Comparison of Conditional and Unconditional Tests for the Log-Linear Model

Ana S. Tehrani, Joshua D. Habiger Ph.D

Department of Mathematics, University of Central Oklahoma

Department of Statistics, Oklahoma State University

The Big Question

A conditional-type hypothesis test in conjunction with a log-linear model was utilized to determine if bacteria is related to productivity, but what if we were to use an unconditional hypothesis test?

Abstract

In Anderson and Habiger 2011, a conditional hypothesis test, along with an FDR method, was utilized to determine which bacteria in the rihizosphere are related to productivity. However, an unconditional hypothesis test could have been utilized. This poster describes the methodology in the aforementioned paper as well as the relevant conditional and unconditional hypothesis tests in detail. A simulation study for determine which test is more powerful, i.e. will more likely reject the null hypothesis that a bacteria is not related to productivity, is presented. The study suggests that the conditional hypothesis test is more powerful than the unconditional test. R code used for the study is provided.

Introduction

-Our goal was to determine if the presence or absence of bacteria in the rhizosphere is related to the productivity of a wheat plant for each of 700 plus diferent bacteria.

-Determine the most powerful test.

Literature Methodology

A conditional hypothesis test, in conjunction with a log-linear model was used. Assuming $log(\mu_i) = \beta_0 + \beta_1 x_i$. Where $H_0 : \mu_1 = \mu_2 = ... = \mu_5$ or $H_0 : \beta_1 = 0$ is the null hypothesis tested against $H_1 : \beta_1 \neq 0$. Also assume that under the null $Y_1, Y_2, ...Y_5 | \sum_{i=1}^{5} Y_i = y$. has a multinomial distribution. In Anderson and Habiger the p-value was adjusted using

$$FDR = \frac{\text{number of type 1 errors}}{\text{number of rejected null hypothesis}}$$

.

Computing a P-value

For both methods we assume $Y_i \sim Poisson(\mu_i)$ and to compute a sufficent test statistic we use:

$$T_{obs} = |\sum_{i=1}^{5} x_i y_i - \overline{y} \sum_{i=1}^{5} x_i.|$$

The p-value can be estimated by

$$p - value = P(\widehat{T > T_{obs}}).$$

For the conditional method we compare T_{obs} to the distribution of T given Y. by sampling $Y_1^{(b)}, Y_2^{(b)}, ..., Y_5^{(b)}$ from a multinomial distribution with a probability vector (1/5,1/5,1/5,1/5,1/5) of size y.

For the unconditional method we use the marginal distribution of T estimated by shuffling Y_1,Y_2,...,Y_5 to obtain $Y_1^{(b)}, Y_2^{(b)}, ..., Y_5^{(b)}$.

Methodology

The powers were compared at a 0.05 level by simulation. The steps for simulation are as follows:

-Step 1: Estimate the power for a specified β_0 and β_1, by generating $y_1^{(k)}, y_2^{(k)}, ..., y_5^{(k)}$ from a Poisson (μ_i) distribuition, where $\mu_i = e^{\beta_o + \beta_1 x_i}$.

-Step 2: Compute the conditional and unconditional p-values denoted by $p_{cond}^{(k)}$ and $p_{unc}^{(k)}$, for k = 1, 2, ..., 5000 as previously described.

-Step 3: Compute the powers as follows.

$$Pow_{(cond)} = \frac{\text{number of } p_{cond}^{(k)} < 0.05}{10000}$$

and

$$Pow_{(unc)} = \frac{\text{number of } p_{unc}^{(k)} < 0.05}{10000}.$$

Specific values of β_1 and β_0 were chosen with the intent of ensuring that $\mu_i < 10$ and to allow for positive, negative and no relationship with x.

The Connection

For both methods the test statistic is computed as $T^{(b)} = |\sum_{i=1}^{5} x_i {y_i}^{(b)} - \overline{y}^{(b)} \sum_{i=1}^{5} x_i|$. This process is repeated 10000 times to obtain $T^{(1)}, T^{(2)}, ..., T^{(10000)}$. The p-value is estimated via $p - value = \frac{\#{T^{(b)}}' s > T_{obs}}{10000}$.

