사례

공원에서 낚시꾼들이 몇 마리의 물고기를 잡았는지 분석

- data on 250 groups that went to a park.

- **종속변수**

- count: 각 그룹이 잡은 물고기의 수

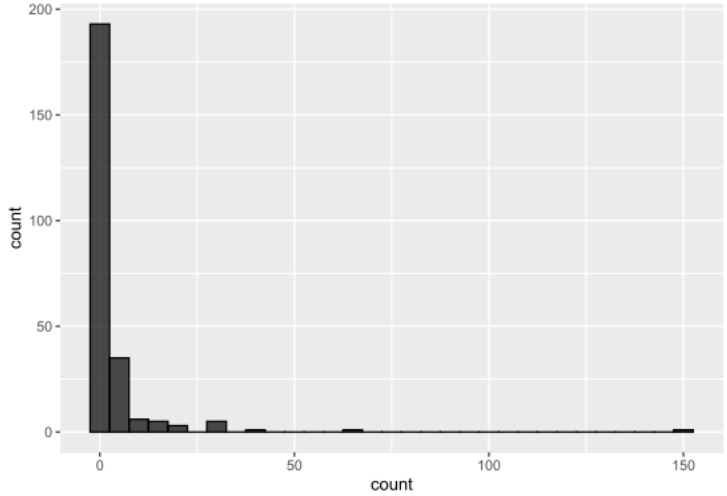
- **독립변수**

1. child: 그룹 내 아이들의 수
2. persons: 그룹 내 인원 수
3. camper: 캠핑카를 가져온 여부(0: no, 1: yes)

→ 조사 대상 그룹이 낚시를 했는지 안했는지 여부에 대한 정보는 포함하고 있지 않음
→ 낚시를 하지 않은 그룹은 perfect state

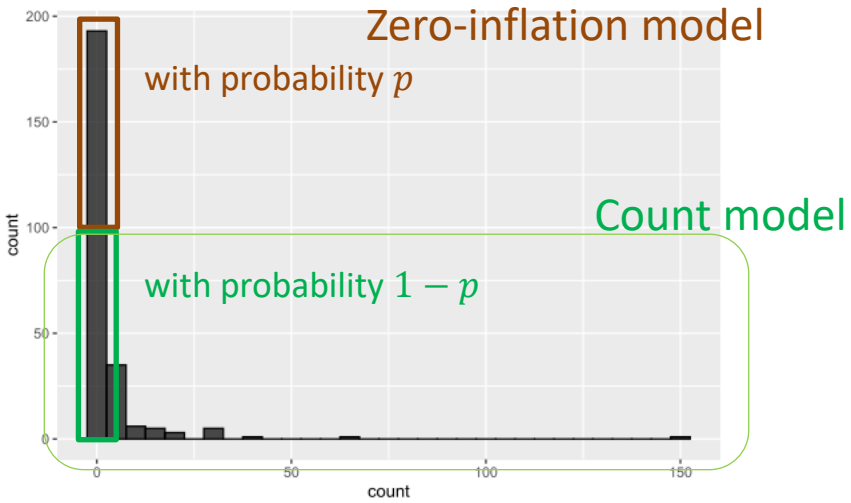
사례

- 히스토그램: count



사례

- 히스토그램: count



사례

- Zero-inflated Poisson regression

- **'pscl'** package: **zeroinfl**()

Usage

```
zeroinfl(formula, data, subset, na.action, weights, offset,
  dist = c("poisson", "negbin", "geometric"),
  link = c("logit", "probit", "cloglog", "cauchit", "log"),
  control = zeroinfl.control(...),
  model = TRUE, y = TRUE, x = FALSE, ...)
```

영과잉 포아송모형을 비롯해 영과잉 음이항모형 등 Zero-inflated count data regression을 위한 함수

사례

- Zero-inflated Poisson regression

```
m1 <- zeroinfl(count ~ child + camper + persons, data = data)
summary(m1)
```

```
##
## Call:
## zeroinfl(formula = count ~ child + camper + persons, data = data)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -3.05440 -0.74336 -0.44275 -0.07559 27.99304
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.79826    0.17081  -4.673 2.96e-06 ***
## child       -1.13666    0.09299 -12.224  < 2e-16 ***
## camper1      0.72425    0.09314   7.776 7.51e-15 ***
## persons      0.82904    0.04395  18.862  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6636     0.5155   3.227  0.00125 **
## child         1.9046     0.3261   5.840 5.21e-09 ***
## camper1      -0.8336     0.3527  -2.364  0.01808 *
## persons      -0.9228     0.1992  -4.632 3.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mnull <- update(m1, .~1)
pchisq(2*(logLik(m1) - logLik(mnull)), df=6, lower.tail = F)
```

```
## 'log Lik.' 1.972431e-158 (df=8)
```

- **Count model**: 포아송 분포를 따르는 count에 대한 포아송 회귀모형

- **Zero-inflation model**: 그룹이 0인지 여부를 예측하는 로지스틱 모델

사례

- Zero-inflated Poisson regression

```
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.79826    0.17081  -4.673 2.96e-06 ***
## child       -1.13666    0.09299 -12.224  < 2e-16 ***
## camper1      0.72425    0.09314   7.776 7.51e-15 ***
## persons      0.82904    0.04395  18.862  < 2e-16 ***
```

