Norwegian University
of Science and Technology
Department of Mathematical Sciences

ST2304 Statistical modelling for biologists and biotechnologists - solution

4. juni, 2012

Problem 1

a) Since X is discrete, the event $X > 8 \cap X < 10$ is equivalent with the event $X = 9$. The probability that X takes this single specific value can be computed using the expression

```
dbinom(9,size=50,prob=.2)
```

b) By default, `pbinom` computes probabilities of the form $P(X \leq q)$. If using the optional argument `lower.tail=FALSE`, R computes the complement of this, that is, $P(X > q)$. We thus need to transform the problem into this form. Since X is discrete $P(X \geq 11) = P(X > 10)$. This can be computed in R using the expression

```
pbinom(10,size=50,prob=.2,lower.tail=F)
```

c) The median is defined as the 50%-quantile of the distribution. All three quantiles can be computed using a single expression using the fact that `qbinom` operates elementwise on vector arguments.

```
qbinom(c(.05,.5,.95),prob=.2,size=50)
```

d)
```
hist(rbinom(1000,prob=.2,size=50))
```

or perhaps

```
hist(rbinom(1000,prob=.2,size=50),breaks=0:21-.5)
```

if we want the bins of the histograms centered around each discrete value X might take.

Problem 2

a) Linear regression is based on the assumption that the response variable y has a normal distribution with a constant variance σ^2. This assumptions is not satisfied since y is discrete and better modelled using a Poisson distribution (see point b) for which the variance equals the mean. Extrapolating a linear regression would furthermore lead to

negative predictions for some values of x which is clearly not realistic since y is necessarily non-negative. A partial remedy could be to work with $\log y$ as the response but this would be problematic since the log of y-values equal to zero are not defined.

b) The model assumes that

$$Y_i \sim \text{pois}(\lambda_i) \tag{1}$$

and that

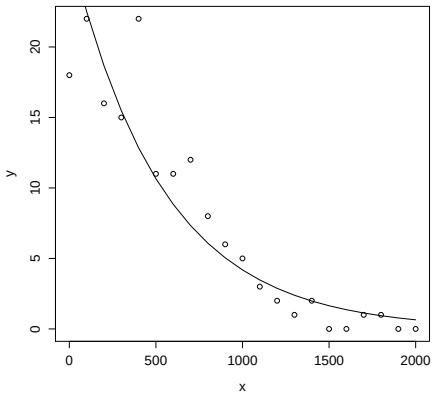
$$\ln \lambda_i = \beta_0 + \beta_1 x_i \tag{2}$$

for each observation $i = 1, 2, \ldots, n$.

If the number of plants inside different disjoint areas within each sampling site are independent, and indvidual plants occur with an approximately constant rate inside each sampling site, then the individual plants occur according to a spatial homogeneous Poissonprocess. The total number of plants within each sampling site is then Poisson distributed with parameter equal to the rate of the Poissonprocess multiplied by the area of each sampling site. The choice of log link-function ensures that the model always (for all values of β_0 and β_1) predicts a positive expected number of plants inside each sampling site, since (2) implies that

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i). \tag{3}$$

c)



d) Under the null hypothesis of no overdispersion, the deviance is chi-square with $n - p = 19$ degrees of freedom. A high deviance would support the alternative hypothesis of overdispersion. We thus reject H_0 if the deviance is larger than the upper 5%-quantile of the chi-square distribution which equals 30.14. Since the observed deviance is smaller than the critical value we cannot reject H_0.

Possible mechanisms which could generate overdispersion would be heterogeneity between the different sampling sites (for example due to local variation in concentration of nutrients), non-independence in the underlying spatial Poissonprocess (clusters of individuals in space), and an incorrect functional relationship between λ and x.

e)
```
lnL <- function(par,x,y) {
  x0 <- par[1]
  c <- par[2]
  K <- par[3]
  lambda <- K/(1+exp(c*(x-x0)))
  -sum(dpois(y,rate=lambda,log=T))
}
```

f) The glm (H_0) involving $p_0 = 2$ parameters (the slope β_0 and the intercept β_1) can be tested against the more general sigmoid model (H_1) involving $p_1 = 3$ parameters (x_0, c and K) using the test statistic

$$2(\ln L_1 - \ln L_0) \tag{4}$$

which is approximately chi-square with $p_1 - p_0 = 3 - 2 = 1$ degrees of freedom. The critical value becomes the upper 0.05 quantile of the chi-square distribtuion, that is, 3.84. Since the observed value of (4),

$$2(-36.32 - (-42.67)) = 12.7 \tag{5}$$

exceeds the critical value we reject H_0 in favour of H_1.

Relying on assymptotic theory, the estimate of carrying capacity K (the third element of the parameter vector) has variance 10.96 and a standard deviation 3.31.

We are not asked to verify that the models are nested so the following is outside the scope of the stated problem. The simpler model is equivalent to the more general sigmoid model in a special limiting case obtained by letting $K \to \infty$ and $x_0 \to -\infty$ jointly in the following manner. We let K be a function of x_0, namely, the function $K(x_0) = K^* e^{-cx_0}$. Then, when we send x_0 to minus infinity, λ goes to the limiting value

$$\lim_{x_0 \to -\infty} \lambda(x; x_0, K(x_0), c) = \lim_{x_0 \to -\infty} \frac{K(x_0)}{1 + e^{c(x-x_0)}} = \lim_{x_0 \to -\infty} \frac{K^* e^{-cx_0}}{1 + e^{c(x-x_0)}} = K^* e^{-cx} \tag{6}$$

for all x. This relationship between λ and x is equivalent to model (3), the only difference being the parameterization used.