MAST30027: Modern Applied Statistics

Week 6 Lab

1. In the question 2 of the week 5 lab, we fitted a rate model (a type of Poisson regression) to data on the effect of gamma radiation on chromosomal abnormalities.

   Show that these data are overdispersed compared to a Poisson distribution. Next test for an interaction between `doserate` and `doseamt`, firstly without allowing for overdispersion (fixing this dispersion $\phi = 1$), and secondly allowing for overdispersion. Do you get different answers?

   Solution: The fitted model from the the question 2:

```
> library(faraway)
> data(dicentric)
> model <- glm(ca ~ offset(log(cells)) + doserate*doseamt, family=poisson, data=dicentric)
> summary(model)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate * doseamt, family = poisson,
    data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7308  -2.2842  -0.6264   3.3487   5.8272

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.29994    0.06160 -53.567  < 2e-16 ***
doserate          0.06401    0.02922   2.191 0.028476 *
doseamt           0.61224    0.01707  35.862  < 2e-16 ***
doserate:doseamt  0.02715    0.00765   3.549 0.000387 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:  270.26  on 23  degrees of freedom
AIC: 453.67

Number of Fisher Scoring iterations: 4
```

   We checked that the deviance of the fitted model was very high last week. To show that these data are overdispersed compared to a Poisson distribution, we estimate the dispersion parameter and get something much larger than 1.

```
> (phi <- sum(residuals(model, type="pearson")^2)/23)

[1] 12.97226
```

   From the summary of the fitted model above (without allowing for overdispersion), we can see that the interaction between `doserate` and `doseamt` appears significant.

   We then fit a quasi-Poisson regression and perform testing for the interaction term using an F test. We see that the interaction no longer appears significant.

```
> model.f <- glm(ca ~ offset(log(cells)) + doserate*doseamt, family=quasipoisson, data=dicentric)
> model.r <- glm(ca ~ offset(log(cells)) + doserate + doseamt, family=quasipoisson, data=dicentric)
> anova(model.f, model.r, test="F")
```

```
Analysis of Deviance Table

Model 1: ca ~ offset(log(cells)) + doserate * doseamt
Model 2: ca ~ offset(log(cells)) + doserate + doseamt
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        23     270.26
2        24     282.95 -1  -12.688 0.9781  0.333
```

Results from testing for the interaction between `doserate` and `doseamt` with/without allowing for overdispersion are different.

2. 1) Simulate 1000 samples from the following mixture model. For $i = 1 \ldots, 1000$,

$$Z_i \sim \text{categorical } (0.3, 0.7),$$

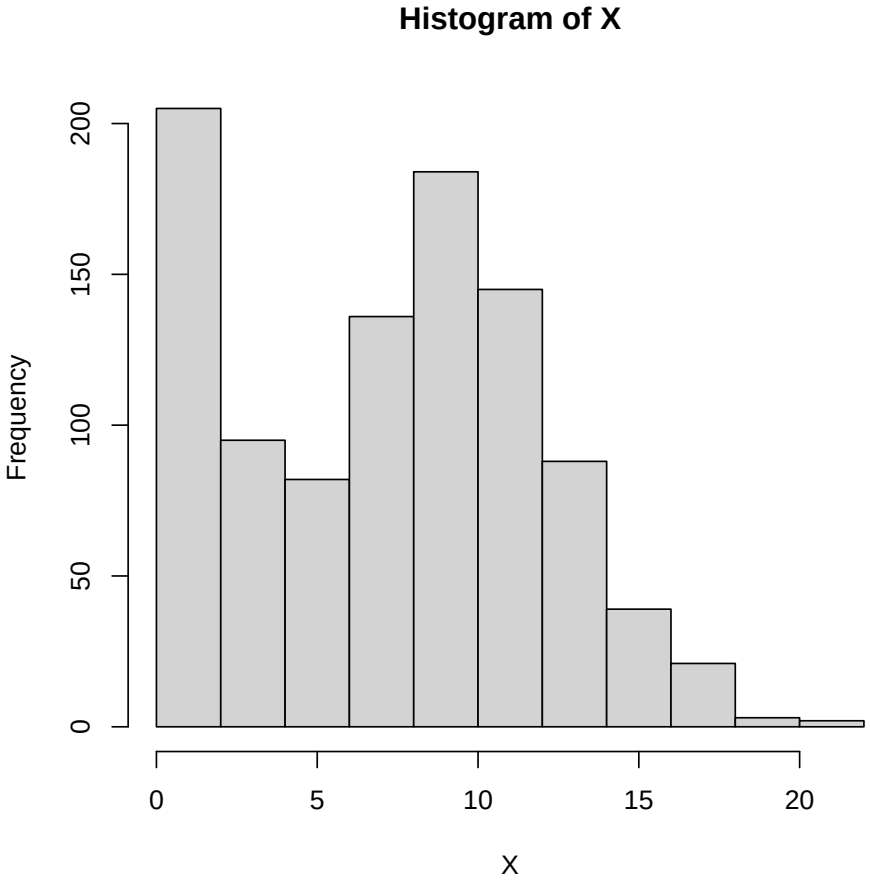
$$X_i|Z_i = 1 \sim \text{Pois}(\lambda = 2) \text{ and } X_i|Z_i = 2 \sim \text{Pois}(\lambda = 10).$$

Please set a seed using 'set.seed(30027)'. Save the information about which component each observation come from. Make a histogram of the simulated samples.

2) We pretend that we just obtain the simulated samples without their true component information. Cluster the simulated samples into two clusters using the mixture of two Poisson models. We assume we have already obtained MLE of the parameters in the model. We will use true values of parameters as MLE here ($\hat{\pi}_1 = 0.3, \hat{\lambda} = 2$ for the first component, and $\hat{\lambda} = 10$ for the second component). If $P(Z_i = 1|X_i) > 0.5$, assign a sample X_i to the first cluster. How many samples are assigned to wrong clusters?

Solution: 1)

```
> set.seed(30027)
> lambda.1   <- 2
> lambda.2   <- 10
> pi.1 <- 0.3
> pi.2 <- 1-pi.1
> NUM.SAMPLES <- 1000
> X <- rep(NA, NUM.SAMPLES)
> Z <- rep(NA, NUM.SAMPLES)
> for(i in seq_len(NUM.SAMPLES)) {
+   Z[i] <- rbinom(1,1,pi.2)
+   if(Z[i] == 0) X[i] <- rpois(1, lambda = lambda.1)
+   else X[i] <- rpois(1, lambda = lambda.2)
+ }
> hist(X)
```



X contains the simulated data and Z contains the information about which component each observation came from.

2)

```
> L = matrix(NA, nrow=length(X), ncol= 2)
> L[, 1] = dpois(X, lambda=lambda.1)
> L[, 2] = dpois(X, lambda=lambda.2)
> L[, 1] = L[, 1]*pi.1
> L[, 2] = L[, 2]*pi.2
> CP = L/rowSums(L)
> # The number of observations that are incorrectly assigned to the second cluster.
> sum((Z==0)&(CP[,2]>0.5))

[1] 19

> # The number of observations that are incorrectly assigned to the first cluster.
> sum((Z==1)&(CP[,2]<=0.5))

[1] 21
```

40 samples (out of 1000) are assigned to the wrong clusters.