STATS 500 HW10

Minxuan Chen

2023-12-06

Table of contents

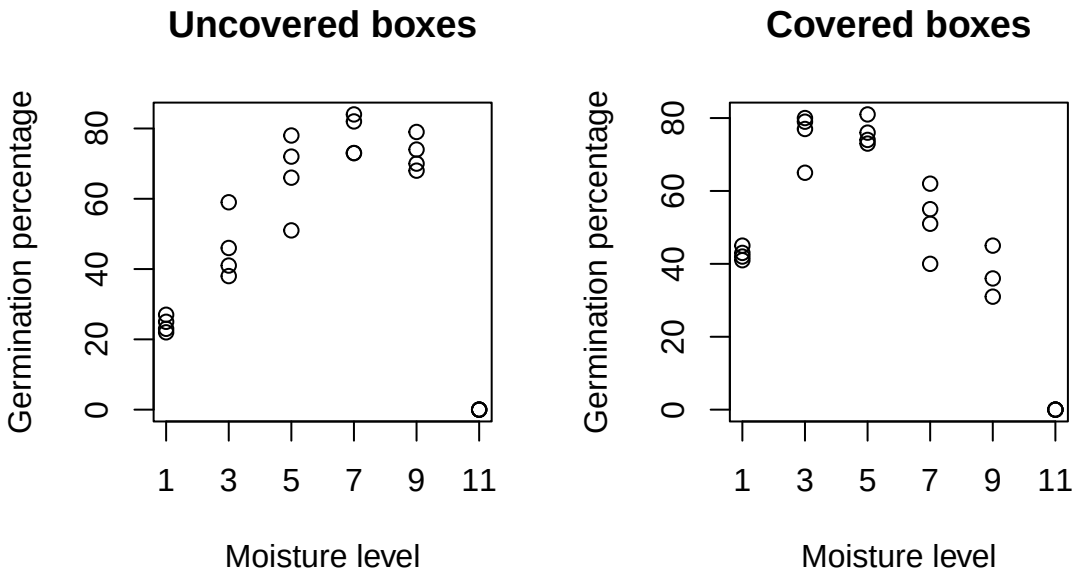
Github repo: https://github.com/PKUniiiiice/STATS_500

Problem 1

(a)

```r
library(faraway)
data(seeds)

par(mfrow=c(1,2))
plot(seeds[seeds$covered=='no', 2],
     seeds[seeds$covered=='no', 1],
     ylab='Germination percentage', xlab='Moisture level',
     main='Uncovered boxes', xaxt='n'
     )
axis(1,at=seq(1,11,by=2))
plot(seeds[seeds$covered!='no', 2],
     seeds[seeds$covered!='no', 1],
     ylab='Germination percentage', xlab='Moisture level',
     main='Covered boxes',xaxt='n')
axis(1,at=seq(1,11,by=2))
```



We find that, in both uncovered and covered boxes, the germination percentage seems to increase with moisture level when the level is relative low (<5), and the germination percentage seems to decrease with moisture level when the level is relative high (>5). And if the moisture level is in a relative high level (=7, =9). If the level is too high (=11), no seed can germinate.

So, considering the whole range of moisture level, the relationship seems in a quadratic form with a negative leading term.

(b)

