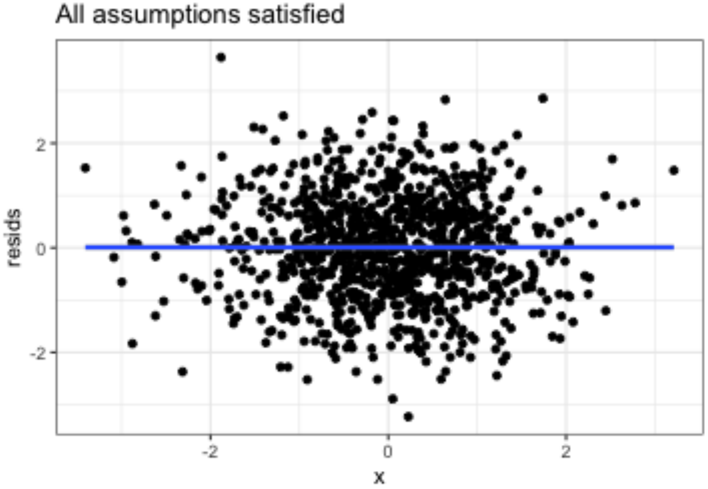
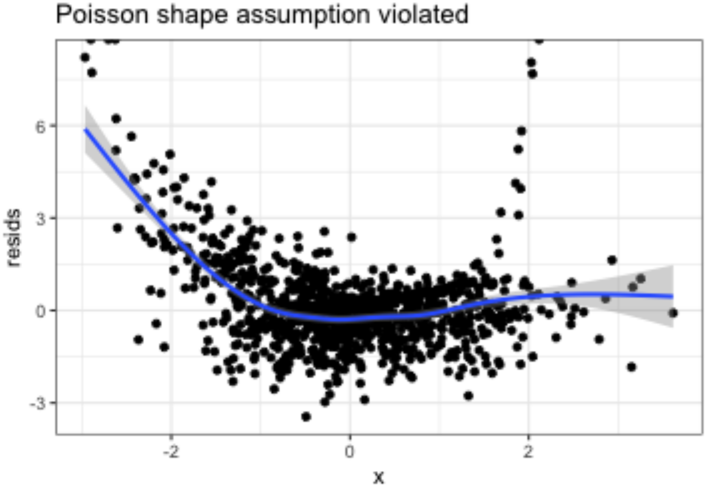
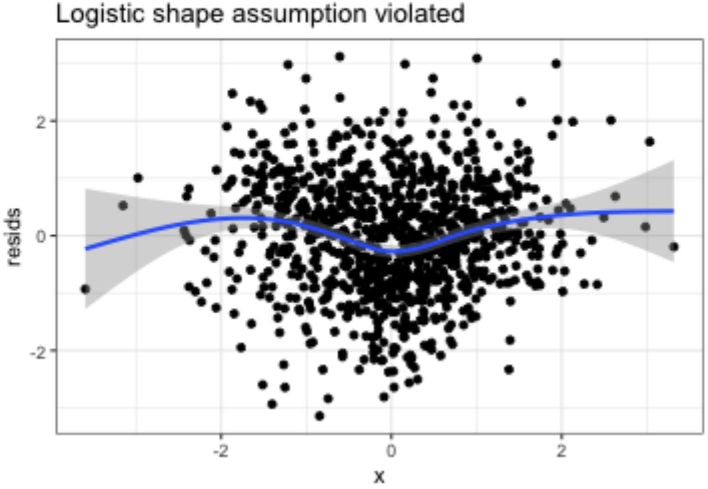
# ZIP models

# Recap: Assessing the shape assumption


All assumptions satisfied


Poisson shape assumption violated


Logistic shape assumption violated

Quantile residual plots show violations of shape assumption

# Logistic component vs. Poisson component

1) Fit a ZIP model to the full data

2) Create a quantile residual plot for the ZIP
Any violations?

