Lecture 26

Data

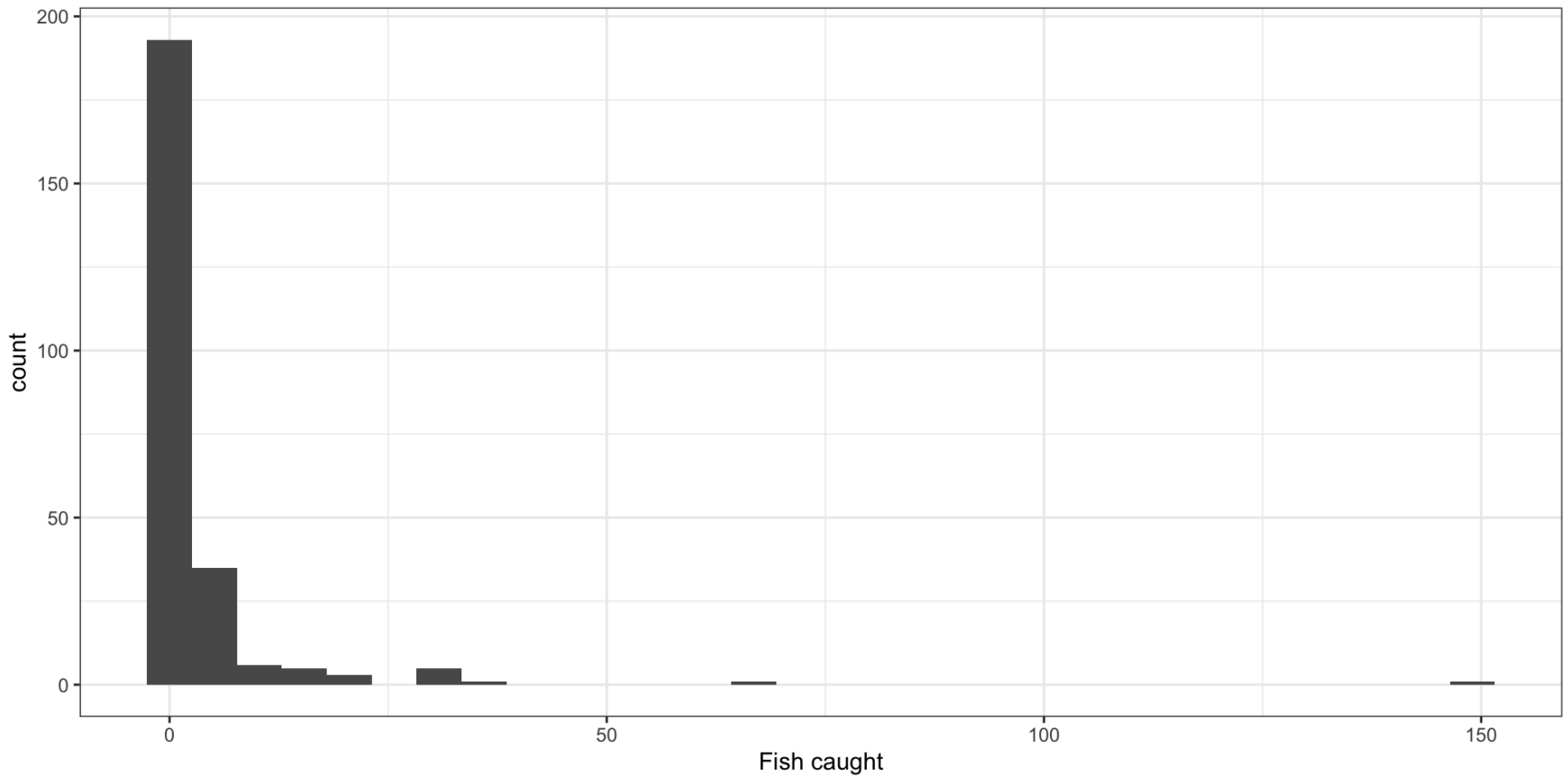
Two sources of zeros:
- bad anglers
- conscientious objectors

Data on the number of fish caught by campers in a state park. We have a sample of 250 groups of park guests who visited the state park. For each group, we record:

- `count`: the number of fish caught by the group
- `camper`: whether the group brought a camper van
- `child`: the number of children in the group
- `persons`: the total number of people in the group
- `LOS`: length of stay (in days)

What model is appropriate if the number of fish is our response?

