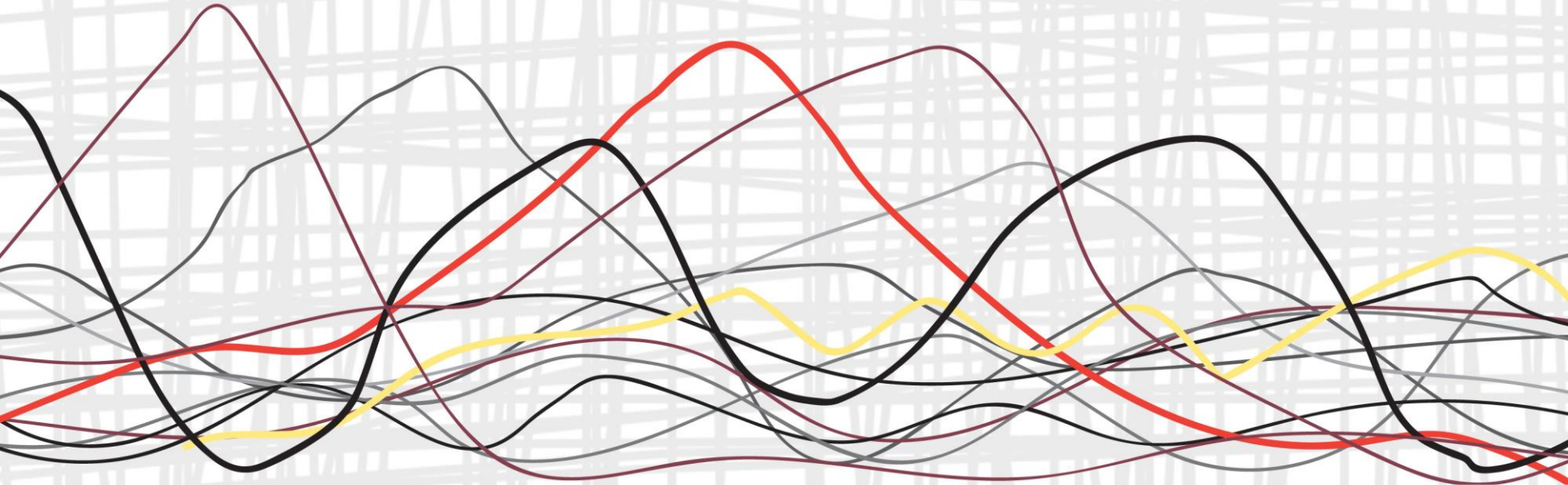




# Modeling Zero-Inflated Count Data

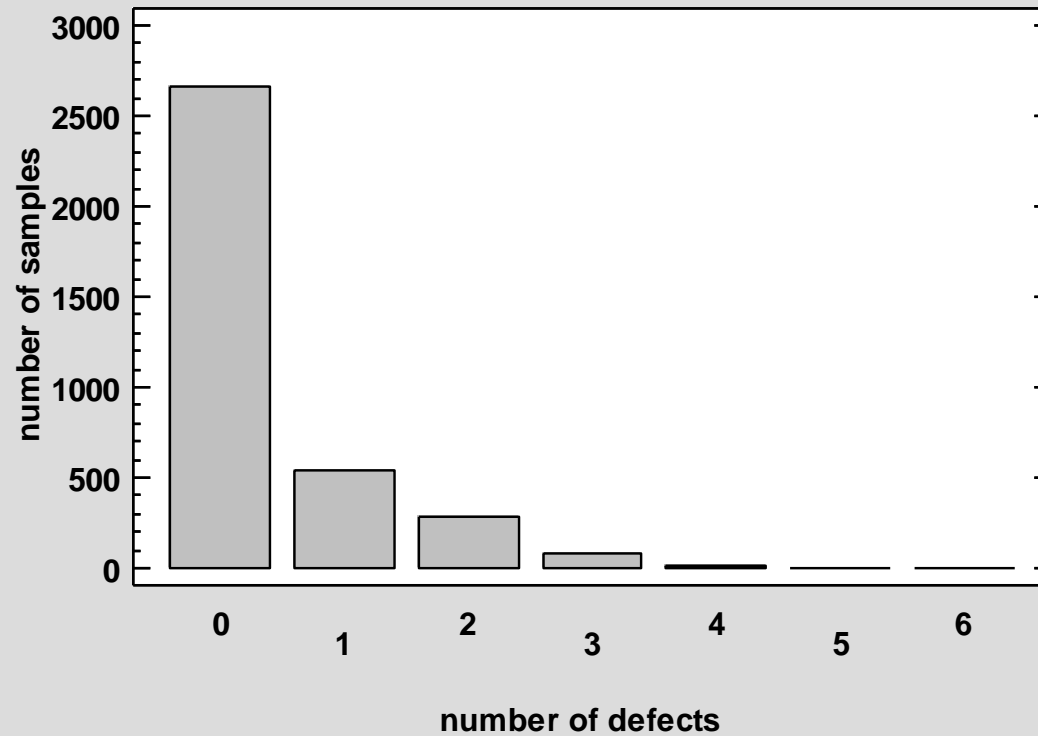


# Zero-Inflated Distributions

- Arise when count data have more zeroes than would be expected from the usual Poisson and negative binomial distributions
- Examples include:
  - Number of days a student is absent from school
  - Number of fish caught by visitors to a state park
  - Number of times a machine fails each month
  - Number of defects in samples from a production line
- Zeroes are often caused by two separate phenomena and are sometimes referred to as “true” zeroes and “excess” zeroes

# Sample Data (n=3600)

Barchart for Defects



# Zero-Inflated Poisson Distribution

$$p(0) = \alpha + (1 - \alpha)e^{-\lambda}$$

$$p(y) = (1 - \alpha) \frac{\lambda^y e^{-\lambda}}{y!} \quad \text{for } y > 0$$

$\lambda$  = conditional mean (excluding excess zeroes)

$\alpha$  = probability of excess zero

# Data Input Dialog Box

Distribution Fitting (Uncensored Data) ✕

Defects

Data:

[Select:]

Sort column names

# Analysis Options Dialog Box

Distribution Fitting Options

Distribution

<input type="checkbox"/> Bernoulli	<input type="checkbox"/> Exponential (2-parameter)	<input type="checkbox"/> Lognormal
<input type="checkbox"/> Binomial	<input type="checkbox"/> Exponential Power	<input type="checkbox"/> Lognormal (3-parameter)
<input type="checkbox"/> Discrete Uniform	<input type="checkbox"/> F (Variance Ratio)	<input type="checkbox"/> Maxwell (2-parameter)
