Dealing with the badness of goodness-of-fit

Rishika Chopara   Dr Ben Stevenson   Professor Rachel Fewster

Department of Statistics, University of Auckland

30 November 2023

Introduction

Introduction

Under H_0 that a model fits the data well, model deviance $\sim \chi^2$

Introduction

Under H_0 that a model fits the data well, model deviance $\sim \chi^2$

```
Call:
glm(formula = ofp ~ ., family = poisson, data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4055  -1.9962  -0.6737   0.7049  16.3620

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.028874   0.023785  43.258   <2e-16 ***
hosp             0.164797   0.005997  27.478   <2e-16 ***
healthpoor       0.248307   0.017845  13.915   <2e-16 ***
healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
numchron         0.146639   0.004580  32.020   <2e-16 ***
gendermale      -0.112320   0.012945  -8.677   <2e-16 ***
school           0.026143   0.001843  14.182   <2e-16 ***
privinsyes       0.201687   0.016860  11.963   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 23168  on 4398  degrees of freedom
AIC: 35959

Number of Fisher Scoring iterations: 5
```

Introduction

Under H_0 that a model fits the data well, model deviance $\sim \chi^2$

```
Call:
glm(formula = ofp ~ ., family = poisson, data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4055  -1.9962  -0.6737   0.7049  16.3620

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.028874   0.023785  43.258   <2e-16 ***
hosp             0.164797   0.005997  27.478   <2e-16 ***
healthpoor       0.248307   0.017845  13.915   <2e-16 ***
healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
numchron         0.146639   0.004580  32.020   <2e-16 ***
gendermale      -0.112320   0.012945  -8.677   <2e-16 ***
school           0.026143   0.001843  14.182   <2e-16 ***
privinsyes       0.201687   0.016860  11.963   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 23168  on 4398  degrees of freedom
AIC: 35959

Number of Fisher Scoring iterations: 5
```

Introduction

Under H_0 that a model fits the data well, model deviance $\sim \chi^2$

```
Call:
glm(formula = ofp ~ ., family = poisson, data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4055  -1.9962  -0.6737   0.7049  16.3620

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.028874   0.023785  43.258   <2e-16 ***
hosp             0.164797   0.005997  27.478   <2e-16 ***
healthpoor       0.248307   0.017845  13.915   <2e-16 ***
healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
numchron         0.146639   0.004580  32.020   <2e-16 ***
gendermale      -0.112320   0.012945  -8.677   <2e-16 ***
school           0.026143   0.001843  14.182   <2e-16 ***
privinsyes       0.201687   0.016860  11.963   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 23168  on 4398  degrees of freedom
AIC: 35959

Number of Fisher Scoring iterations: 5
```

Introduction

Under H_0 that a model fits the data well, model deviance $\sim \chi^2$

```
Call:
glm(formula = ofp ~ ., family = poisson, data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4055  -1.9962  -0.6737   0.7049  16.3620

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.028874   0.023785  43.258   <2e-16 ***
hosp             0.164797   0.005997  27.478   <2e-16 ***
healthpoor       0.248307   0.017845  13.915   <2e-16 ***
healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
numchron         0.146639   0.004580  32.020   <2e-16 ***
gendermale      -0.112320   0.012945  -8.677   <2e-16 ***
school           0.026143   0.001843  14.182   <2e-16 ***
privinsyes       0.201687   0.016860  11.963   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 23168  on 4398  degrees of freedom
AIC: 35959

Number of Fisher Scoring iterations: 5
```

```
1 - pchisq(23168, 4398)
```

Definitions

Definitions

- Traditionally, deviance $= 2(\log(\hat{L}_{sat}) - \log(\hat{L}_{simple}))$

Definitions

- Traditionally, deviance $= 2(\log(\hat{L}_{sat}) - \log(\hat{L}_{simple}))$
  - L_{simple} evaluated at MLEs

Definitions

- Traditionally, deviance $= 2(\log(\hat{L}_{sat}) - \log(\hat{L}_{simple}))$
  - L_{simple} evaluated at MLEs
  - Empirical deviance, D_E

Definitions

- Traditionally, deviance $= 2(\log(\hat{L}_{sat}) - \log(\hat{L}_{simple}))$
  - L_{simple} evaluated at MLEs
  - Empirical deviance, D_E
- Alternatively, could evaluate L_{simple} at true parameter values

Definitions

- Traditionally, deviance $= 2(\log(\hat{L}_{sat}) - \log(\hat{L}_{simple}))$
  - L_{simple} evaluated at MLEs
  - Empirical deviance, D_E
- Alternatively, could evaluate L_{simple} at true parameter values
  - Generative deviance, D_G

Distribution of D_E

Distribution of D_E

Under H_0 that a model fits the data well,

Distribution of D_E

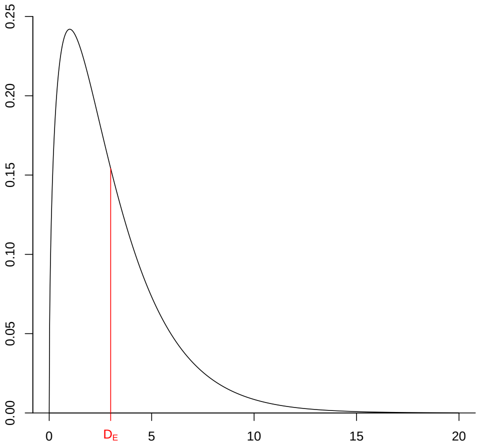
Under H_0 that a model fits the data well,

$$D_E \sim \chi^2_{df},$$

df = difference in free parameters for saturated and simple model

Distribution of D_E

Under H_0 that a model fits the data well,
$D_E \sim \chi^2_{df}$,
df = difference in free parameters for saturated and simple model

