Data Analysis

Ming Yang

October 29, 2014

```
library(ZIM)
```

```
## Loading required package: MASS
```

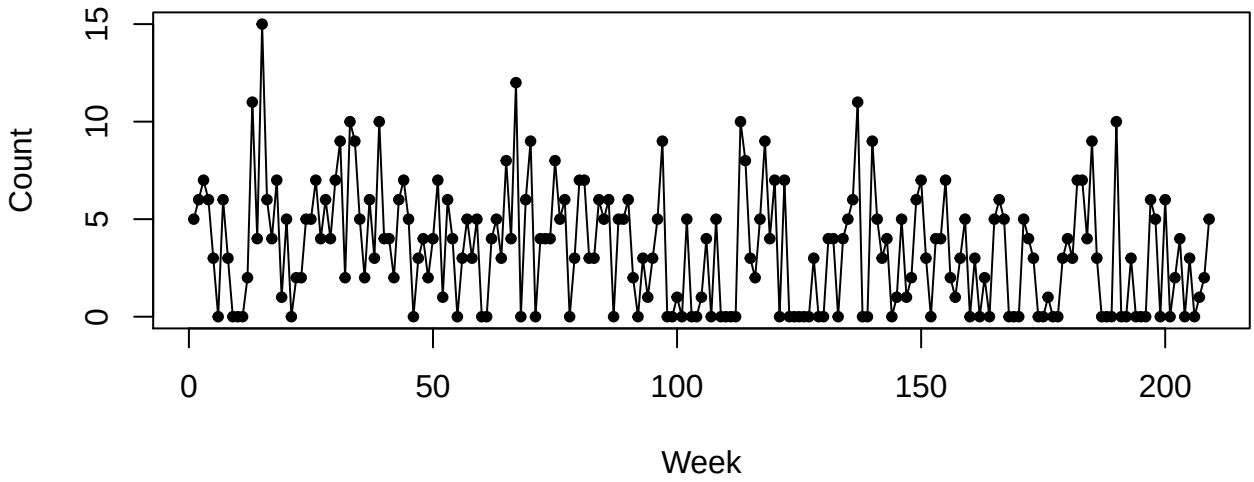
```
library(lattice)

data(syph)
count.old <- syph$a33

count.new <- c(5, 5, 0, 0, 11, 0, 0, 9, 8, 0, 3, 0,
  0, 0, 5, 1, 5, 0, 5, 5, 0, 5, 2, 0, 5, 0, 5, 6, 6, 5, 13, 0, 0, 2, 3, 2)
n1 <- length(count.old)
n2 <- length(count.new)
count <- c(count.old, count.new)
trend <- 1:(n1 + n2) / 1000

table(count.old)
```

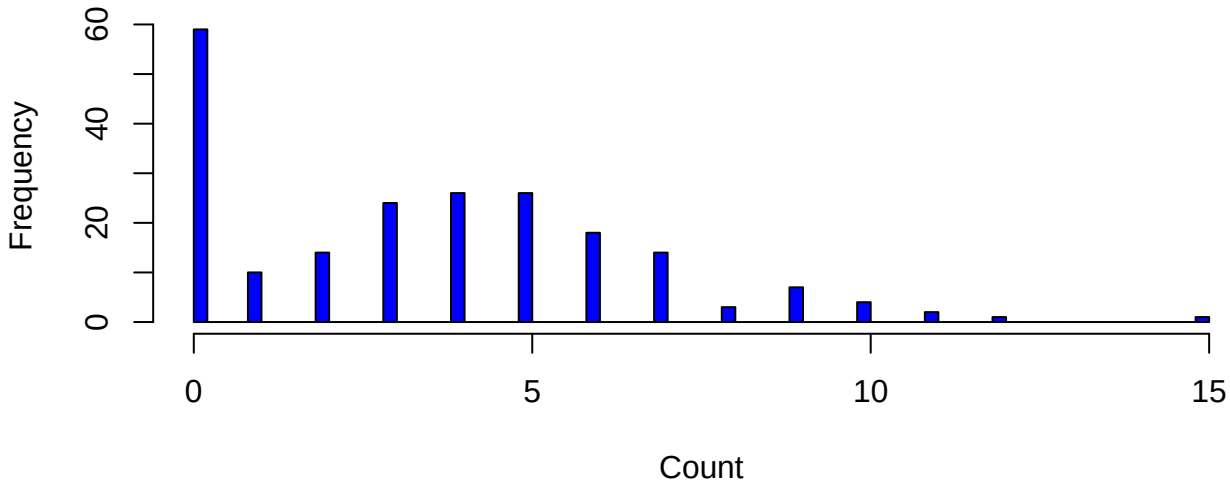
```
## count.old
##  0  1  2  3  4  5  6  7  8  9 10 11 12 15
## 59 10 14 24 26 26 18 14  3  7  4  2  1  1
```

```
table(count.new)
```

```
## count.new
##  0  1  2  3  5  6  8  9 11 13
## 14  1  3  2 10  2  1  1  1  1
```

```
par(mfrow = c(2, 1))
plot(count.old, type = "o", pch = 20, xlab = "Week", ylab = "Count")
hist(count.old, breaks = 100, xlab = "Count", col = "blue", main ="")
```






```
m <- 4
M <- cbind(1, sapply(1:m, bshift, x = count.old) > 0)

aic <- matrix(NA, 1 + m, 1 + m)
tic <- matrix(NA, 1 + m, 1 + m)
for(i in 1:(1 + m)) {
  for(j in 1:(1 + m)) {
    fit <- zim.fit(count.old[(1 + m):n1],
      cbind(M[(1 + m):n1, 1:i], trend[(1 + m):n1]),
      cbind(M[(1 + m):n1, 1:j], trend[(1 + m):n1]))
    aic[i, j] <- fit$aic
    tic[i, j] <- fit$tic
  }
}

sum(tic > aic)
```

```
## [1] 25
```

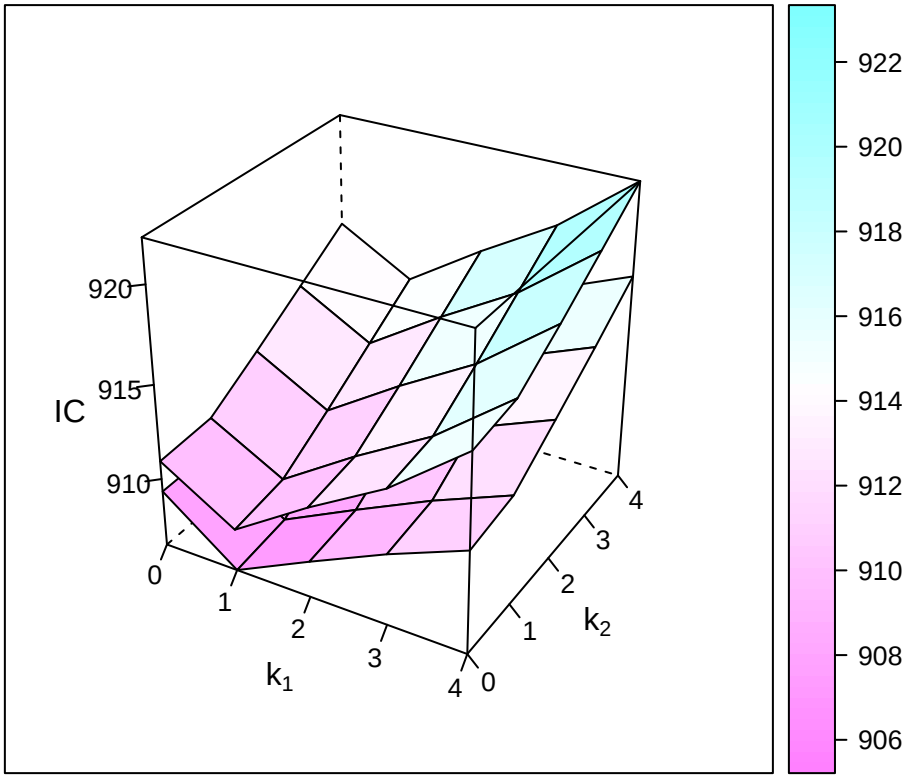