<input type="checkbox"/> Geometric	<input type="checkbox"/> Folded Normal	<input type="checkbox"/> Noncentral Chi-Square
<input type="checkbox"/> Hypergeometric	<input type="checkbox"/> Gamma	<input type="checkbox"/> Noncentral F
<input type="checkbox"/> Negative Binomial	<input type="checkbox"/> Gamma (3-parameter)	<input type="checkbox"/> Noncentral t
<input checked="" type="checkbox"/> Poisson	<input type="checkbox"/> Generalized Gamma	<input type="checkbox"/> Normal
<input type="checkbox"/> Zero-Inflated Neg. Binomial	<input type="checkbox"/> Generalized Logistic	<input type="checkbox"/> Pareto
<input checked="" type="checkbox"/> Zero-Inflated Poisson	<input type="checkbox"/> Half Normal (2-parameter)	<input type="checkbox"/> Pareto (2-parameter)
<input type="checkbox"/> Beta	<input type="checkbox"/> Inverse Gaussian	<input type="checkbox"/> Rayleigh (2-parameter)
<input type="checkbox"/> Beta (4-parameter)	<input type="checkbox"/> Johnson	<input type="checkbox"/> Smallest Extreme Value
<input type="checkbox"/> Birnbaum-Saunders	<input type="checkbox"/> Laplace	<input type="checkbox"/> Student's t
<input type="checkbox"/> Cauchy	<input type="checkbox"/> Largest Extreme Value	<input type="checkbox"/> Triangular
<input type="checkbox"/> Chi-Square	<input type="checkbox"/> Logistic	<input type="checkbox"/> Uniform
<input type="checkbox"/> Erlang	<input type="checkbox"/> Loglogistic	<input type="checkbox"/> Weibull
<input type="checkbox"/> Exponential	<input type="checkbox"/> Loglogistic (3-parameter)	<input type="checkbox"/> Weibull (3-parameter)

