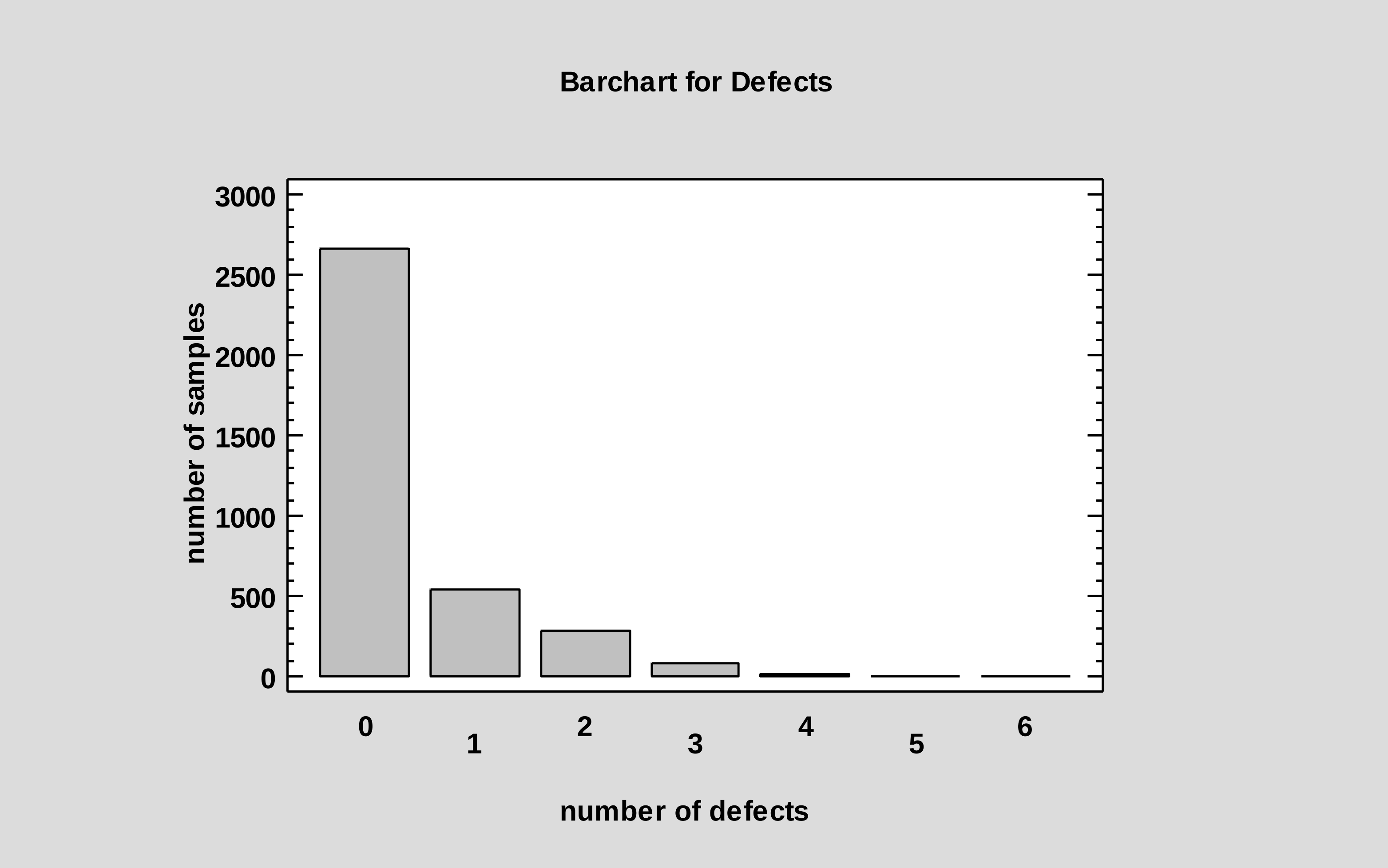
statgraphics 19®

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)