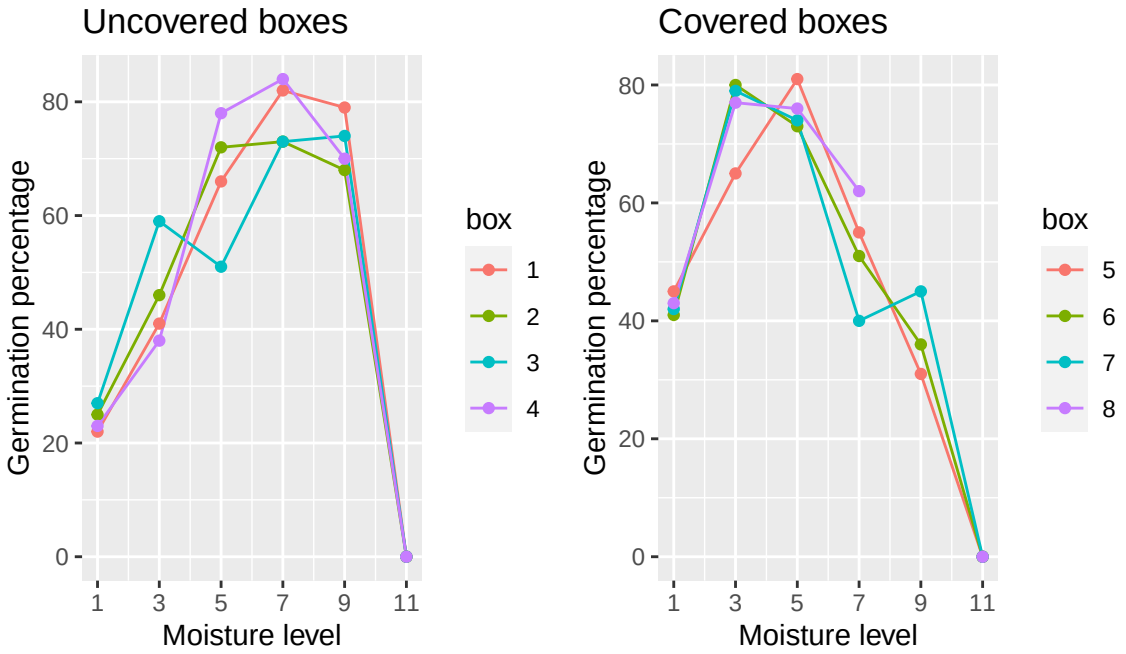
```
seeds$box <- factor(rep(1:8,each=6))

library(ggplot2)
# Create the ggplot objects
plot_uncovered <- ggplot(subset(seeds, covered == 'no'), aes(x = moisture, y = germ, col
  geom_line() +
  geom_point() +
  labs(title = 'Uncovered boxes', x = 'Moisture level', y = 'Germination percentage') +
  scale_x_continuous(breaks = seq(1, 11, 2))

plot_covered <- ggplot(subset(seeds, covered != 'no'), aes(x = moisture, y = germ, color
  geom_line() +
  geom_point() +
  labs(title = 'Covered boxes', x = 'Moisture level', y = 'Germination percentage') +
  scale_x_continuous(breaks = seq(1, 11, 2))

# Arrange the plots side by side
gridExtra::grid.arrange(plot_uncovered, plot_covered, ncol = 2)
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```



There is no indication of a box effect. Since we observe the same tendency among different boxes (in uncovered boxes or in covered boxes.)

(c)

In (a), we observe a quadratic relationship between moisture and germination percentage.

Therefore, we can add a quadratic term of moisture. That is

```r
#dropna
seeds <- na.omit(seeds)

logitm <- glm(cbind(germ, 100-germ) ~ moisture + I(moisture^2) + box + covered,
              family=binomial(link=logit), data=seeds)
summary(logitm)
```

```
Call:
glm(formula = cbind(germ, 100 - germ) ~ moisture + I(moisture^2) +
    box + covered, family = binomial(link = logit), data = seeds)

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.867630   0.135282 -13.805   <2e-16 ***
moisture       1.153897   0.043495  26.529   <2e-16 ***
I(moisture^2) -0.109683   0.003815 -28.747   <2e-16 ***
box2          -0.052195   0.131908  -0.396    0.692
box3          -0.052195   0.131908  -0.396    0.692
box4           0.026146   0.132027   0.198    0.843
box5          -0.112965   0.131856  -0.857    0.392
box6          -0.078252   0.131881  -0.593    0.553
box7          -0.086933   0.131874  -0.659    0.510
box8           0.099087   0.141349   0.701    0.483
coveredyes           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1790.99  on 46  degrees of freedom
Residual deviance:  559.22  on 37  degrees of freedom
AIC: 768.96