Binomial Trials  
Sample Size n:  
100

Hypergeometric Trials  
Sample Size n:  
100  
 Estimate N  
 Specify N  
1000

Negative Binomial Trials  
 Estimate k  
 Specify k  
10

Extended Threshold Parameters  
 Estimate  
 Specify lower/upper  
0.0 1.0

OK  
Cancel  
Help  
Estimation

# Tables and Graphs

Tables and Graphs

**TABLES**

- Analysis Summary
- Tests for Normality
- Goodness-of-Fit Tests
- Tail Areas
- Critical Values
- Normal Tolerance Limits
- Distribution-Free Limits
- Comparison of Alternative Distributions

**GRAPHS**

- Density Trace
- Symmetry Plot
- Frequency Histogram
- Quantile Plot
- Quantile-Quantile Plot
- Distribution Functions 1
- Distribution Functions 2

OK

Cancel

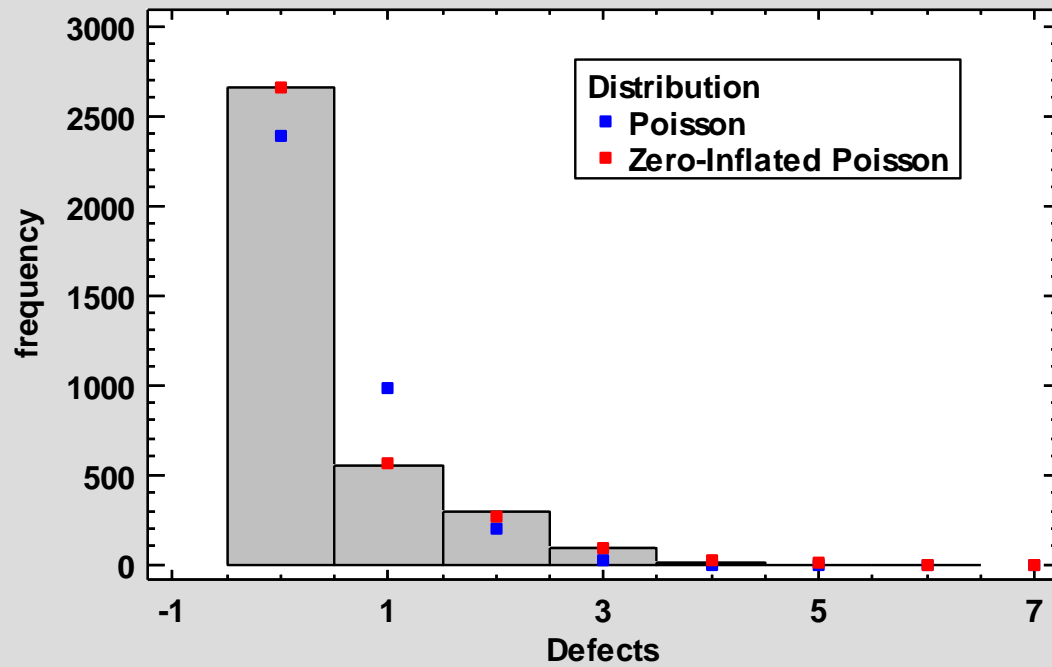
All

Store

Help

# Fitted Distributions

Histogram for Defects





# Analysis Summary

## Distribution Fitting (Uncensored Data) - Defects

Data variable: Defects

3600 values ranging from 0.0 to 6.0

Fitted Distributions

<b>Poisson</b>	<b>Zero-Inflated Poisson</b>
mean = 0.410833	conditional mean = 0.976894
	P(structural zero) = 0.57945