Zero-Inflated Poisson Distribution

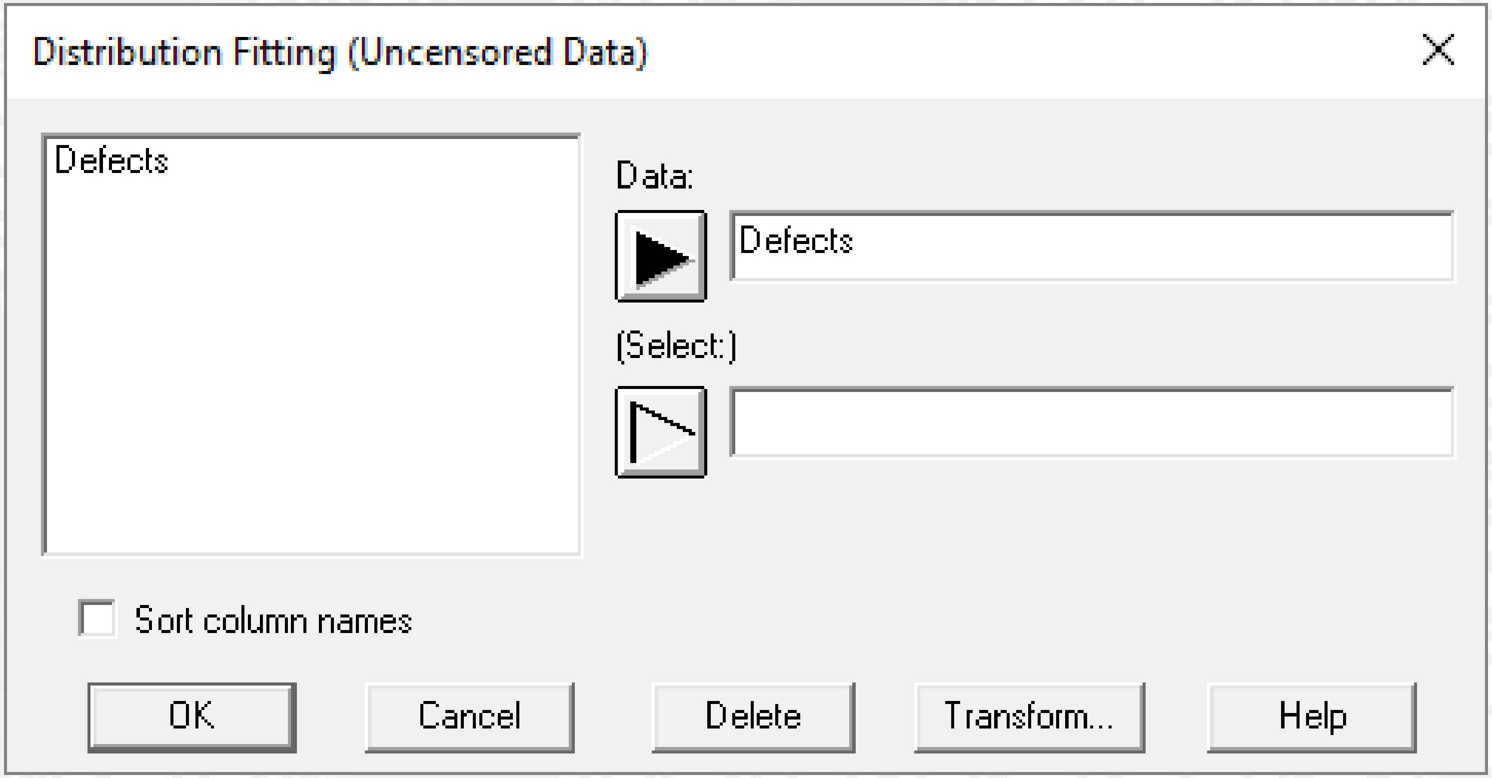
$$p(0) = \alpha + (1-\alpha)e^{-\lambda}$$