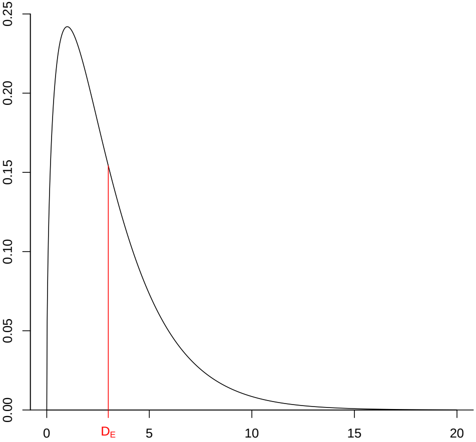


Distribution of D_E

Under H_0 that a model fits the data well,

$$D_E \sim \chi^2_{df},$$

df = difference in free parameters for saturated and simple model



Doesn't hold if data doesn't contain enough information...

Distribution of D_E?

$\mathrm{E}[Y_1]$	0.29
$\mathrm{E}[Y_2]$	1.3
$\mathrm{E}[Y_3]$	3.2
$\mathrm{E}[Y_4]$	0.3
$\mathrm{E}[Y_5]$	2.2
$\mathrm{E}[Y_6]$	29.1
$\mathrm{E}[Y_7]$	5.2
$\mathrm{E}[Y_8]$	14.2
...	...
$\mathrm{E}[Y_{90}]$	3.5

Distribution of D_E?

Distribution of D_E?

For capture-recapture/spatial capture-recapture, sparse sets of counts are common

Distribution of D_E?

For capture-recapture/spatial capture-recapture, sparse sets of counts are common

$E[Y_1]$	0.001
$E[Y_2]$	0.04
$E[Y_3]$	0.09
$E[Y_4]$	0.025
$E[Y_5]$	1.1
$E[Y_6]$	0.0003
$E[Y_7]$	0.29
$E[Y_8]$	0.047
...	...
$E[Y_{90}]$	0.99

Bootstrapping?

Bootstrapping?

Can approximate the distribution of D_E, without assuming χ^2

Bootstrapping?

Can approximate the distribution of D_E, without assuming χ^2

1. Fit model to dataset

Bootstrapping?

Can approximate the distribution of D_E, without assuming χ^2

1. Fit model to dataset
2. Calculate model deviance

Bootstrapping?

Can approximate the distribution of D_E, without assuming χ^2

1. Fit model to dataset
2. Calculate model deviance
3. Using MLEs, simulate m datasets

Bootstrapping?

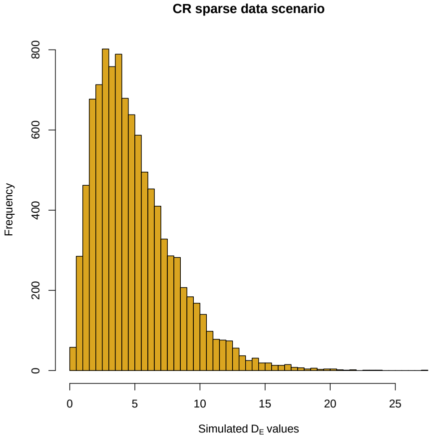
Can approximate the distribution of D_E, without assuming χ^2

1. Fit model to dataset
2. Calculate model deviance
3. Using MLEs, simulate m datasets
4. Fit the same model to each dataset, calculate deviance

Bootstrapping?

Can approximate the distribution of D_E, without assuming χ^2

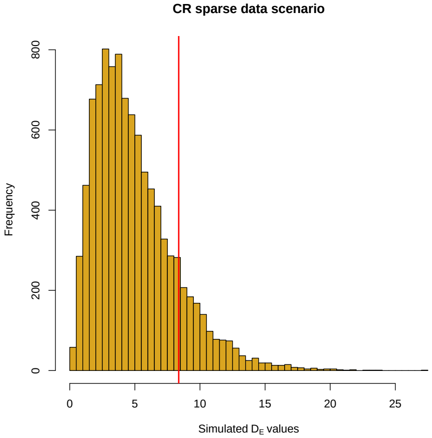
1. Fit model to dataset
2. Calculate model deviance
3. Using MLEs, simulate m datasets
4. Fit the same model to each dataset, calculate deviance



Bootstrapping?

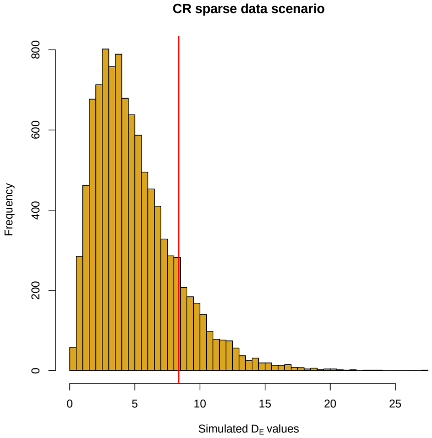
Can approximate the distribution of D_E, without assuming χ^2

1. Fit model to dataset
2. Calculate model deviance
3. Using MLEs, simulate m datasets
4. Fit the same model to each dataset, calculate deviance



Bootstrapping?