# Goodness-of-Fit Tests

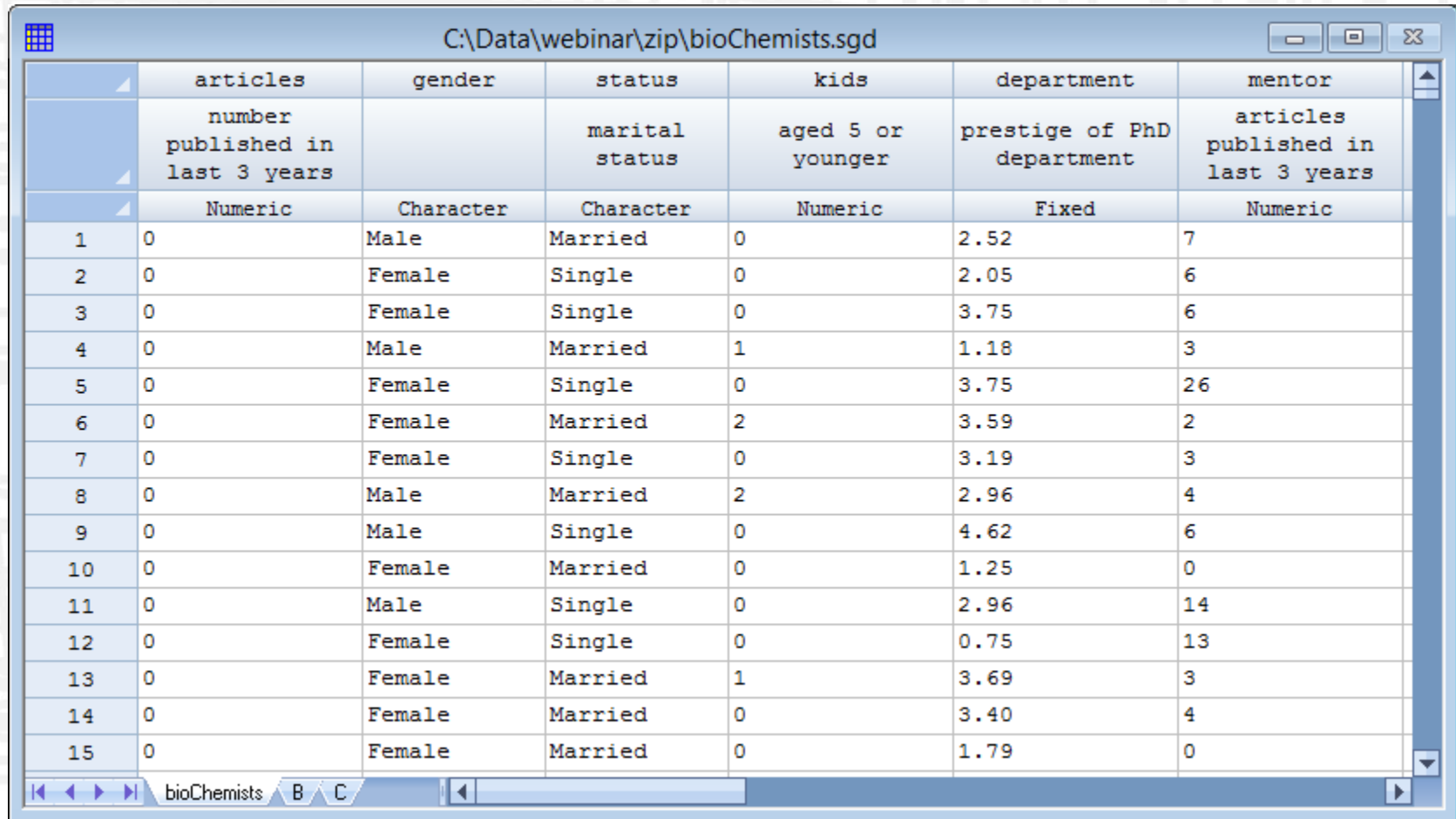
## Goodness-of-Fit Tests for Defects

### Chi-Square Test

	<i>Poisson</i>	<i>Zero-Inflated Poisson</i>
Chi-Square	496.55	3.16622
D.f.	3	3
P-Value	0.0	0.366697