Some EDA




```
1 mean(fish$count == 0)
```

```
[1] 0.568
```

Research question

Park rangers at the state park wonder whether groups with many children tend to catch fewer fish. They ask you to fit a model to investigate their hypothesis, and they want you to account for the total number of visitors in the group and whether the group brought a camper van (they suspect that camper vans make noise that scares away the fish).

Model

Ideally: might want to know where they are fishing

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1 - p_i) + p_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1 - p_i) & y > 0 \end{cases}$$

where

total # of fish caught

$$\log\left(\frac{p_i}{1 - p_i}\right) = \gamma_0 + \gamma_1 \text{Camper}_i + \gamma_2 \text{Child}_i + \gamma_3 \text{Persons}_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Camper}_i + \beta_2 \text{Child}_i + \beta_3 \text{Persons}_i$$

Question: Is there anything else we should add?

expected # of fish caught (if group went fishing)

Groups who stay for longer have more opportunity to catch fish!

Idea: compare the rate of fish caught (# fish per day)

Offsets

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Camper}_i + \beta_2 \text{Child}_i + \beta_3 \text{Persons}_i + \underbrace{\log(\text{LOS}_i)}_{\text{offset term (no } \beta!)}$$

$$\Rightarrow \log(\lambda_i) - \log(\text{LOS}_i) = \beta_0 + \beta_1 \text{Camper}_i + \beta_2 \text{Child}_i + \beta_3 \text{Persons}$$

$$\Rightarrow \log\left(\frac{\lambda_i}{\text{LOS}_i}\right) = \beta_0 + \beta_1 \text{Camper}_i + \beta_2 \text{Child}_i + \beta_3 \text{Persons}_i$$

Expected # of Fish caught per day $= \frac{\lambda_i}{\text{LOS}_i} = \exp\{\ldots\ldots\}$

Interpretation changes! But still modeling Y_i

$$Y_i \mid (Z_i = 0) \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta^T X_i + \log(LOS_i)$$

$$\Rightarrow \log\left(\frac{\lambda_i}{LOS_i}\right) = \beta^T X_i$$

$\Rightarrow$ response variable is still Y_i (not $\frac{Y_i}{LOS_i}$)

βs interpreted in terms of rate, not raw counts

Fitting a model with an offset

```
1 m1 <- glm(count ~ camper + child + persons,
2           offset = log(LOS), data = fish, family = poisson)
3 summary(m1)$coefficients
```