```
which(aic == min(aic), arr.ind = TRUE)
```

```
##      row col
## [1,]   2   1
```

```
which(tic == min(tic), arr.ind = TRUE)
```

```
##      row col
## [1,]   2   1
```

```
k1 <- 0:m
k2 <- 0:m
dat <- expand.grid(k1, k2, 1:2)
dat$IC <- c(aic, tic)
colnames(dat) <- c("k1", "k2", "gr", "IC")

print(wireframe(IC ~ k1 * k2, data = dat, groups = gr, alpha.regions = 1,
  scales = list(arrows = FALSE), xlab = expression(k[1]),
  ylab = expression(k[2]), drape = TRUE, screen = list(z = -30, x = -60)))
```




```
ar1 <- bshift(count, 1) > 0
syph.md <- data.frame(count, trend, ar1)
zim(count ~ ar1 + trend | trend,
  data = syph.md, subset = 2:n1)
```

```
##
## Call:
## zim(formula = count ~ ar1 + trend | trend, data = syph.md, subset = 2:n1)
##
## Coefficients (log-linear):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.489      0.120   12.42   <2e-16 ***
## ar1TRUE        0.221      0.101    2.20    0.028 *
## trend         -1.010      0.667   -1.51    0.130
##
## Coefficients (logistic):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.933      0.372   -5.20    2e-07 ***
## trend          8.605      2.808    3.06   0.0022 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Test for overdispersion (H0: ZIP vs. H1: ZINB)
## score.test: 2.603
## p.value: 0.0046
##
## Criteria for assessing goodness of fit
## loglik: -454.4
## aic: 918.8
## bic: 935.5
## tic: 920.8
##
## Number of EM-NR iterations: 11
## Maximum absolute gradient: 2.975e-14
```