# Zero-Inflated Count Regression

- Fits a regression model where the dependent variable has excess zeroes



	articles	gender	status	kids	department	mentor
	number published in last 3 years		marital status	aged 5 or younger	prestige of PhD department	articles published in last 3 years
	Numeric	Character	Character	Numeric	Fixed	Numeric
1	0	Male	Married	0	2.52	7
2	0	Female	Single	0	2.05	6
3	0	Female	Single	0	3.75	6
4	0	Male	Married	1	1.18	3
5	0	Female	Single	0	3.75	26
6	0	Female	Married	2	3.59	2
7	0	Female	Single	0	3.19	3
8	0	Male	Married	2	2.96	4
9	0	Male	Single	0	4.62	6
10	0	Female	Married	0	1.25	0
11	0	Male	Single	0	2.96	14
12	0	Female	Single	0	0.75	13
13	0	Female	Married	1	3.69	3
14	0	Female	Married	0	3.40	4
15	0	Female	Married	0	1.79	0

# Two Models for Zero-Inflated Regression

## 1. Zero-inflated model

- *Count component*: Poisson or negative binomial distribution to describe the distribution of counts ( $Y=0,1,2,\dots$ )
- *Zero component*: binomial distribution to describe the probability of excess zeroes

## 2. Hurdle model

- *Count component*: zero-truncated Poisson, negative binomial or geometric distribution to describe the distribution of positive counts ( $Y=1,2,3,\dots$ )
- *Zero component*: binomial or censored Poisson, negative binomial or geometric distribution to describe the probability of all zeroes

# ZIP Regression

- For the count component, a log-linear function links the conditional Poisson mean  $\lambda$  to the independent variables

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i}$$

- For the excess zeroes, a logit function links the binomial probability parameter  $\alpha$  to the independent variables

$$\alpha_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i})}$$

# Data Input Dialog Box

Zero Inflated Count Regression

articles  
gender  
status  
kids  
department  
mentor

Dependent Variable:  
▶ articles

Categorical Factors:  
▶ gender  
status

Quantitative Factors:  
▶ kids  
department  
mentor

[Weights:]  
▶

[Select:]  
▶

Sort column names

OK Cancel Delete Transform... Help

# Model Specification

Model Factors ×

Count component

gender  
status  
kids  
department  
mentor

Zero-inflation component

gender  
status  
kids  
department  
mentor

Include quadratic effects       Include 2-factor interactions       Include nested effects

# Analysis Summary

Zero Inflated Count Regression Options

Model

- Zero inflated
- Hurdle

Count distribution

- Poisson
- Negative binomial
- Geometric

Link function

- Logit
- Probit
- Complementary log-log
- Cauchit
- Log

Zero distribution

- Binomial
- Poisson
- Negative binomial
- Geometric

Optimization

- BFGS
- Nelder-Mead
- Congugate gradient

Maximum iterations:

Estimate starting values using EM

Diagnostics

- Compare to null model
- Compare to model without zero inflation

OK Cancel Help



# Tables and Graphs

Tables and Graphs ✕

TABLES	GRAPHS
<input checked="" type="checkbox"/> Analysis Summary	<input checked="" type="checkbox"/> Probability Distribution
<input type="checkbox"/> Probability Distribution	<input checked="" type="checkbox"/> Means Plot
<input checked="" type="checkbox"/> Predictions	<input checked="" type="checkbox"/> Count Component
<input type="checkbox"/> Unusual Residuals	<input type="checkbox"/> Zero Component
<input type="checkbox"/> R Script and Messages	<input type="checkbox"/> Observed versus Predicted
	<input type="checkbox"/> Residual Plots

OK  
Cancel  
All  
Store  
Help

# Analysis Summary

```
## Count model coefficients (poisson with log link):
##      Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.535427   0.112484   4.760 1.94e-06 ***
## genderMale   0.209144   0.063405   3.299 0.000972 ***
## statusSingle -0.103752   0.071111  -1.459 0.144563
## kids        -0.143320   0.047429  -3.022 0.002513 **
## department  -0.006160   0.031009  -0.199 0.842543
## mentor       0.018098   0.002294   7.888 3.08e-15 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##      Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.821111   0.480292  -1.710 0.08734 .
## genderMale  -0.109743   0.280082  -0.392 0.69519
## statusSingle 0.354033   0.317610   1.115 0.26499
## kids        0.217099   0.196481   1.105 0.26919
## department  0.001182   0.145270   0.008 0.99351
## mentor     -0.134103   0.045243  -2.964 0.00304 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Model Comparisons