Can approximate the distribution of D_E, without assuming χ^2

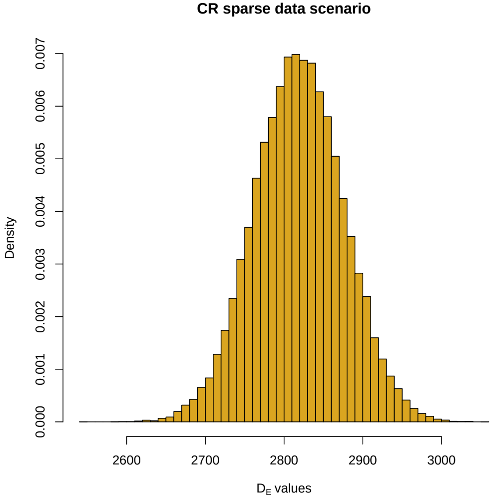


1. Fit model to dataset
2. Calculate model deviance
3. Using MLEs, simulate *m* datasets
4. Fit the same model to each dataset, calculate deviance

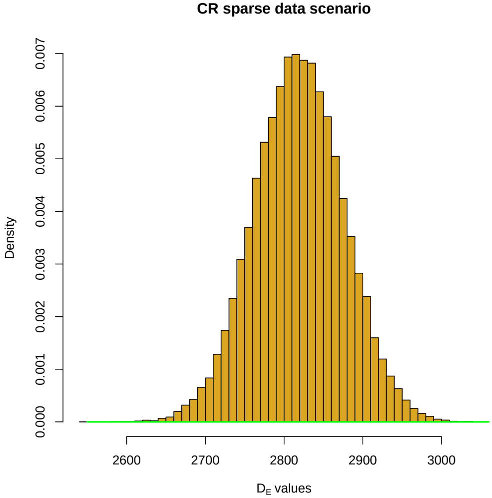
Can require a lot of time/resources

Chi-squared approximation when data is sparse

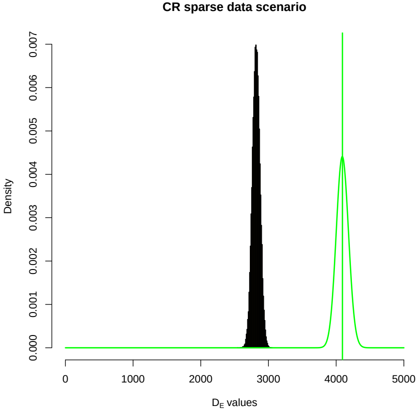
Chi-squared approximation when data is sparse



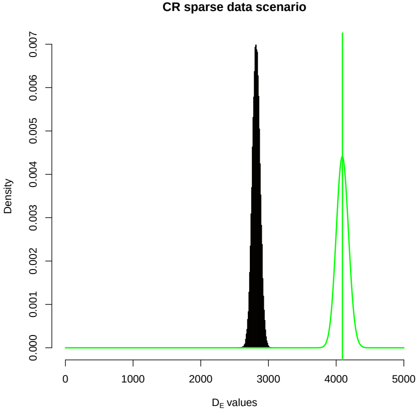
Chi-squared approximation when data is sparse



Chi-squared approximation when data is sparse

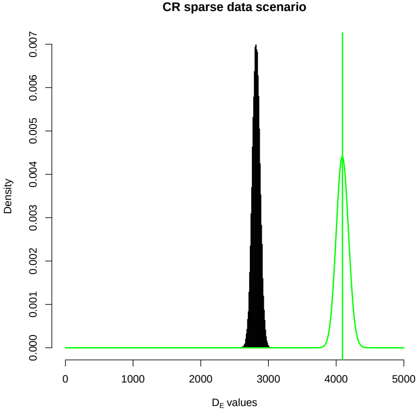


Chi-squared approximation when data is sparse



- Could adjusting the mean of the Chi-squared approximation work?

Chi-squared approximation when data is sparse



- Could adjusting the mean of the Chi-squared approximation work?
  - Need to find $E[D_E]$

Correct mean for the Chi-squared approximation?

Correct mean for the Chi-squared approximation?

- Looked at $D_G - D_E$

Correct mean for the Chi-squared approximation?

- Looked at $D_G - D_E$
  - Recall: $D_G = 2(\log(\hat{L}_{sat}) - \log(L_{simple}))$, L_{simple} evaluated at true parameter values

Correct mean for the Chi-squared approximation?

- Looked at $D_G - D_E$
  - Recall: $D_G = 2(\log(\hat{L}_{sat}) - \log(L_{simple}))$, L_{simple} evaluated at true parameter values
- $D_G - D_E \sim \chi^2_p$

Correct mean for the Chi-squared approximation?

- Looked at $D_G - D_E$
  - Recall: $D_G = 2(\log(\hat{L}_{sat}) - \log(L_{simple}))$, L_{simple} evaluated at true parameter values
- $D_G - D_E \sim \chi^2_p$
  - $p =$ number of parameters in the model we are considering

Correct mean for the Chi-squared approximation?

- Looked at $D_G - D_E$
  - Recall: $D_G = 2(\log(\hat{L}_{sat}) - \log(L_{simple}))$, L_{simple} evaluated at true parameter values
- $D_G - D_E \sim \chi^2_p$
  - $p =$ number of parameters in the model we are considering
- Therefore, $E[D_G - D_E] = p$

Correct mean for the Chi-squared approximation?