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

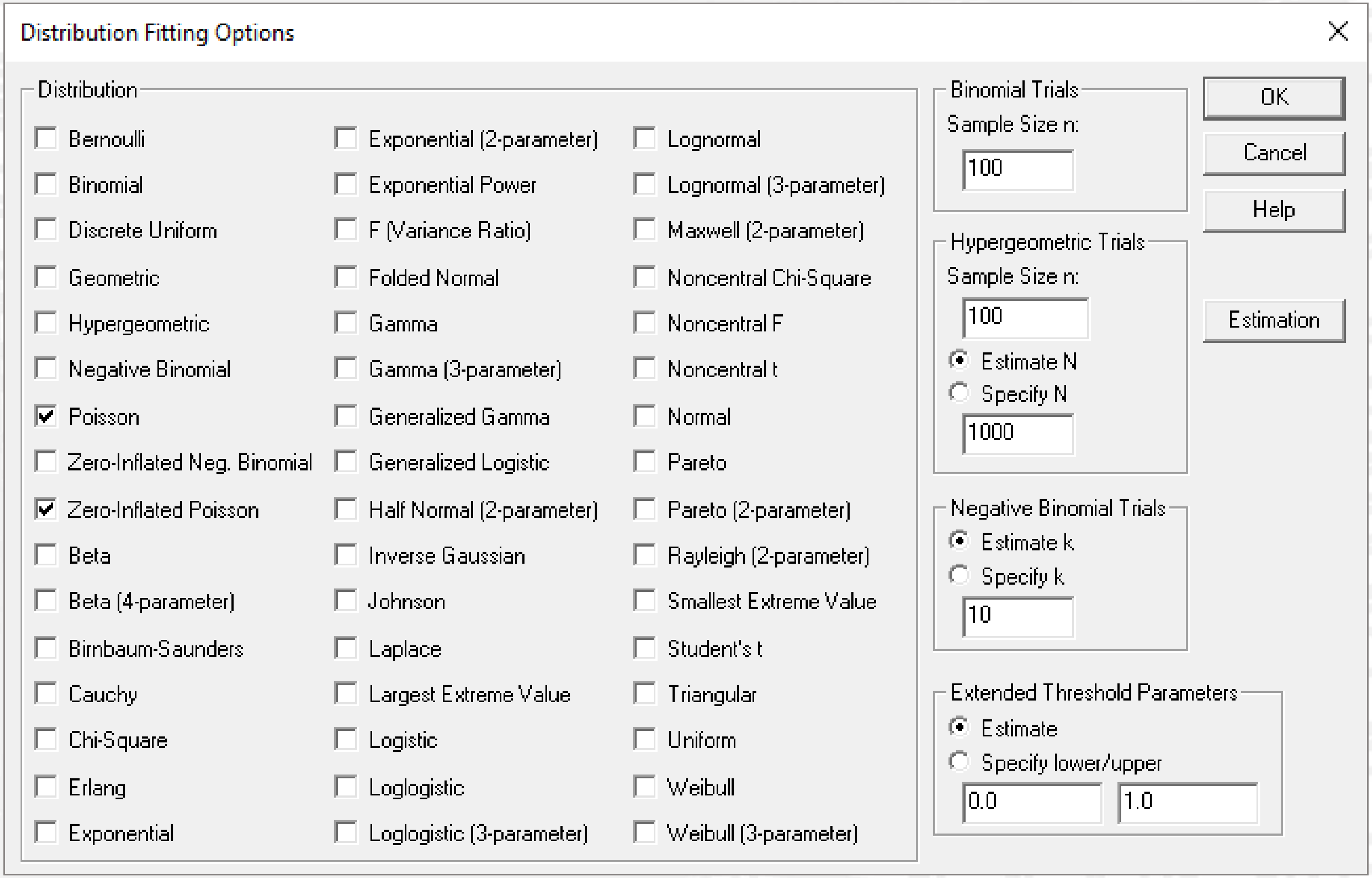
$\lambda = conditional\ mean\ (excluding\ excess\ zeroes)$