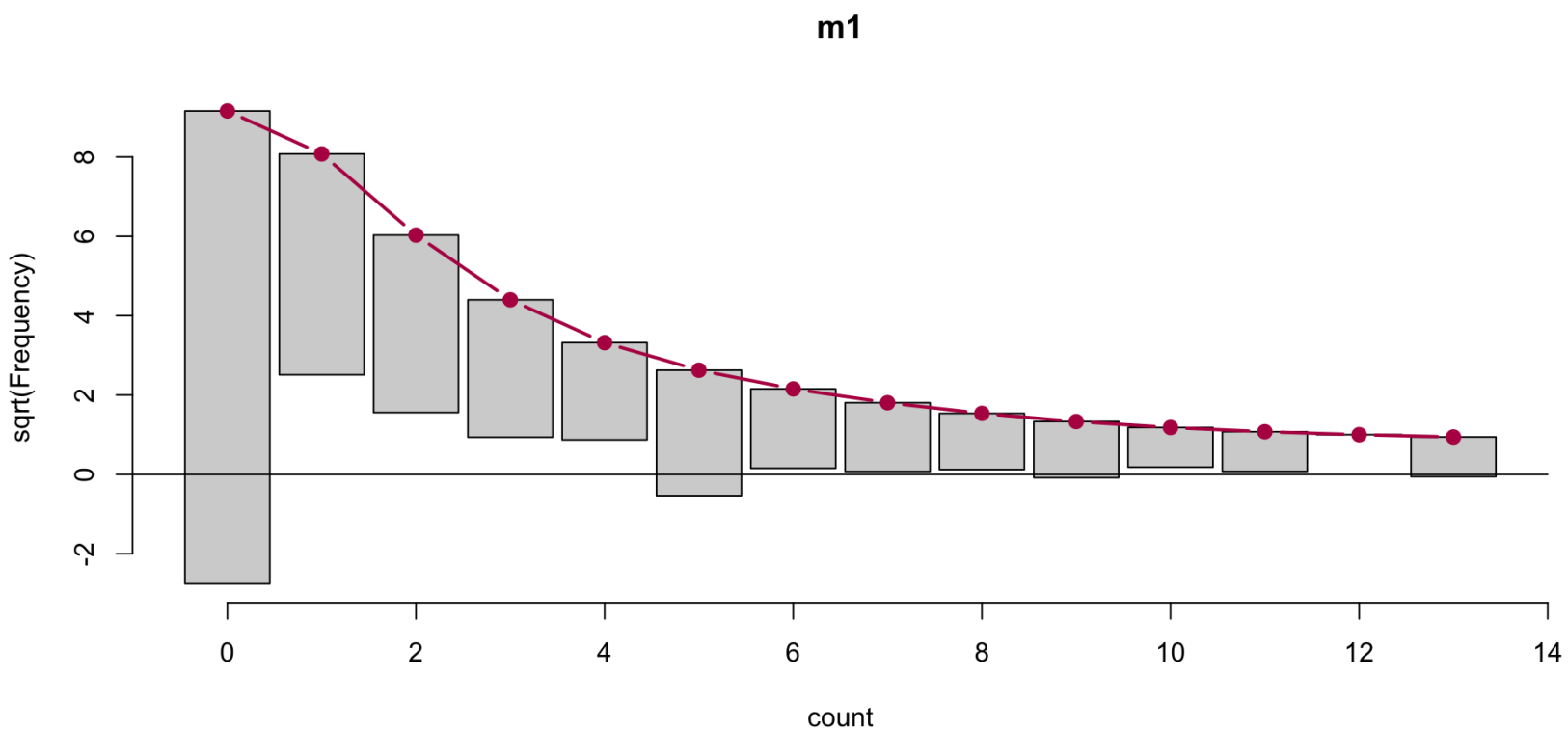
```
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.0387342 0.14014138 -14.547695 6.040389e-48
camper       0.2782194 0.09159030   3.037651 2.384296e-03
child       -1.1001418 0.07838568 -14.034985 9.521603e-45
persons      0.6307856 0.03791446  16.637073 3.755316e-62
```

(no offset term in the coefficients)

holding camper & child fixed, an increase of 1 person in group size is associated with an increase in the expected # of fish caught per day by a factor of $\exp\{0.6313\}$

Assessing the Poisson model

```
1 library(countreg)
2
3 rootogram(m1)
```



ZIP model (with offset)

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1 - p_i) + p_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1 - p_i) & y > 0 \end{cases}$$

where

$$\log\left(\frac{p_i}{1 - p_i}\right) = \gamma_0 + \gamma_1 \text{Camper}_i + \gamma_2 \text{Child}_i + \gamma_3 \text{Persons}_i$$