- Looked at $D_G - D_E$
  - Recall: $D_G = 2(\log(\hat{L}_{sat}) - \log(L_{simple}))$, L_{simple} evaluated at true parameter values
- $D_G - D_E \sim \chi^2_p$
  - $p =$ number of parameters in the model we are considering
- Therefore, $E[D_G - D_E] = p$
- By linearity of expectations, $E[D_E] = E[D_G] - p$

Correct mean for the Chi-squared approximation?

- We have: $E[D_E] = E[D_G] - p$

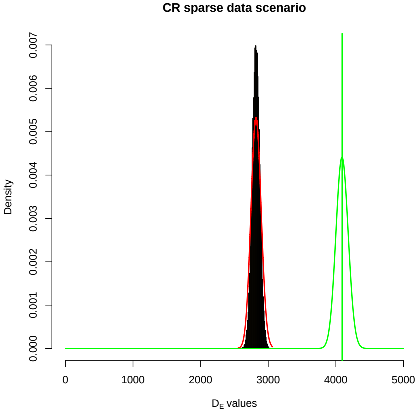
If we fit a closed-population CR model using a Poisson likelihood:

$$E[D_G] = \sum_{i=1}^{k} \sum_{n=1}^{\infty} \frac{\mu_i^n e^{-\mu_i}}{n!} \cdot 2\left[n\left(\log\left(\frac{n}{\mu_i}\right) - 1\right) + \mu_i\right]$$

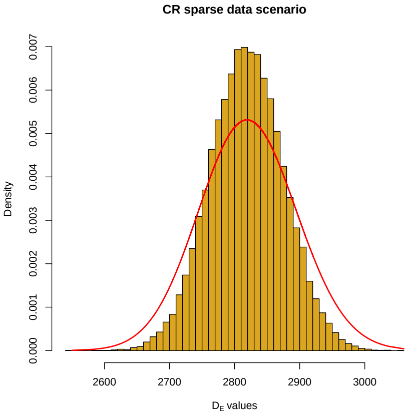
where:

- k is the number of observable capture histories;
- μ_i is the expected count for the ith observable capture history, evaluated at the generating parameter values.

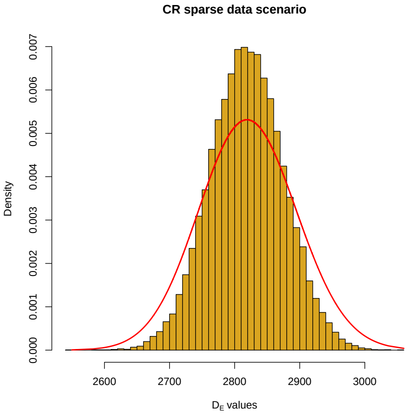
Correct mean for the Chi-squared approximation?



Correct mean for the Chi-squared approximation?

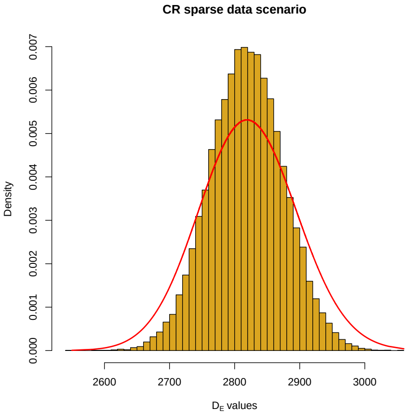


Correct mean for the Chi-squared approximation?



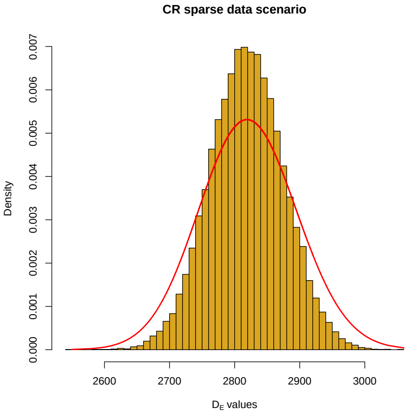
- Variance wrong

Correct mean for the Chi-squared approximation?



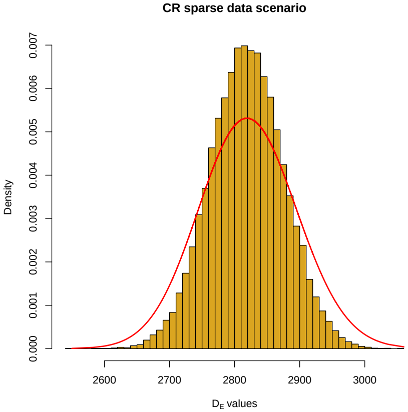
- Variance wrong
- χ^2 distribution is a special case of the Gamma distribution

Correct mean for the Chi-squared approximation?



- Variance wrong
- χ^2 distribution is a special case of the Gamma distribution
  - Can separately specify mean and variance for Gamma distribution

Correct mean for the Chi-squared approximation?



- Variance wrong
- χ^2 distribution is a special case of the Gamma distribution
  - Can separately specify mean and variance for Gamma distribution
  - Would a Gamma approximation work?