Number of Fisher Scoring iterations: 5
```

We find that the p-value of moisture squared is strongly significant, so it's reasonable to add this term.

Moreover, the effect of `coverage` is not identifiable since the box number and coverage are completely correlated, i.e. we can tell coverage or not totally from box number (1,2,3,4=uncovered, 5,6,7,8=covered).

To solve this problem, we may need to remove the `box` predictor from the model.

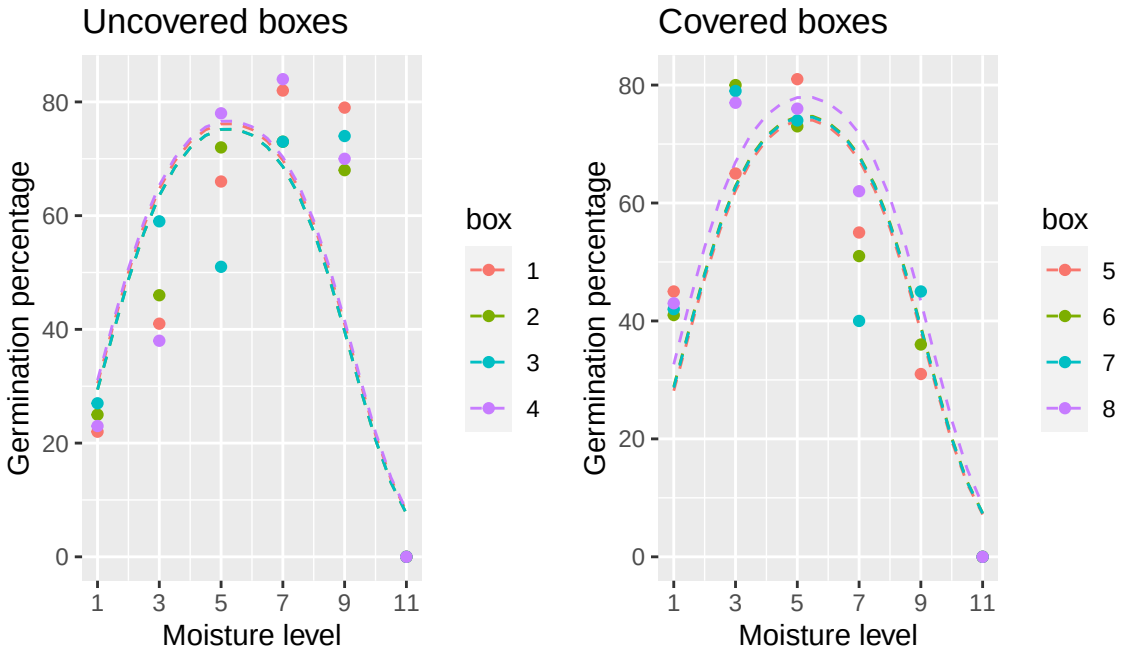
We draw a plot of the predictions.

```
xnew <- data.frame(
  moisture = rep(seq(1,11,0.5),8),
  covered = factor(rep(c('no','yes'),each=84)),
  box = factor(rep(1:8, each=21))
)
pred.logit <- predict(logitm, xnew, type='response')
temp <- cbind(xnew, pred.logit)

plot_uncovered <- ggplot(subset(temp, covered == 'no'), aes(x = moisture, color=box)) +
 # geom_point(aes(x=seeds$moisture, y = seeds$germ))+
  geom_line(aes(y = pred.logit*100,), linewidth = 0.5,
            linetype=2) +
  geom_point(data=subset(seeds, covered == 'no'),
             aes(x=moisture, y = germ))+
  labs(title = 'Uncovered boxes', x = 'Moisture level', y = 'Germination percentage') +
  scale_x_continuous(breaks = seq(1, 11, 2))

plot_covered <- ggplot(subset(temp, covered != 'no'), aes(x = moisture, color=box)) +
 # geom_point(aes(x=seeds$moisture, y = seeds$germ))+
  geom_line(aes(y = pred.logit*100,), linewidth = 0.5,
            linetype=2) +
  geom_point(data=subset(seeds, covered != 'no'),
             aes(x=moisture, y = germ))+
  labs(title = 'Covered boxes', x = 'Moisture level', y = 'Germination percentage') +
  scale_x_continuous(breaks = seq(1, 11, 2))

# Arrange the plots side by side
gridExtra::grid.arrange(plot_uncovered, plot_covered, ncol = 2)
```



(d)

To test the significance of a box effect in the model, we remove the box predictor and use anova.