```
# Comparison to model with only a constant
mnull <- update(model, . ~ 1)
pchisq(2 * (logLik(model) - logLik(mnull)), df = 10, lower.tail = FALSE)

## 'log Lik.' 5.351991e-27 (df=12)

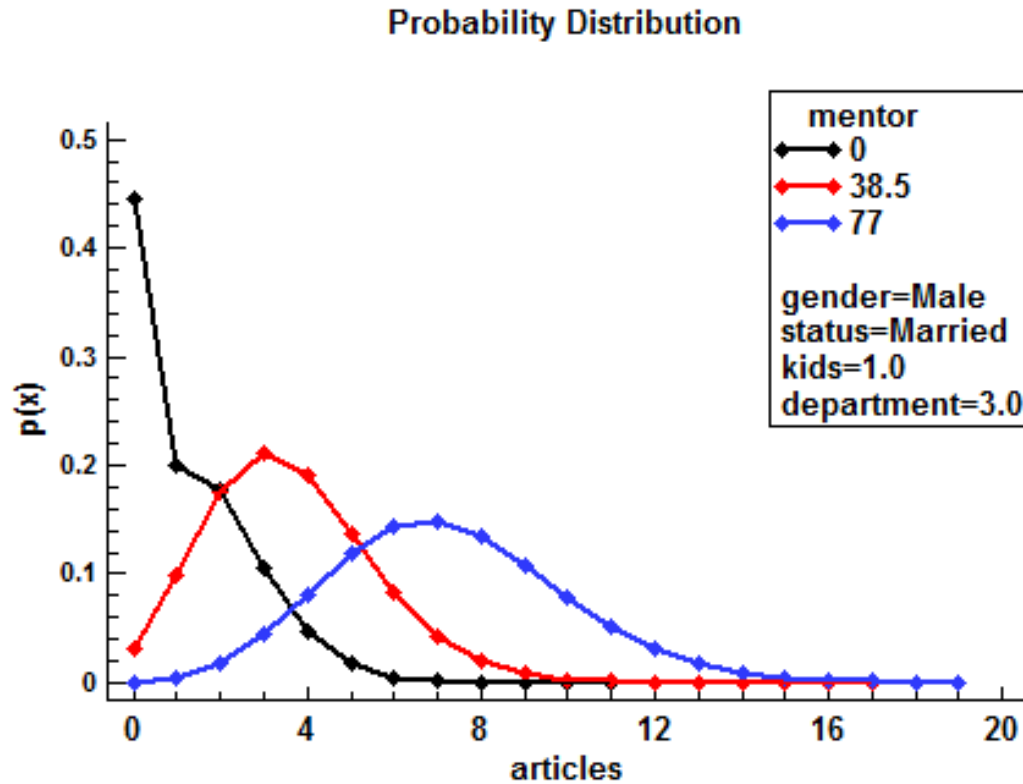
# Comparison to model without zero inflation
p1<-glm(articles~gender+status+kids+department+mentor,family="poisson",data=d)
vuong(p1,model)

## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
## null that the models are indistinguishable)
## -----
##              Vuong z-statistic              H_A      p-value
## Raw              -4.180476 model2 > model1 1.4545e-05
## AIC-corrected    -3.638531 model2 > model1 0.0001371
## BIC-corrected    -2.332734 model2 > model1 0.0098310
```

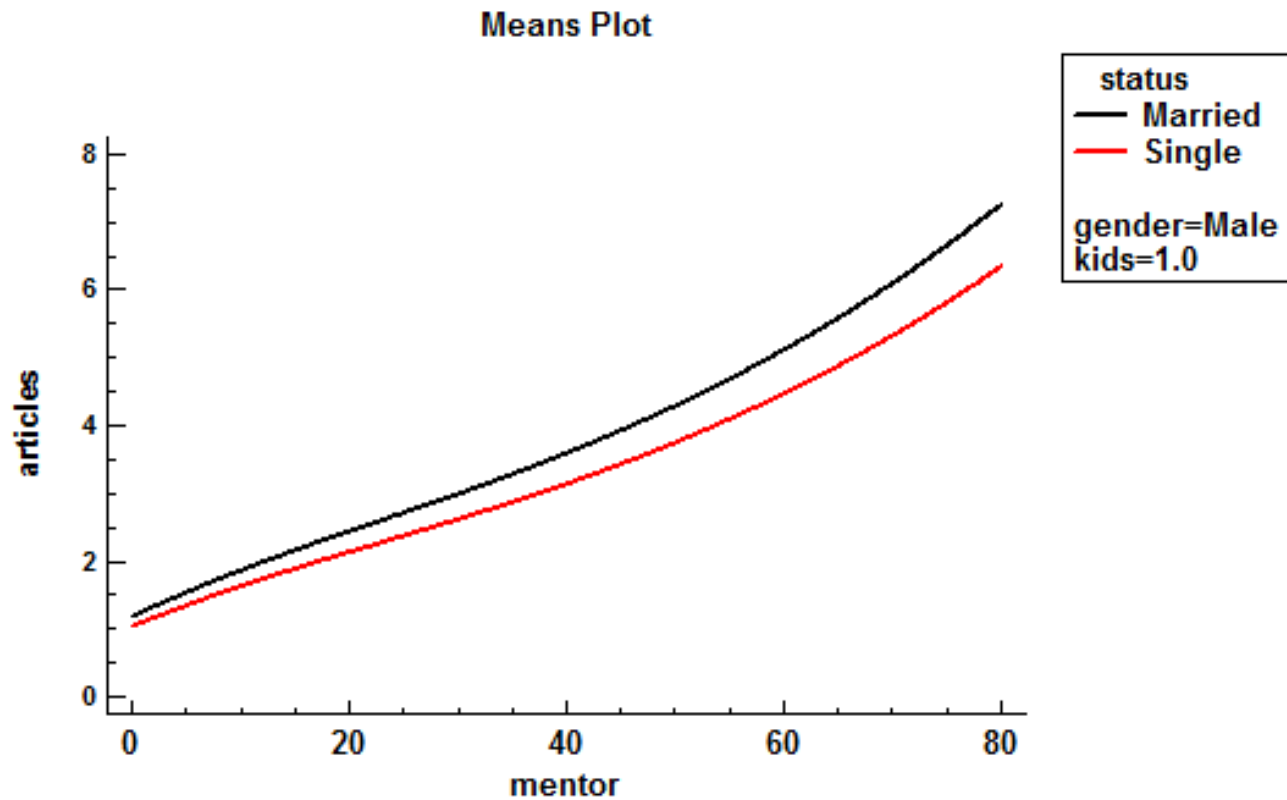
# Simplifying the Model

```
## Count model coefficients (poisson with log link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.52604    0.06139   8.569 < 2e-16 ***
## genderMale   0.21826    0.05878   3.713 0.000205 ***
## statusSingle -0.13483    0.06587  -2.047 0.040670 *
## kids        -0.16277    0.04337  -3.753 0.000175 ***
## mentor       0.01819    0.00221   8.227 < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.68569    0.20548  -3.337 0.000847 ***
## mentor      -0.13007    0.04023  -3.233 0.001224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