Variance for a Gamma approximation

Variance for a Gamma approximation

- If $\mathrm{Cov}(D_E, D_G - D_E) = 0 \implies \mathrm{Var}(D_E) = \mathrm{Cov}(D_E, D_G)$

Variance for a Gamma approximation

- If $\mathrm{Cov}(D_E, D_G - D_E) = 0 \implies \mathrm{Var}(D_E) = \mathrm{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\mathrm{Var}(D_G - D_E) = 2p$

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

$$\implies 2p = \text{Var}(D_G) + \text{Var}(D_E) - 2\text{Var}(D_E)$$

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

$$\implies 2p = \text{Var}(D_G) + \text{Var}(D_E) - 2\text{Var}(D_E)$$

$$\implies 2p = \text{Var}(D_G) - \text{Var}(D_E)$$

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

$$\implies 2p = \text{Var}(D_G) + \text{Var}(D_E) - 2\text{Var}(D_E)$$

$$\implies 2p = \text{Var}(D_G) - \text{Var}(D_E)$$

$$\implies \text{Var}(D_E) = \text{Var}(D_G) - 2p$$

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

$$\implies 2p = \text{Var}(D_G) + \text{Var}(D_E) - 2\text{Var}(D_E)$$

$$\implies 2p = \text{Var}(D_G) - \text{Var}(D_E)$$

$$\implies \text{Var}(D_E) = \text{Var}(D_G) - 2p$$

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

$$\implies 2p = \text{Var}(D_G) + \text{Var}(D_E) - 2\text{Var}(D_E)$$

$$\implies 2p = \text{Var}(D_G) - \text{Var}(D_E)$$

$$\implies \text{Var}(D_E) = \text{Var}(D_G) - 2p$$

- Using Wilks' Theorem, we can show this holds for a Normal response

Variance for a Gamma approximation

- If $\text{Cov}(D_E, D_G - D_E) = 0 \implies \text{Var}(D_E) = \text{Cov}(D_E, D_G)$
- $D_G - D_E \sim \chi^2_p$, so $\text{Var}(D_G - D_E) = 2p$
- $\text{Var}(D_G - D_E) = Var(D_G) + Var(D_E) - 2\text{Cov}(D_E, D_G)$

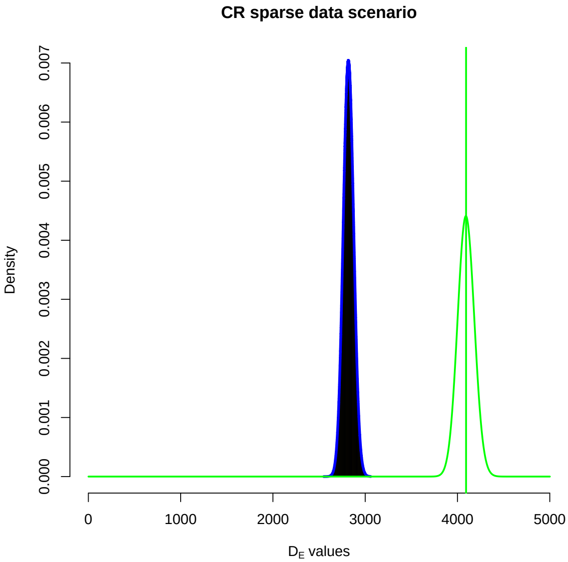
$$\implies 2p = \text{Var}(D_G) + \text{Var}(D_E) - 2\text{Var}(D_E)$$

$$\implies 2p = \text{Var}(D_G) - \text{Var}(D_E)$$

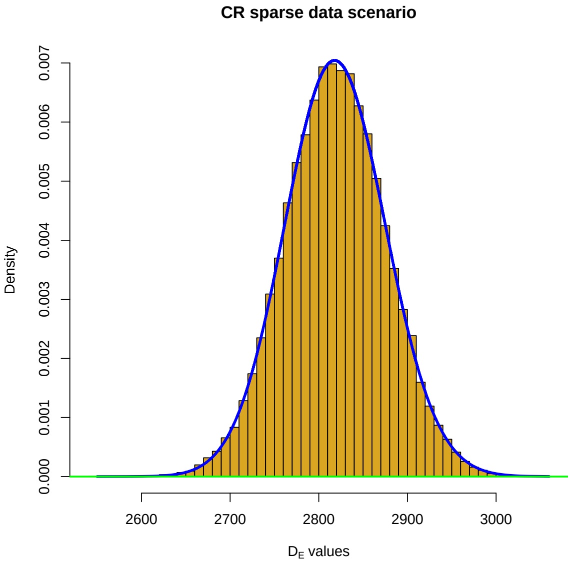
$$\implies \text{Var}(D_E) = \text{Var}(D_G) - 2p$$

- Using Wilks' Theorem, we can show this holds for a Normal response
- For other models (e.g. CR), have found by simulation this holds approximately