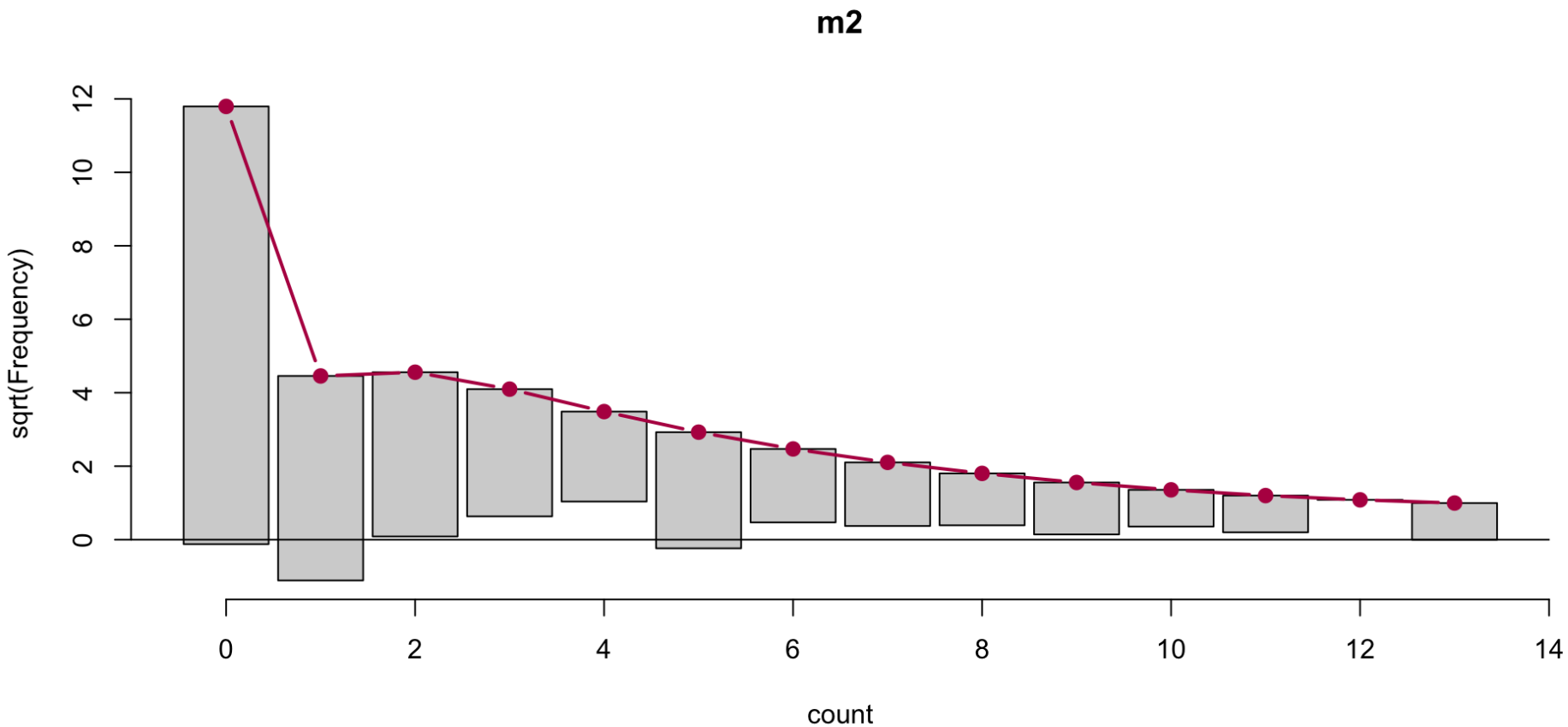
$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Camper}_i + \beta_2 \text{Child}_i + \beta_3 \text{Persons}_i + \log(\text{LO}$$

log(LOS$_i$)

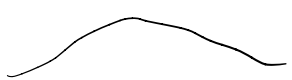
```
m2 <- zeroinfl(count ~ camper + child + persons,
               offset = log(LOS),
               data = fish)
```