α = *probability of excess zero*

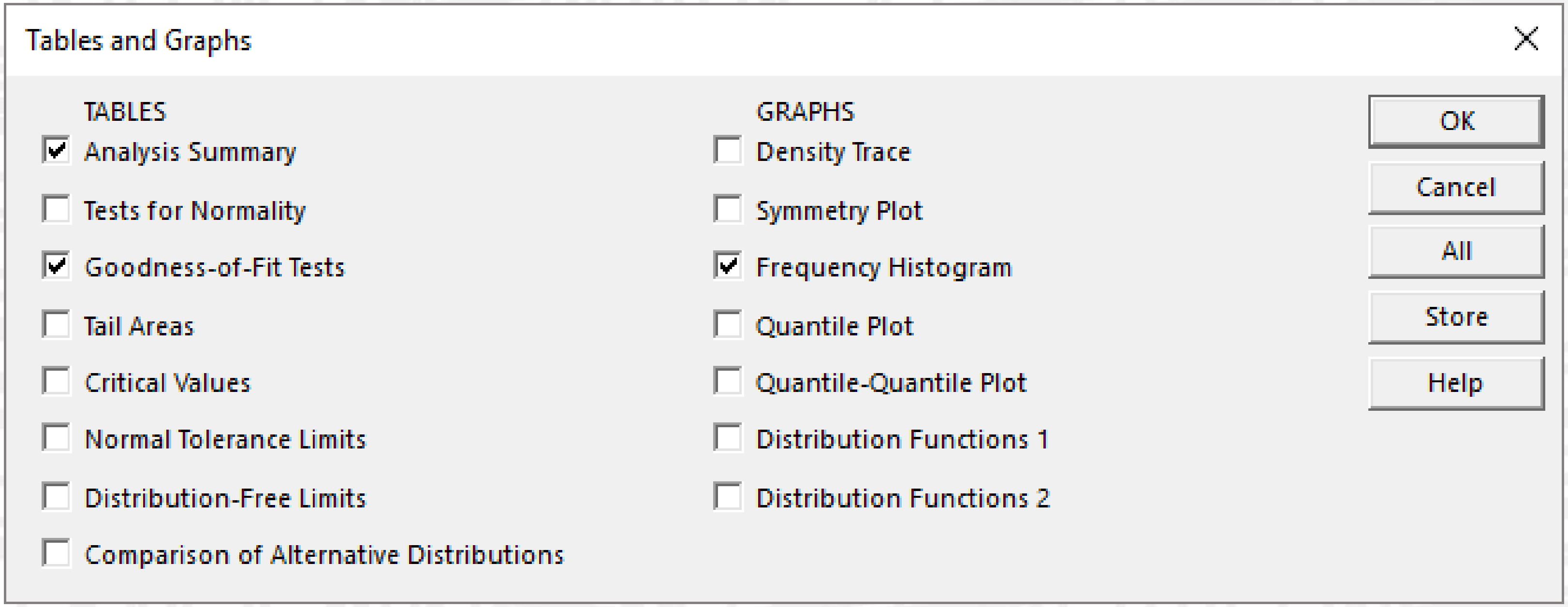
Data Input Dialog Box



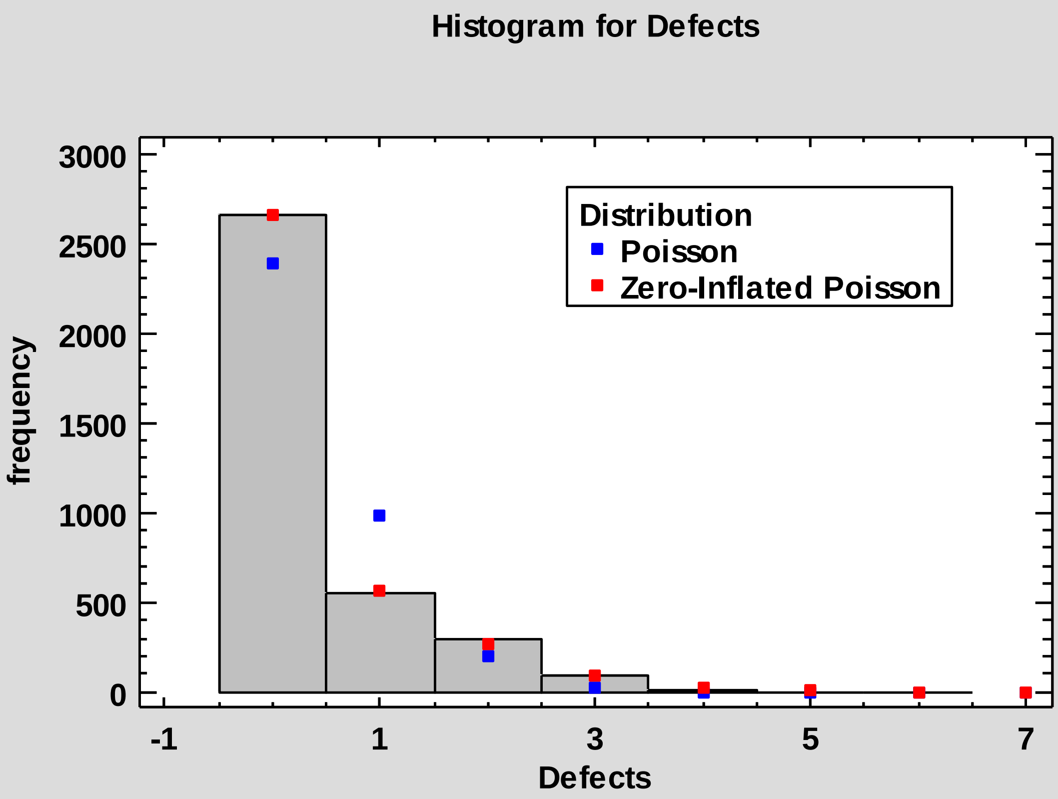
Analysis Options Dialog Box



Tables and Graphs



Fitted Distributions



Analysis Summary

Distribution Fitting (Uncensored Data) - Defects

Data variable: Defects

3600 values ranging from 0.0 to 6.0

Fitted Distributions

Poisson	***Zero-Inflated Poisson***
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

	Poisson	*Zero-Inflated Poisson*
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