```
pois.tic <- function(fit) {
  y <- fit$model[, 1]
  X <- cbind(1, fit$model[, -1])
  lambda <- fitted(fit)
  v <- y - lambda
  J <- t(X) %*% as.matrix(as.vector(v * v) * X)
  loglik <- (fit$aic - 2 * length(fit$coef)) / (-2)
  (-2) * loglik + 2 * sum(diag(J %*% summary(fit)$cov.scaled))
}

pois.fit <- glm(count ~ ar1 + trend,
  data = syph.md, subset = 2:n1, family = poisson)
summary(pois.fit)
```

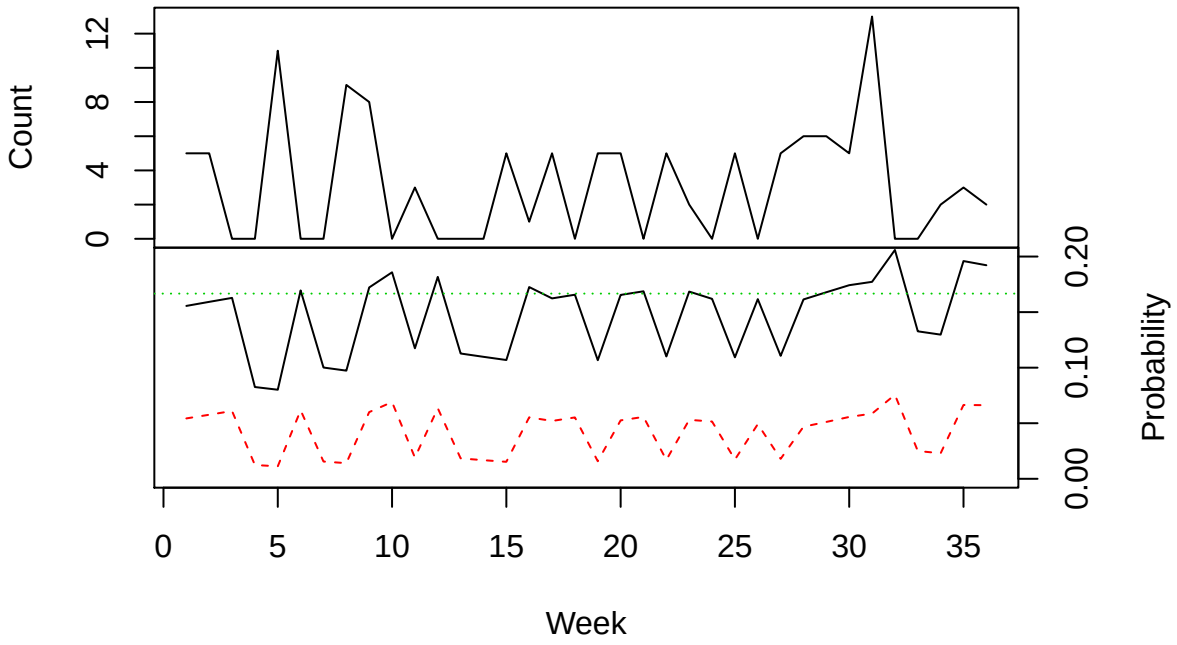
```
##
## Call:
## glm(formula = count ~ ar1 + trend, family = poisson, data = syph.md,
##     subset = 2:n1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.176  -2.061  -0.072   0.920   4.030
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2822     0.1126   11.39  < 2e-16 ***
## ar1TRUE       0.3544     0.0952    3.72    2e-04 ***
## trend        -3.1174     0.6448   -4.83  1.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 670.85  on 207  degrees of freedom
## Residual deviance: 621.24  on 205  degrees of freedom
## AIC: 1121
##
## Number of Fisher Scoring iterations: 5
```

```
pois.tic(pois.fit)
```

```
## [1] 1130
```

```
f <- function(c, ylim = c(0, 1)) {
  syph.md$p0.zip <- NA
  syph.md$gr.zip <- NA
  syph.md$p0.pois <- NA