Diagnostics

```
1 rootogram(m2)
```

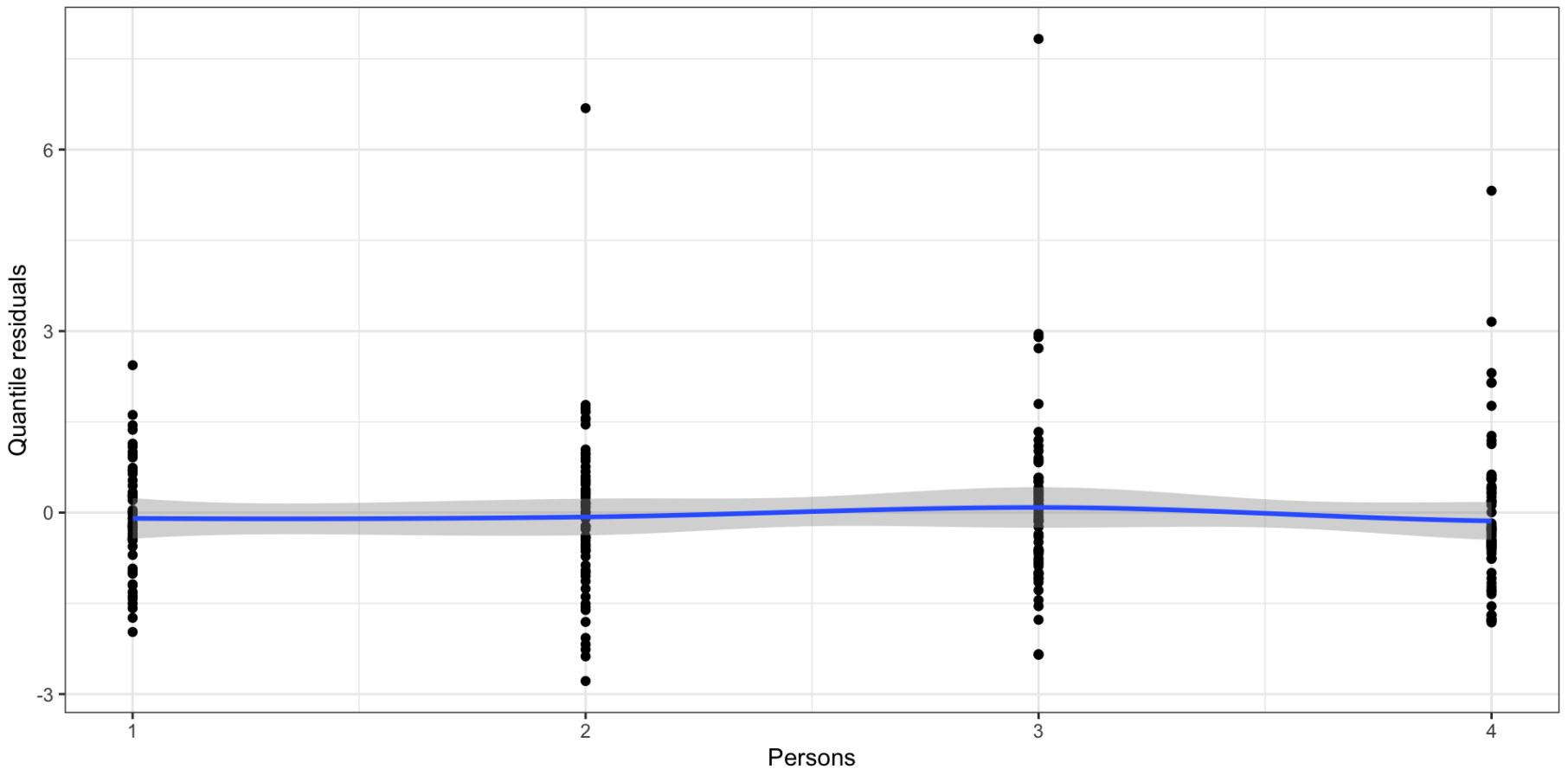


little bit of a wave pattern



Diagnostics

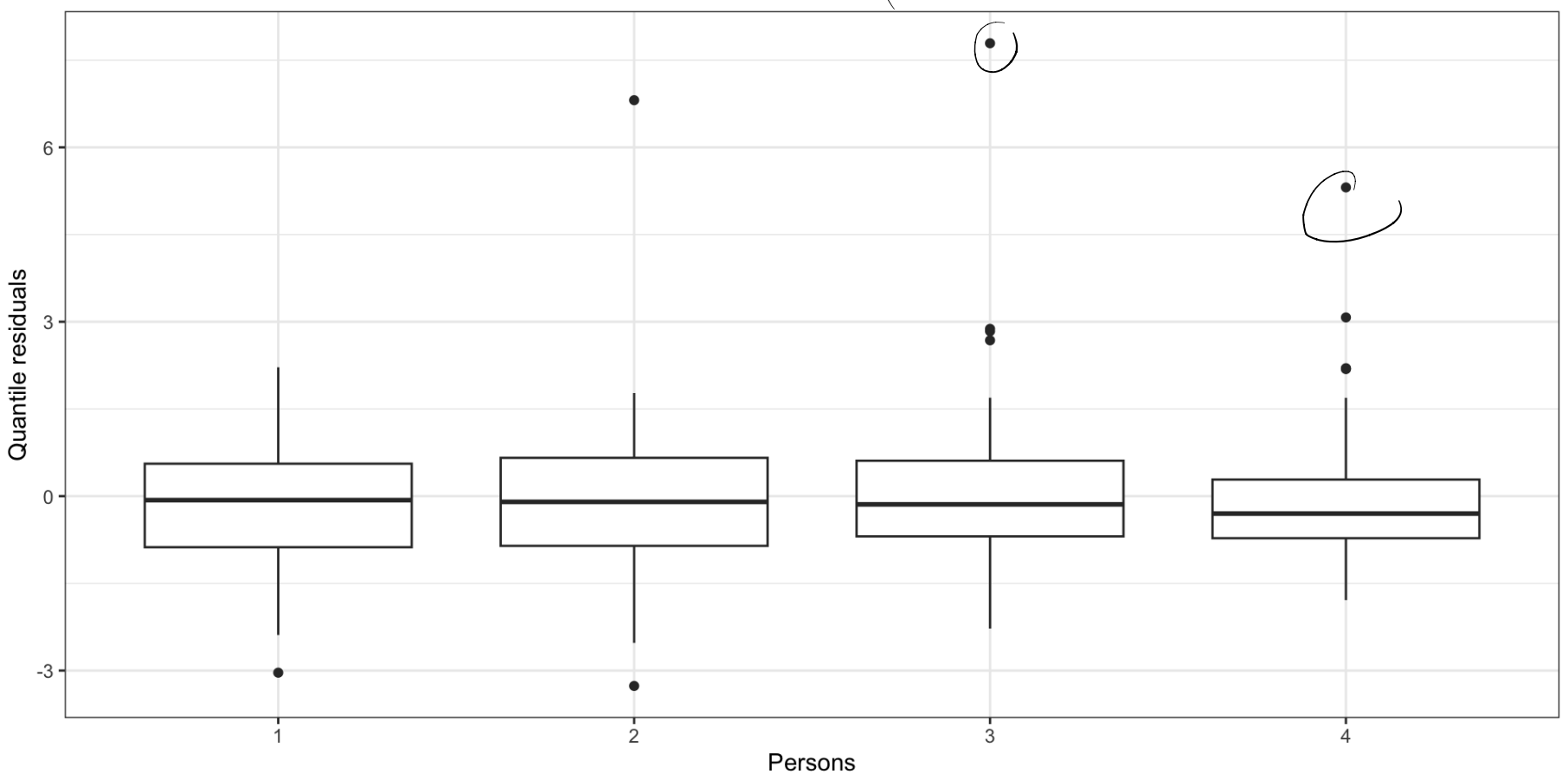
Quantile residual plot:



Diagnostics

Quantile residual plot:

check for influential points



Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_26.html