C:\Data\webinar\zip\bioChemists.sgd

	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

bioChemists B C

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,…)
   - *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,…)
   - *Zero component*: binomial or censored Poisson, negative binomial or geometric distribution to describe the probability of all zeroes

ZIP Regression

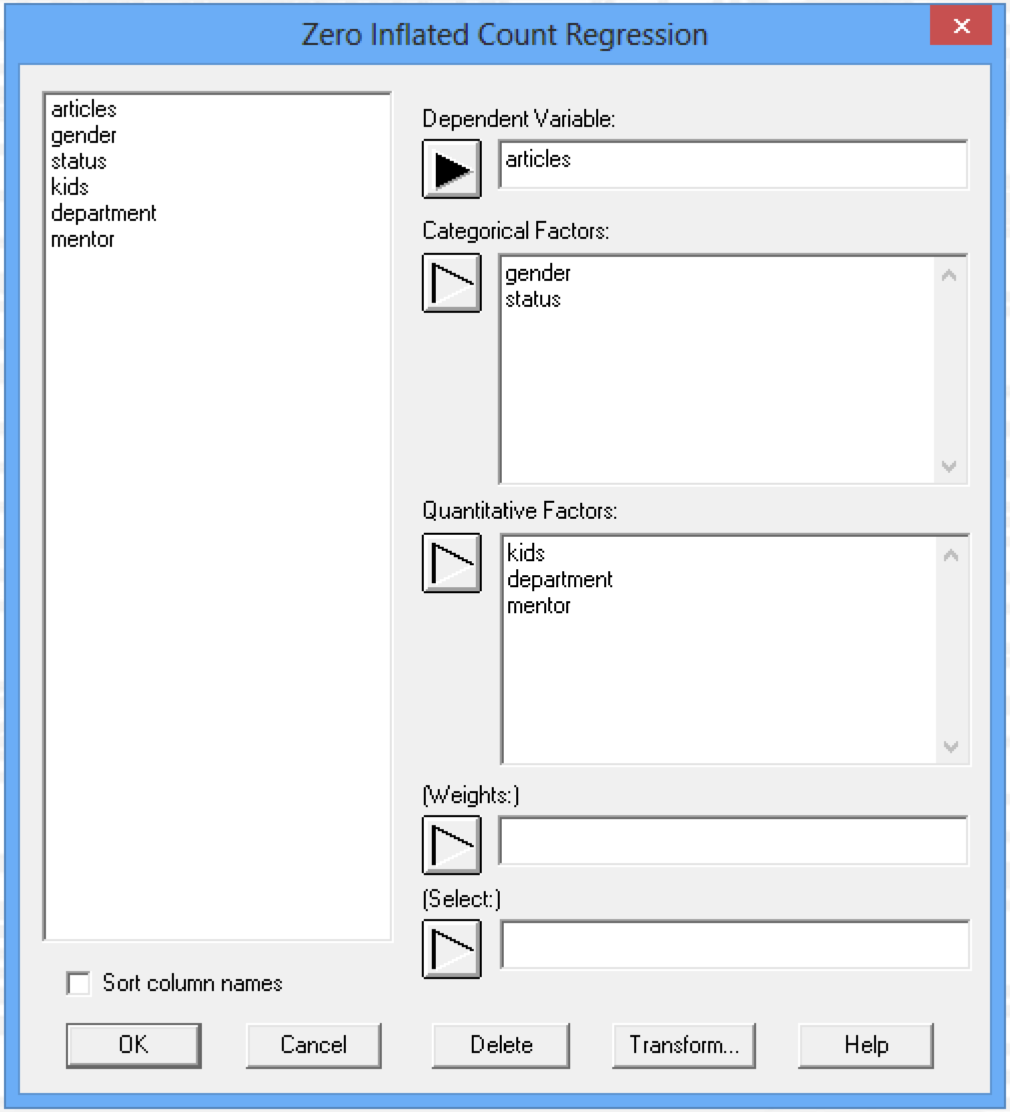
- For the count component, a log-linear function links the conditional Poisson mean λ to the independent variables