```
logitm2 <- glm(cbind(germ, 100-germ) ~ moisture + I(moisture^2) + covered,
               family=binomial(link=logit), data=seeds)
anova(logitm2, logitm)
```

```
Analysis of Deviance Table

Model 1: cbind(germ, 100 - germ) ~ moisture + I(moisture^2) + covered
Model 2: cbind(germ, 100 - germ) ~ moisture + I(moisture^2) + box + covered
  Resid. Df Resid. Dev Df Deviance
1        43     562.43
2        37     559.22  6   3.2148
```

The deviance is 3.2148, by $D \approx \chi^2_{n-s}$, the p-value is

```
pchisq(3.2148, df=6, lower.tail=F)
```

```
[1] 0.7814444
```

The p-value is larger than 0.05, so we conclude that we do not reject the null model. That is, we accept the model without box effect.

If using Pearson's Chi-squared test, we need Pearson residuals.

```r
resM2 <- residuals(logitm2, type='pearson')
resM1 <- residuals(logitm, type='pearson')
pchisq(sum(resM2^2)-sum(resM1^2), df=6, lower.tail=F)
```

```
[1] 0.621187
```

The p-value is still larger than 0.05, so we accept the model without box effect.

(e)

```r
#get prediction, full model is logitm
seeds$pred <- predict(logitm, type='response')
```

In uncovered boxes

```r
seeds.no <- subset(seeds, covered=='no')
seeds.no[seeds.no$pred==max(seeds.no$pred),]
```

```
   germ moisture covered box      pred
21   78        5      no   4 0.7660103
```

The predicted maximum germination for uncovered boxes occurs at moisture level=5.

In covered boxes

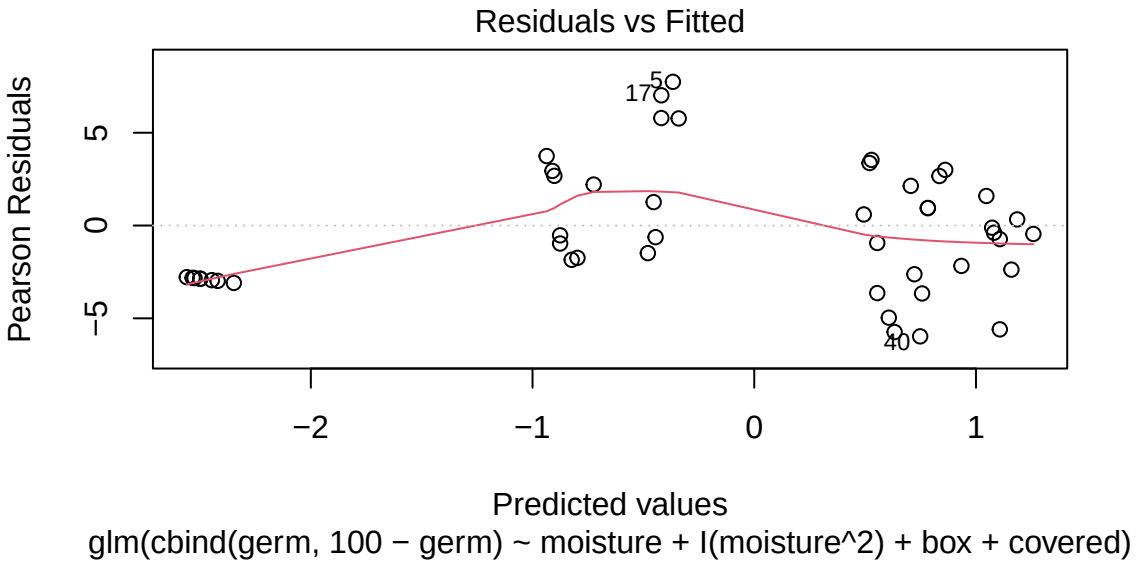
```r
seeds.yes <- subset(seeds, covered=='yes')
seeds.yes[seeds.yes$pred==max(seeds.yes$pred),]
```