  syph.md$gr.pois <- NA
  for(t in (n1 + 1):(n1 + n2)) {
    fit.zip <- zim(count ~ trend + ar1 | trend,
      data = syph.md, subset = 2:(t - 1))
    fit.pois <- summary(glm(count ~ trend + ar1,
      data = syph.md, subset = 2:(t - 1), family = poisson))
    zip.lambda <- exp(sum(c(1, trend[t], ar1[t]) * fit.zip$para[1:3]))
    zip.omega <- plogis(sum(c(1, trend[t]) * fit.zip$para[4:5]))
    pois.lambda <- exp(sum(c(1, trend[t], ar1[t]) * fit.pois$coef[, 1]))
    syph.md$p0.zip[t] <- dzip(0, zip.lambda, zip.omega)
    syph.md$gr.zip[t] <- pzip(c, zip.lambda, zip.omega, lower.tail = FALSE)
    syph.md$p0.pois[t] <- dpois(0, pois.lambda)
    syph.md$gr.pois[t] <- ppois(c, pois.lambda, lower.tail = FALSE)
  }
  p0.obs <- sum(count.new == 0) / n2
  gr.obs <- sum(count.new > c) / n2
  par(mfrow = c(2, 1))
  par(mar = c(0, 5, 5, 5))
  plot(ts(count.new), xaxt = "n", ylab = "Count")
  par(mar = c(5, 5, 0, 5))
  plot(ts(syph.md$gr.zip[(n1 + 1):(n1 + n2)]),
    yaxt = "n", xlab = "Week", ylab = "", ylim = ylim)
    lines(ts(syph.md$gr.pois[(n1 + 1):(n1 + n2)]), lty = 2, col = 2)
    abline(h = gr.obs, lty = 3, col = 3)
    axis(side = 4, at = seq(0, 1, 0.05))
    mtext("Probability", side = 4, line = 3)
  list(p0.obs = p0.obs, gr.obs = gr.obs)
}

f(5, c(0, 0.2))
```




```
## $p0.obs
## [1] 0.3889
##
## $gr.obs
## [1] 0.1667
```