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

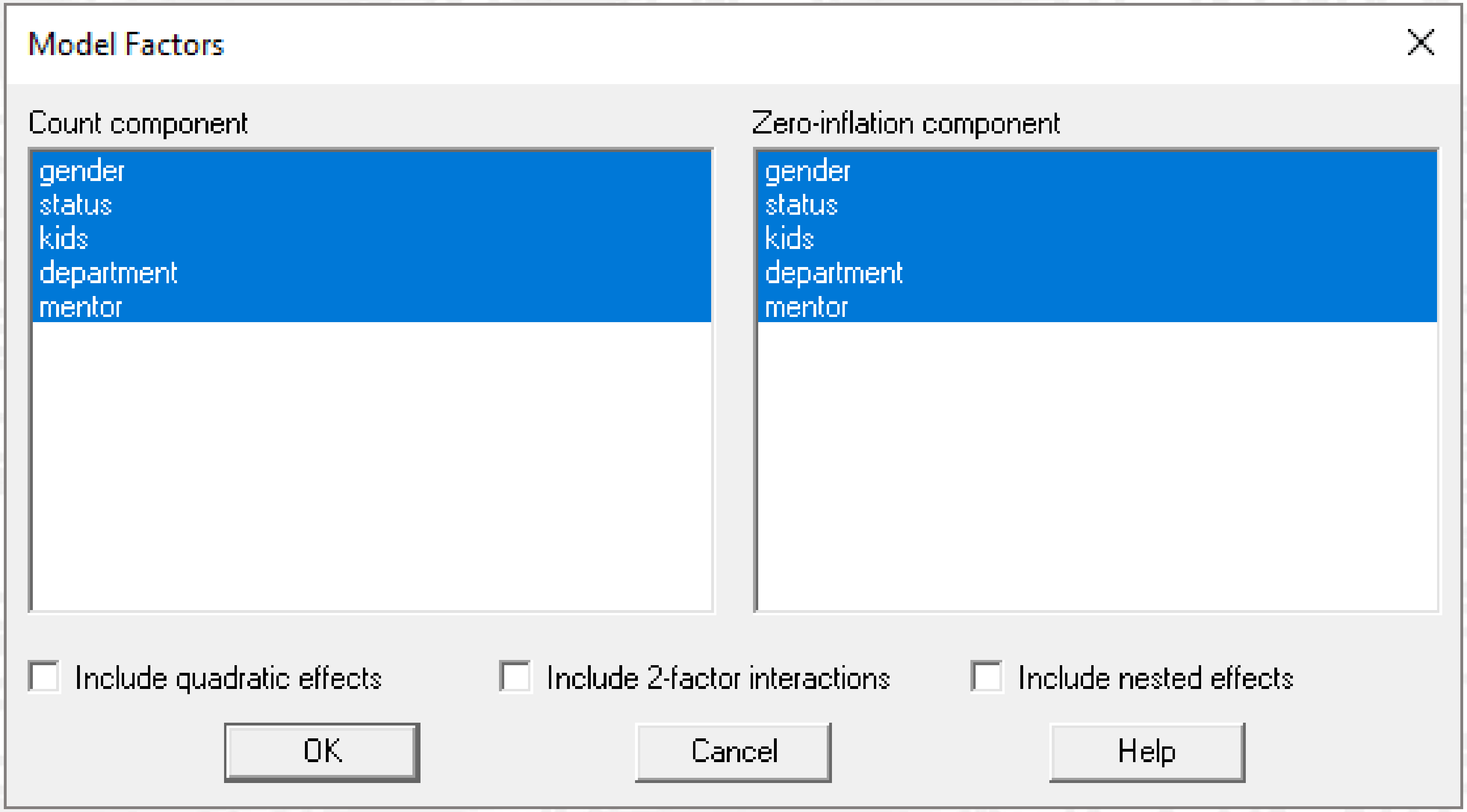
- For the excess zeroes, a logit function links the binomial probability parameter α to the independent variables

