the model. Can you reject this model based on the obseved counts? How does this compare with goodness-of-fit test for hypothesis 1 in handout 2?

5. Would you be able to assess the goodness-of-fit of a model involving three instead of two unknown parameters?

Problem 2 In this exercise we shall reanalyse the data used in problem 3 in assignment 7. First load the data into R using the commnad

```
moose <- read.table("https://www.math.ntnu.no/~jarlet/statmod/ovul2.dat")
```

1. Instead of fitting the model used in assignment 7 using the `glm` function, first write a likelihood function for model
$$\text{probit}\, p = \beta_0 + \beta_1 x \tag{1}$$
and maximise the log likelihood numerically. Verify that you obtain the same estimates of β_0 and β_1.

   Hints: You may need to study the example listed in section 3 in handout 5 but note that this model involves a probit and not a logit link so p is a different function of the linear predictor. Which function? Also you'll need to take into account the number of bernoulli trials on which each observation is based when computing the log likelihood.

2. Equation (1) implies that
$$p = \phi(\beta_0 + \beta_1 x) \tag{2}$$
where ϕ is the cumulative standard normal density. In assignment 7 we saw that this model fitted the data poorly. A possible explanation for this may be that not all the individuals but only a proportion q ovulates during the rut each year. If we treat q as an unknown parameter and build this assumption into the model the relationship between ovulation probability p and time x becomes
$$p = q\phi(\beta_0 + \beta_1 x) \tag{3}$$
Modify the code you wrote in the previous point so that the likelihood function instead involves the three unknown parameters q, β_0, β_1 and the above relationship between p and x and refit the model.

   Hint: The MLEs for the simple model not involving q provide reasonable starting values for β_0 and β_1.

3. Compute the standard errors of q, β_0 and β_1.

4. Make a plot of the data, and curves representing the alternative models.

5. Are the models considered above nested? What is the change in two times the log likelihood? Can you based on this reject (1) in favour of (3)? (See section 2.2. in handout 5).