- **Count model** $\log(\hat{\lambda}) = -0.7983 - 1.1367 \times child + 0.7243 \times camper + 0.8290 \times persons$

- 계수에 대한 해석은 일반적인 포아송 회귀 모형과 동일
- 잡은 물고기 수의 기댓값은 해당 독립변수가 한 단위 증가할 때 exp(coef)만큼 변화

ex) (동일한 camper와 persons에서) 그룹 내 아이의 수가 한 명 증가할 때 잡는 물고기는 exp(−1.1367) = 0.3209배로 줄어든다. 따라서, 그룹 내 아이의 수가 많아질수록 잡는 물고기의 수는 줄어든다

◆ 단, 이때의 그룹은 perfect state에 해당하지 않는 그룹

사례

- Zero-inflated Poisson regression

```
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6636     0.5155   3.227  0.00125 **
## child         1.9046     0.3261   5.840 5.21e-09 ***
## camper1      -0.8336     0.3527  -2.364  0.01808 *
## persons      -0.9228     0.1992  -4.632 3.62e-06 ***
```

- **Zero-inflation model** $\text{logit}(\hat{p}) = 1.6636 + 1.9046 \times child - 0.8336 \times camper - 0.9228 \times persons$

- 계수에 대한 해석은 일반적인 로지스틱 회귀 모형과 동일
- Perfect state에 속할 오즈는 해당 독립변수가 한 단위 증가할 때 exp(coef)만큼 변화

ex) (동일한 camper와 persons에서) 그룹 내 아이의 수가 한 명 증가할 때 해당 그룹이 perfect state 그룹에 속할 오즈는 exp(1.9046)= 6.7167배로 증가
다시 말해, 그룹 내 속한 아이가 많을수록 해당 그룹은 낚시를 하지 않았을 가능성이 높다.

사례

- Zero-inflated Poisson regression

- **Count model** $\log(\hat{\lambda}) = -0.7983 - 1.1367 \times child + 0.7243 \times camper + 0.8290 \times persons$

- **Zero-inflation model** $\text{logit}(\hat{p}) = 1.6636 + 1.9046 \times child - 0.8336 \times camper - 0.9228 \times persons$

ex) (child, camper, persons)=(0,0,1) 인 그룹이 잡은 물고기 수의 기댓값

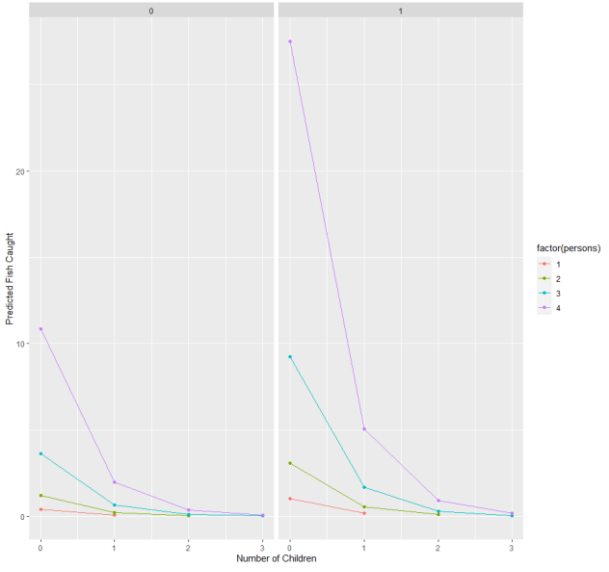
$$\hat{p} = \frac{exp(1.6636 - 0.9228 \times 1)}{1 + exp(1.6636 - 0.9228 \times 1)} = 0.6772$$

$$\hat{\lambda} = \exp(-0.7983 + 0.8290 \times 1) = 1.0312$$

$$E(count|\boldsymbol{x}) = (1 - \hat{p}) \times \hat{\lambda} = 0.3329$$

사례

- 결과 시각화


0
1
20
10
0
Predicted Fish Caught
0
1
2
3
Number of Children
factor(persons)
1
2
3
4

사례

- Zero-inflated Poisson regression

```
m2 <- zeroinfl(count ~ child + camper|persons, data = data)
summary(m2)
```

```
##
## Call:
## zeroinfl(formula = count ~ child + camper | persons, data = data)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.2369 -0.7540 -0.6080 -0.1921 24.0847
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.59789    0.08554  18.680   <2e-16 ***
## child       -1.04284    0.09999 -10.430   <2e-16 ***
## camper1      0.83402    0.09363   8.908   <2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2974     0.3739   3.470 0.000520 ***
## persons      -0.5643     0.1630  -3.463 0.000534 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 10
## Log-likelihood: -1032 on 5 Df
```

- **Count model과 Zero-inflation model에 사용할 변수를 지정가능**

References

Agresti, Alan and Maria Kateri. "Categorical Data Analysis". In: International Encyclopedia ofStatistical Science. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 206–208. isbn: 978-3-642-04898-2. doi: 10.1007/978- 3- 642- 048982_161. url: https://doi.org/10.1007/978-3-642-048982_161.

Lambert, Diane. "Zero-inflated Poisson regression, with an application to defects in manufacturing". In: Technometrics 34.1 (1992), pp. 1–14

ZERO-INFLATED POISSON REGRESSION | R DATA ANALYSIS EXAMPLES. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/r/dae/zip/

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Publications. Everitt, B. S. and Hothorn, T. A Handbook of Statistical Analyses Using R

END