P-value Illustration

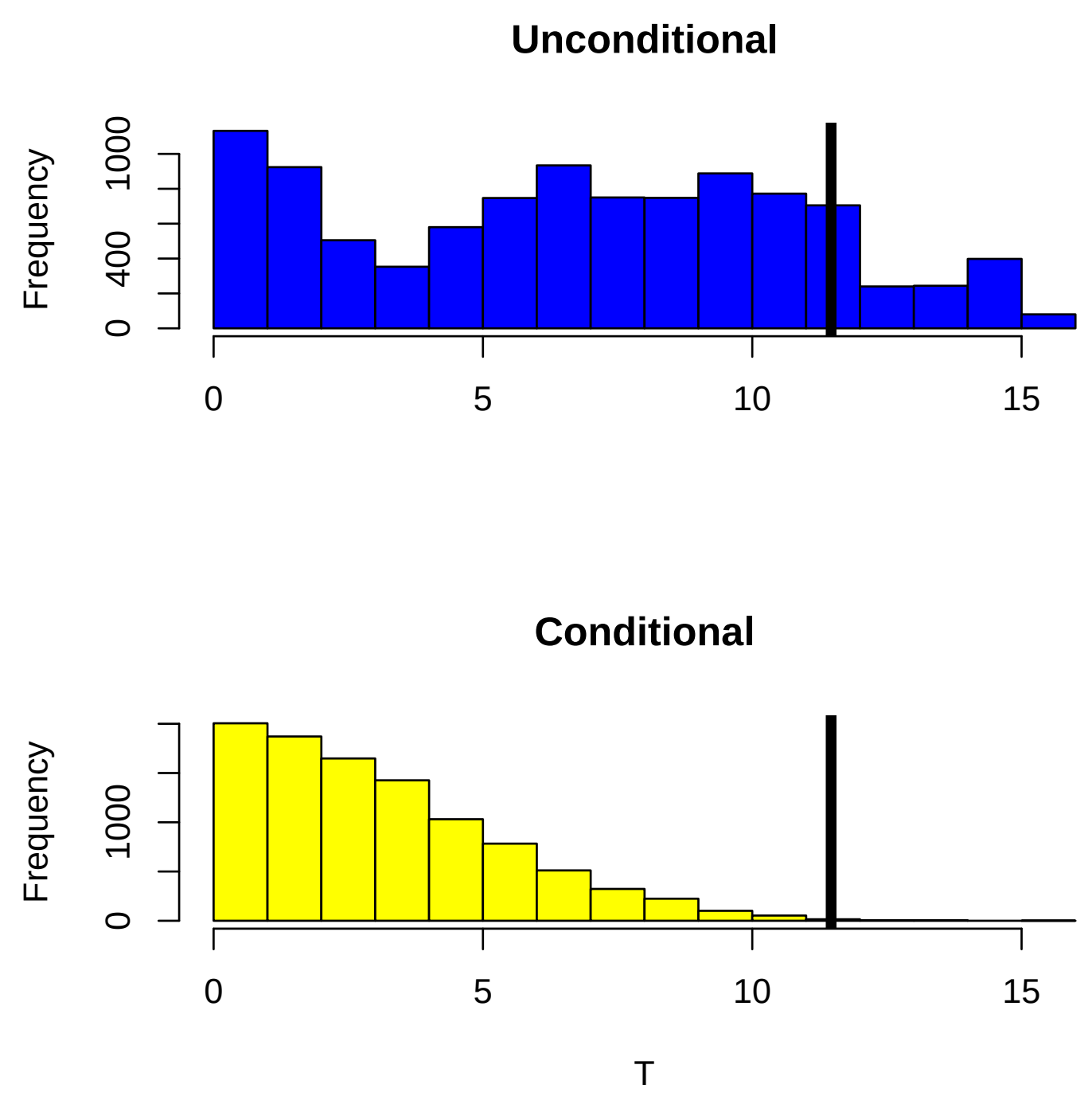

Figure 1: The unconditional (top) and conditional (bottom) distributions of T for 1000 replications for bacteria 9 are above.

Results

β_1	β_0	U. Pow	C. Pow	μ_1	μ_2	μ_3	μ_4	μ_5
1	-2	0.180	0.415	0.323	0.517	0.827	1.448	2.718
2	-4	0.543	0.971	0.104	0.267	0.684	2.096	7.389
-1	2	0.109	0.385	3.096	1.935	1.210	0.691	0.369
-2	4	0.311	0.989	9.583	3.743	1.462	0.477	0.135
0	1	0.000	0.003	2.718	2.718	2.718	2.718	2.718
0	2	0.028	0.039	4.750	4.750	4.750	4.750	4.750

Table 1: Power of conditional and unconditional tests.

For example, when β_1=2 and β_0=-4, the conditional power is 0.97 while the unconditional power is 0.530 making the conditional test a more powerful test. Thus, simulation studies suggest the conditional test is more powerful.

R-Code

```
###Imput the appropriate values of x and y as below:###

x = c(.87,1.34,1.81,2.37,3)

y = c(1,5,6,2,13)

### exact unconditional pval  ####

get.pval1<-function(y){
T1<-rep(0,10000)
T = abs(sum(x*y) - mean(y)*sum(x))
for(i in 1:10000)
  {
  y0<-sample(y)
  T1[i]<-abs(sum(x*y0) - mean(y0)*sum(x))
  }
pval = mean(T1>=T)
return(list(pval,T1))}


#### exact conditional pval###

get.pval2<-function(y){
T1<-rep(0,10000)
T = abs(sum(x*y) - mean(y)*sum(x))
for(i in 1:10000)
  {
  y0<-rmultinom(1,sum(y),c(.2,.2,.2,.2,.2))
  T1[i]<-abs(sum(x*y0) - mean(y0)*sum(x))
  }
pval = mean(T1>=T)
return(list(pval,T1))}

#Example
get.pval1(y=c(1,5,6,2,13)
get.pval2(y=c(1,5,6,2,13)


#### simulation of power ########

simulate<-function(reps = 1000, b1 = 1,b0=-1){
pvalcond<-rep(0, reps)
pvaluncond<-rep(0,reps)
mu = exp(b0+b1*x)
for(i in 1:reps)
  {y = rpois(5, mu)
   pvaluncond[i] = get.pval1(y)[[1]]
   pvalcond[i] = get.pval2(y)[[1]]
   print(i)
  }
power.uncd = mean(pvaluncond<=.05)
power.cond = mean(pvalcond<=.05)
return(c(power.uncd = power.uncd, power.cond=power.cond))}

#Example
b0n2b11<-simulate(reps = 5000,b0=-2,b1=1)
```

References

[1] M. Anderson and J. Habiger. Identifcation of root bacteria assoicate with shoot biomass productivity. Submitted, 2011.

[2] P. McCullagh and J.A. Nelder. Generalized Linear Models. Chapman and Hall, 1989.

[3] J.D. Storey. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64:479 498, 2002.

[4] J.D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. The Annals of Statistics, 31:2012 2035, 2003.

Acknowledgements

Special thanks to Dr. Tracy Morris from the University of Central Oklahoma for supervising this project. This project was supported by the UCO McNair and LS-OKAMP programs.