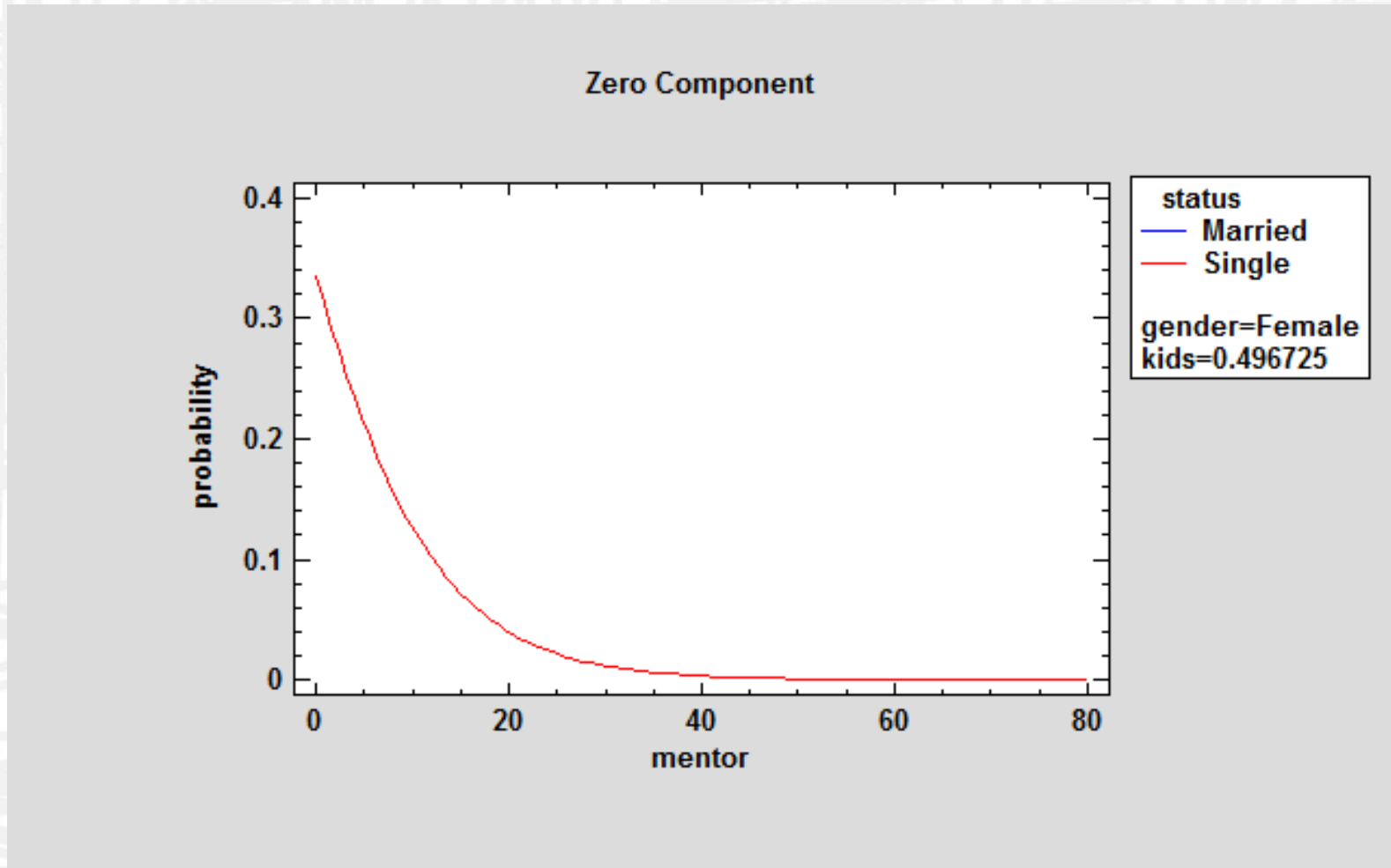
# Probability Distribution



# Means Plot



# Zero Component



# Predictions

## Predictions for articles

Row	Observed Value	Fitted Value	Residual	Pearson Residual	Count Component	Zero Component
916		1.10732			1.36291	0.187532

- Conditional mean:  $\hat{\lambda} = 1.36291$
- Prob. of excess zero:  $\hat{\alpha} = 0.187532$
- Unconditional mean:  $\hat{Y} = (1 - \hat{\alpha})\hat{\lambda} = 1.10732$



# References

- Long, J. Scott. 1990. The origins of sex differences in science. Social Forces. 68(3):1297-1316.
- R Package “MASS” (2016) <https://cran.r-project.org/web/packages/MASS/MASS.pdf>
- R Package “PSCL” (2017) <https://cran.r-project.org/web/packages/pscl/pscl.pdf>

# Slides, Data and Recorded Webinar

Posted at:

[www.statgraphics.com/webinars](http://www.statgraphics.com/webinars)

Also check our website for upcoming webinars.