```
   germ moisture covered box      pred
45   76        5     yes   8 0.7788297
```

The predicted maximum germination for uncovered boxes also occurs at moisture level=5.

(f)

```r
plot(logitm, 1)
```



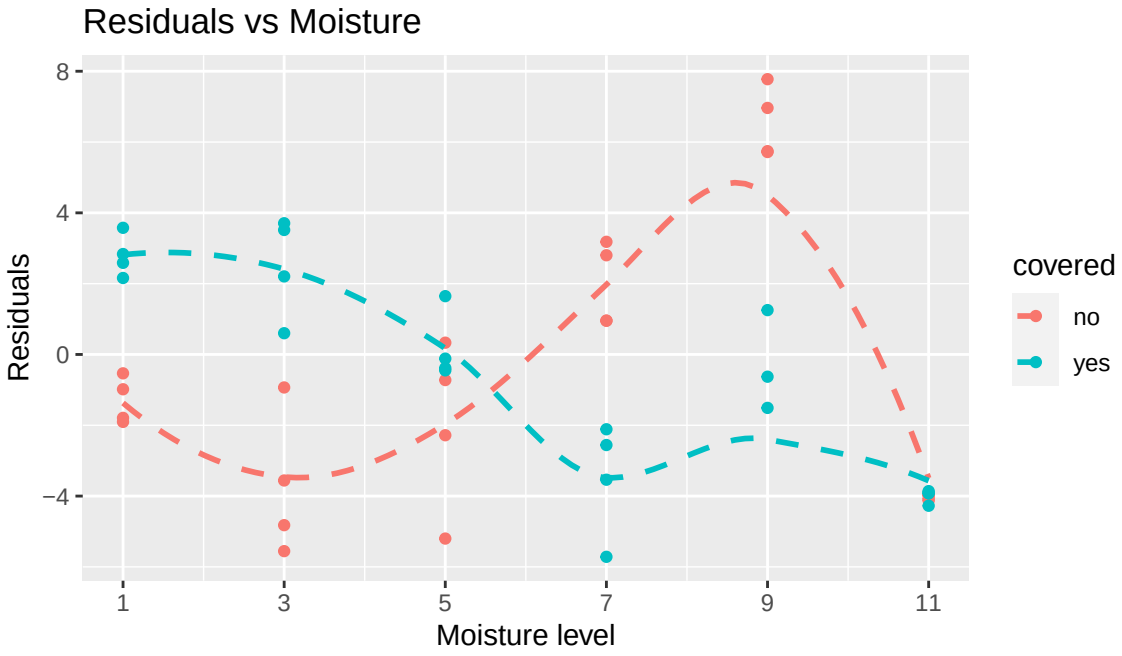
It seems the distribution of residuals is uneven across the predicted values. In other words, there may be non-constant variance in errors. And there is also slightly non-linear in the residuals since the red line is not horizontal.

(g)

```r
seeds$resid <- resid(logitm)

ggplot(seeds, aes(x = moisture, y=resid, color=covered)) +
  geom_point()+
  geom_smooth(formula=y~x, method="loess", se=FALSE, linetype = 2, lindwidth=0.5)+
  labs(title = 'Residuals vs Moisture',
       x = 'Moisture level',
       y = 'Residuals') +
  scale_x_continuous(breaks = seq(1, 11, 2))
```

```
Warning in geom_smooth(formula = y ~ x, method = "loess", se = FALSE, linetype
= 2, : Ignoring unknown parameters: `lindwidth`
```



From this plot, we can still observe non-linear pattern in the residuals. Therefore, the moisture squared tern still does not fit the data well. We may add some higher terms as a trial.

Problem 2

We estimate the dispersion parameter by

$$\hat{\sigma}^2 = \frac{X^2}{n-p}$$

that is

```
sigma2 <- sum(resid(logitm, type='pearson')^2)/logitm$df.residual
sigma2
```

```
[1] 13.76635
```

```
sumary(logitm,dispersion=sigma2)
```

```
               Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -1.867630   0.501938 -3.7208 0.0001986
moisture       1.153897   0.161380  7.1502 8.665e-13
I(moisture^2) -0.109683   0.014156 -7.7480 9.337e-15
box2          -0.052195   0.489417 -0.1066 0.9150691
box3          -0.052195   0.489417 -0.1066 0.9150691
box4           0.026146   0.489859  0.0534 0.9574330
```