3) Now look @ Poisson component directly

- If $Y_i > 0$, $Y_i$ comes from Poisson component
- $Y_i \mid Y_i > 0$ $\underline{\text{not}}$ distributed Poisson($\lambda_i$)

$Y_i \mid Y_i > 0 \sim$ Positive Poisson($\lambda_i$)

$$P(Y_i = y \mid Y_i > 0) = \frac{P(Y_i = y, y > 0)}{P(Y_i > 0)}$$

$$= \frac{e^{-\lambda_i}\lambda_i^{y}}{y!\,(1 - e^{-\lambda_i})}$$

- $Y_i \mid Y_i > 0 \sim$ Positive Poisson

$$\log(\lambda_i) = \underset{\substack{\uparrow}}{\beta^T} X_i$$

same $\beta$ as the Poisson component of the ZIP model

# Class activity

https://sta712-f22.github.io/class_activities/ca_lecture_33.html

# Class activity

$$Z_i = \begin{cases} 1 & \text{nonsmoker} \\ 0 & \text{sometime smoke} \end{cases}$$

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 EducationSome_i + \gamma_2 EducationCollege_i$$
$$\gamma_3 EducationAdv_i + \gamma_4 Diabetes_i + \gamma_5 Age_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 EducationSome_i + \beta_2 EducationCollege_i +$$
$$\beta_3 EducationAdv_i + \beta_4 Diabetes_i + \beta_5 Age_i$$

> Research question: for smokers, does the number of cigarettes smoked per day depend on age?
>
> What are the null and alternative hypotheses?

$H_0: \beta_5 = 0$ $\quad$ $H_A: \beta_5 \neq 0$

# Class activity

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 EducationSome_i + \gamma_2 EducationCollege_i$$
$$\gamma_3 EducationAdv_i + \gamma_4 Diabetes_i + \gamma_5 Age_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 EducationSome_i + \beta_2 EducationCollege_i +$$
$$\beta_3 EducationAdv_i + \beta_4 Diabetes_i + \beta_5 Age_i$$

> Research question: is there a relationship between age and whether someone is a smoker?
>
> What are the null and alternative hypotheses?

$H_0: \quad \gamma_5 = 0$ $\qquad\qquad$ $H_A: \quad \gamma_5 \neq 0$

# Wald tests

$\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix} \overset{\sim}{\cdot}$ Normal for large $n$

$\Rightarrow$ can use Wald tests!

> Research question: is there a relationship between age and whether someone is a smoker?

```
m1 <- zeroinfl(cigsPerDay ~ education + diabetes +
                 age | education + diabetes + age,
               data = heart_data)
summary(m1)
```

```
...
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.49673    0.20977 -11.902   <2e-16 ***
## education2  -0.06100    0.07840  -0.778   0.4366
## education3   0.17141    0.09362   1.831   0.0671 .
## education4   0.03547    0.10749   0.330   0.7414
## diabetes     0.26063    0.20854   1.250   0.2114
## age          0.05071    0.00395  12.838   <2e-16 ***
...
```

$\hat{\gamma}_s$

$z = \dfrac{\hat{\gamma}_s - 0}{SE_{\hat{\beta}_s}}$

p-value $\approx 0$

# Class activity

Test: LRT

Statistic: $2 * (\ell_{full} - \ell_{reduced})$

(not deviance b/c not EDM)

df: 6

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 EducationSome_i + \gamma_2 EducationCollege_i$$

$$\gamma_3 EducationAdv_i + \gamma_4 Diabetes_i + \gamma_5 Age_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 EducationSome_i + \beta_2 EducationCollege_i +$$

$$\beta_3 EducationAdv_i + \beta_4 Diabetes_i + \beta_5 Age_i$$

> Research question: Is there a relationship between education level and the number of cigarettes smoked?
>
> What are the null and alternative hypotheses?

$H_0$: $\gamma_1 = \gamma_2 = \gamma_3 = \beta_1 = \beta_2 = \beta_3 = 0$

$H_A$: at least one of $\gamma_1, \gamma_2, \gamma_3, \beta_1, \beta_2, \beta_3 \neq 0$

# Likelihood ratio test

```r
m1 <- zeroinfl(cigsPerDay ~ education + diabetes +
                 age | education + diabetes + age,
               data = heart_data)
m2 <- zeroinfl(cigsPerDay ~ education + diabetes
               | education + diabetes,
               data = heart_data)

2*(m1$loglik - m2$loglik)
```

```
## [1] 242.281
```

```r
pchisq(242.281, df=6, lower.tail=F)
```

```
## [1] 1.828386e-49
```