Performance of Gamma approximation

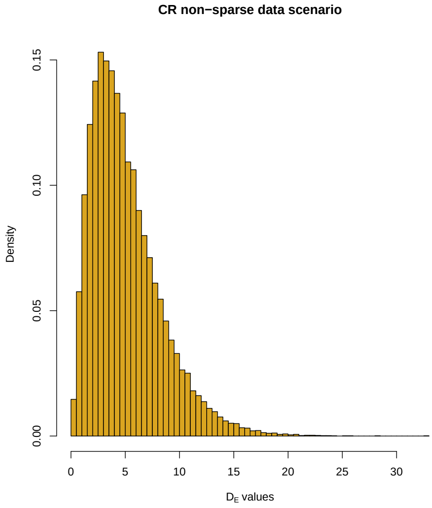


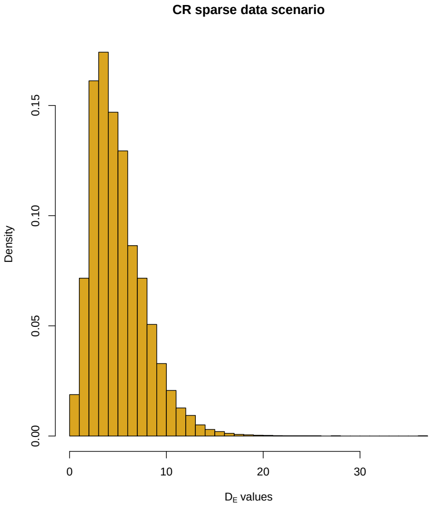
Performance of Gamma approximation



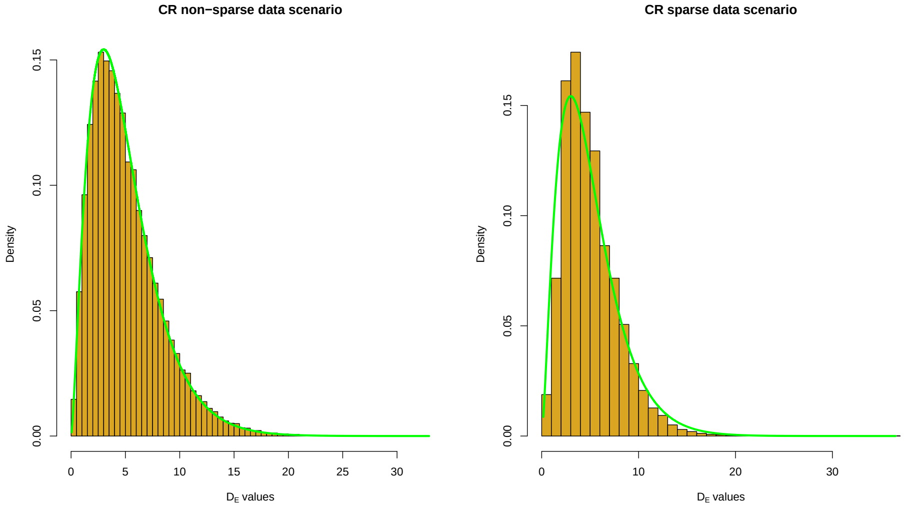
Performance of Gamma approximation

Performance of Gamma approximation

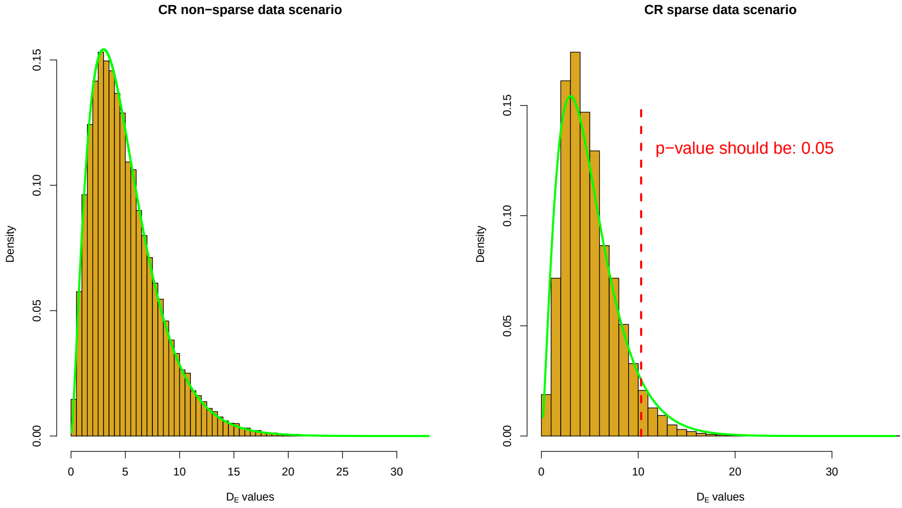




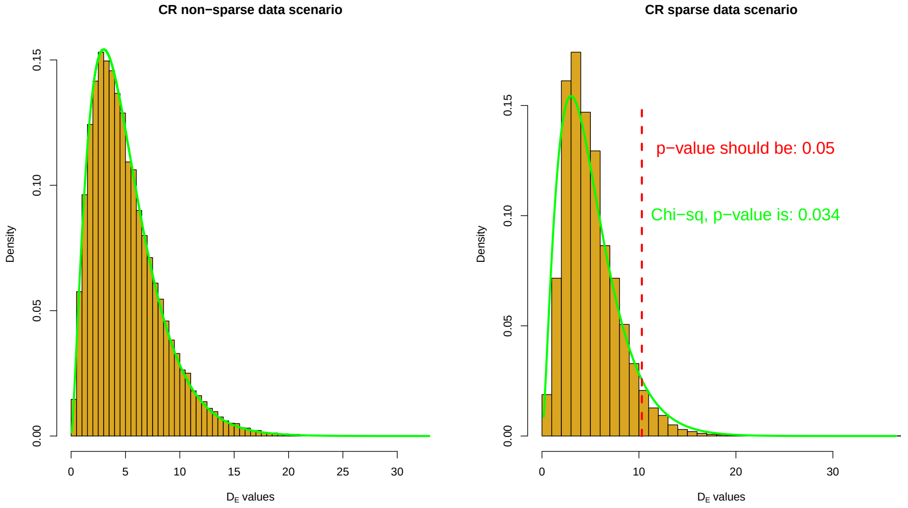
Performance of Gamma approximation



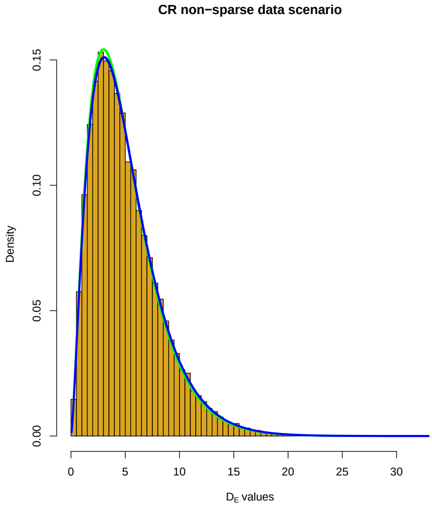
Performance of Gamma approximation

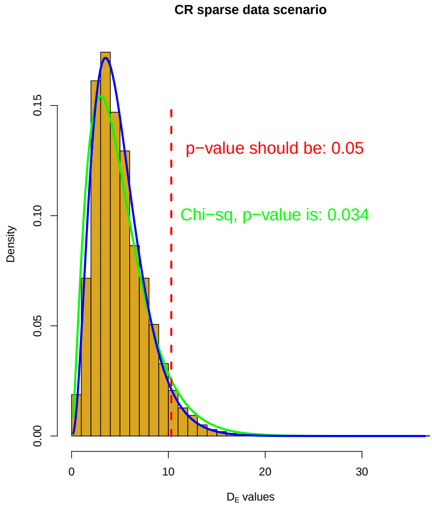


Performance of Gamma approximation

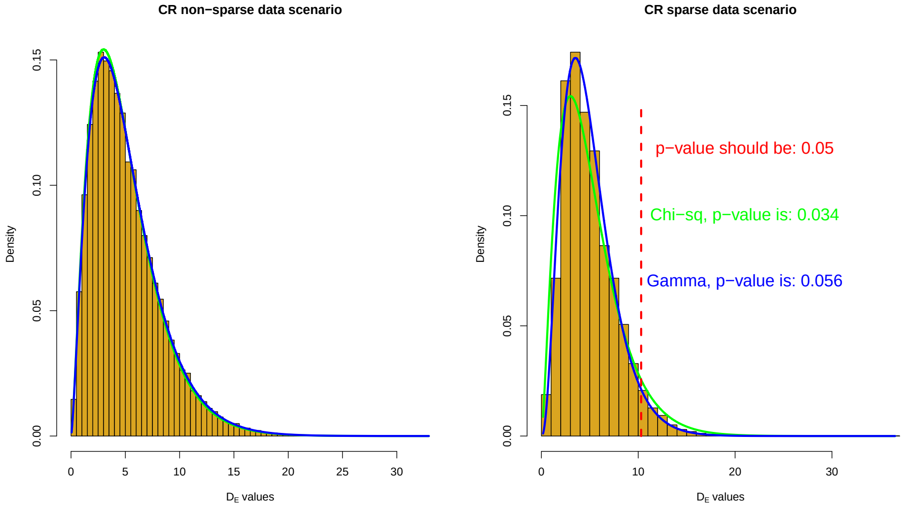


Performance of Gamma approximation

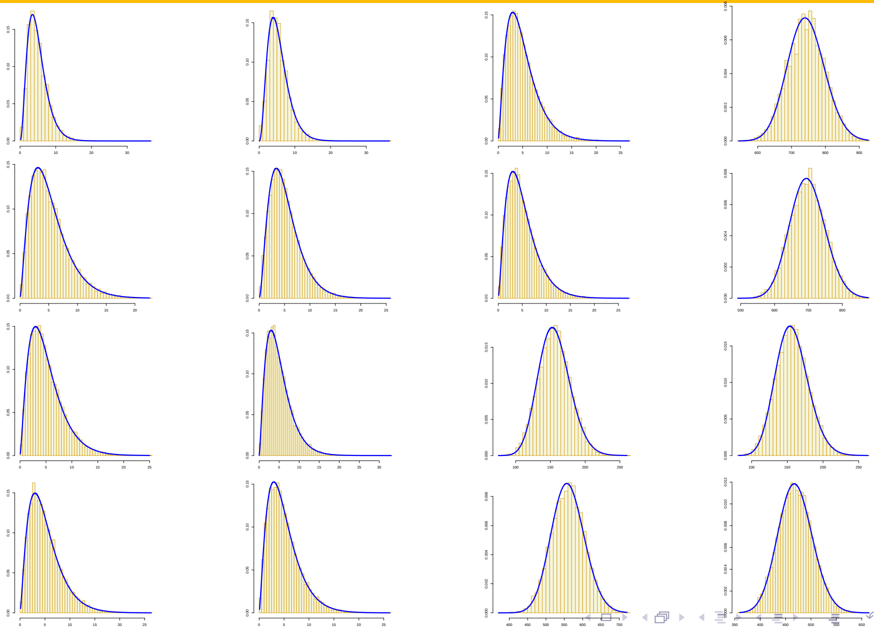




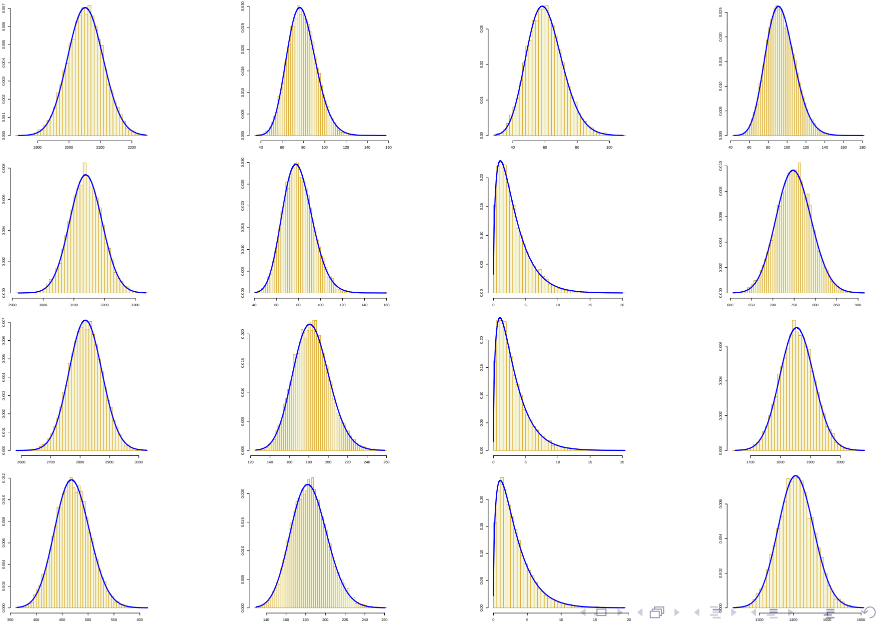
Performance of Gamma approximation



Performance of Gamma approximation



Performance of Gamma approximation



Simulation results

Simulation results

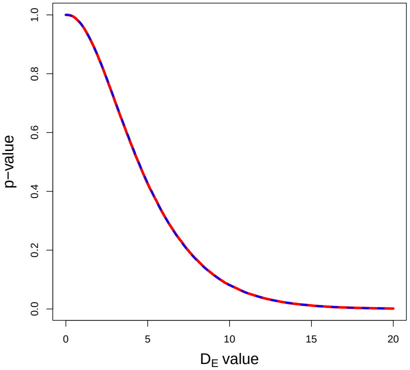