```
box5          -0.112965   0.489225 -0.2309 0.8173884
box6          -0.078252   0.489318 -0.1599 0.8729432
box7          -0.086933   0.489291 -0.1777 0.8589804
box8           0.099087   0.524447  0.1889 0.8501422

Dispersion parameter = 13.76635
n = 47 p = 10
Deviance = 559.21551 Null Deviance = 1790.99166 (Difference = 1231.77615)
```

```r
drop1(logitm, scale=sigma2, test="F")
```

```
Warning in drop1.glm(logitm, scale = sigma2, test = "F"): F test assumes
'quasibinomial' family


Single term deletions

Model:
cbind(germ, 100 - germ) ~ moisture + I(moisture^2) + box + covered

scale:  13.76635

              Df Deviance    AIC F value    Pr(>F)
<none>             559.22 768.96
moisture       1  1389.54 827.27 54.9374 8.065e-09 ***
I(moisture^2)  1  1624.40 844.33 70.4773 4.298e-10 ***
box            6   562.43 757.19  0.0355    0.9998
covered        0   559.22 768.96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated dispersion parameter value is 13.77, which is much larger than the assumption of 1 we used in last problem.

From the summary and F-test results, we find the standard error and p-value of coefficients changed. But the conclusion of no box effect still holds.

Problem 3

First we consider the binomial model.

The deviation is defined as

$$D = 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right], \hat{y}_i = n_i\hat{p}_i$$

We can rewrite the deviance as

$$\begin{aligned}
D &= 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right] \\
&= 2\sum_{i=1}^{n}\left[y_i \log y_i - y_i \log n_i\hat{p}_i + (n_i - y_i)\log(n_i - y_i) - (n_i - y_i)\log(n_i - n_i\hat{p}_i)\right] \\
&= -2\sum_{i=1}^{n} y_i(\log n_i + \log\hat{p}_i) + (n_i - y_i)(\log n_i + \log(1 - \hat{p}_i)) + C_1(n_i, y_i) \\
&= -2\sum_{i=1}^{n} y_i \log\hat{p}_i + (n_i - y_i)\log(1 - \hat{p}_i) + C_2(n_i, y_i)
\end{aligned}$$

in which $C_1(n_i, y_i), C_2(n_i, y_i)$ are terms that only depend on $n_i, y_i, i = 1, 2, ...n$.

AIC of binomial model is defined as

$$\begin{aligned}
AIC &= -2\log(\hat{L}) + 2q \\
&= -2\sum_{i=1}^{n}\left[\log\begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \log\hat{p}_i + (n_i - y_i)\log(1 - \hat{p}_i)\right] + 2q \\
&= -2\sum_{i=1}^{n} y_i \log\hat{p}_i + (n_i - y_i)\log(1 - \hat{p}_i) + 2q + C_3(n_i, y_i)
\end{aligned}$$

Comparing these two expressions, we find that

$$AIC^* = D + 2q = AIC - C_3(n_i, y_i) + C_2(n_i, y_i)$$

Note that the difference between AIC and AIC* is a term that only depend on $n_i, y_i, i = 1, 2, ...n$, which is the same across all models. Therefore, minimize AIC* criteria is equivalent to minimize AIC criteria.

Next, for binary model, deviation is defined as

$$D = -2\sum_{i=1}^{n}\left[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right]$$

Note that in binomial model, if we take all $n_i = 1$, then the deviance will become the same as that in binary situation except for a constant difference, which is independent of model and only depends on data.

$$D = -2\sum_{i=1}^{n} y_i \log\hat{p}_i + (1 - y_i)\log(1 - \hat{p}_i) + C_2(n_i = 1, y_i)$$

Moreover, for AIC of binary model, we can take all $n_i = 1$ in AIC of binomial model to get the result.

Therefore, for both deviance and AIC, binary model is a special case of binomial model. So previous conclusion still holds.

From above, selecting the model via AIC criteria can be done by picking the model that minimizes the AIC* criteria