- Used simulation to test use of p-values to assess goodness-of-fit under H_0

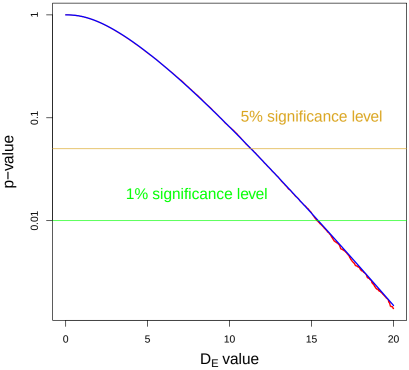
Simulation results

- Used simulation to test use of p-values to assess goodness-of-fit under H_0
- Non-sparse data scenario from capture-recapture:

Simulation results

- Used simulation to test use of p-values to assess goodness-of-fit under H_0
- Non-sparse data scenario from capture-recapture:



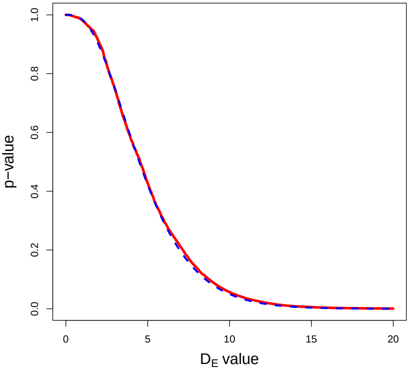


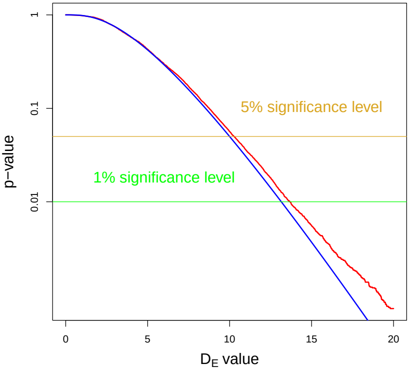
Simulation results

- Used simulation to test use of p-values to assess goodness-of-fit under H_0
- Sparse data scenario from capture-recapture:

Simulation results

- Used simulation to test use of p-values to assess goodness-of-fit under H_0
- Sparse data scenario from capture-recapture:





Summary

- Accurate approximation to the distribution of the deviance

Summary

- Accurate approximation to the distribution of the deviance
    - Whether or not we have sparse data

Summary

- Accurate approximation to the distribution of the deviance
    - Whether or not we have sparse data
    - Doesn't just apply to CR or Poisson models, is general

Future research

- Formalise theory underlying the Gamma approximation

Future research

- Formalise theory underlying the Gamma approximation
  - Formalise when $\text{Cov}(D_E, D_G - D_E) = 0$ for a model

Future research

- Formalise theory underlying the Gamma approximation
  - Formalise when $\text{Cov}(D_E, D_G - D_E) = 0$ for a model
  - Use statistical theory to justify: "why Gamma?"

Calculting p-values using the Gamma approximation

- To calculate p-values using the Gamma approximation:
  - Fit model to data, find deviance
  - Treat MLEs as true parameter values, find $E[D_G]$, $\text{Var}(D_G)$
  - Find $E[D_E]$, $\text{Var}(D_E)$
  - Fit Gamma curve, find p-value associated with model deviance