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

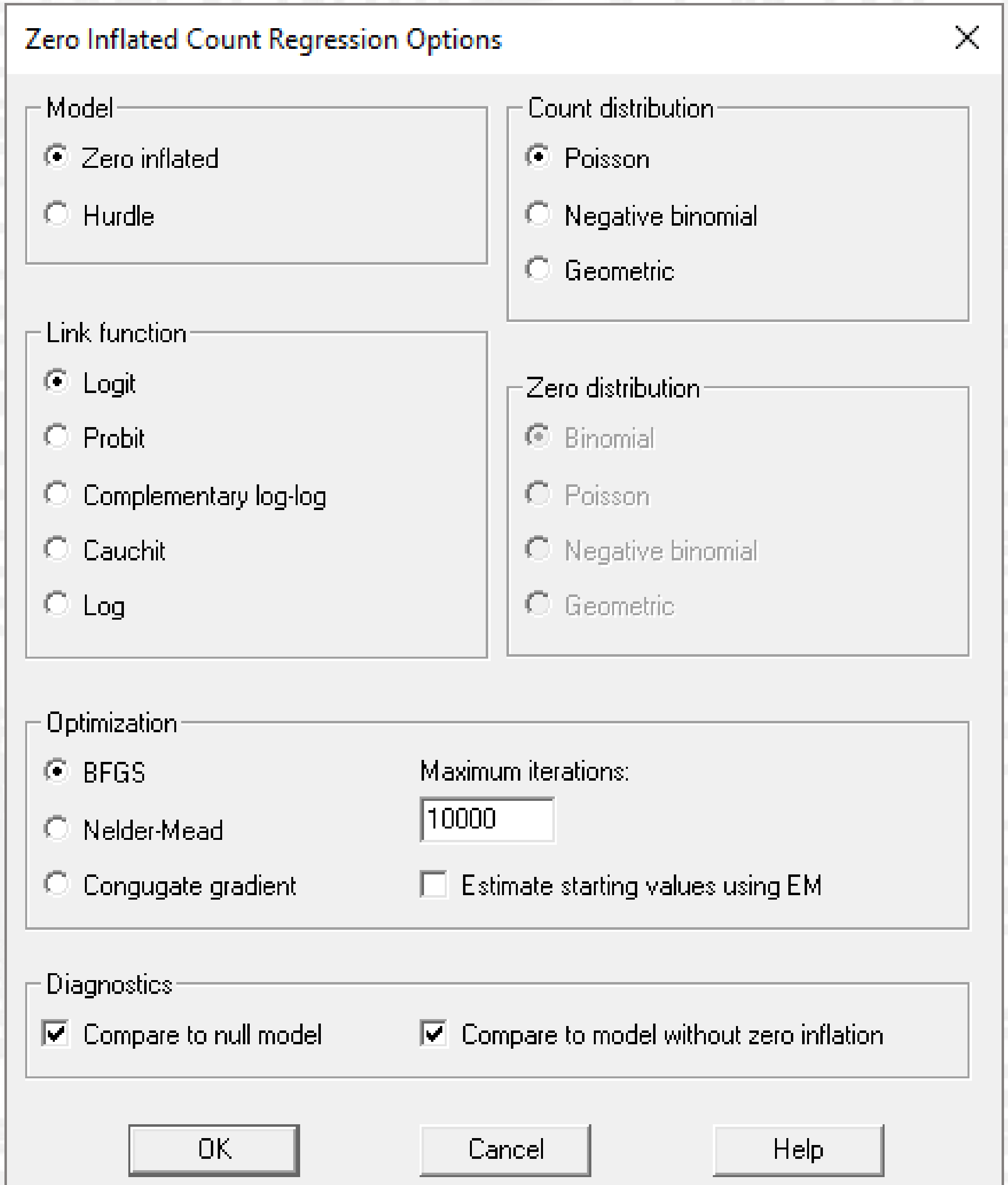
Data Input Dialog Box



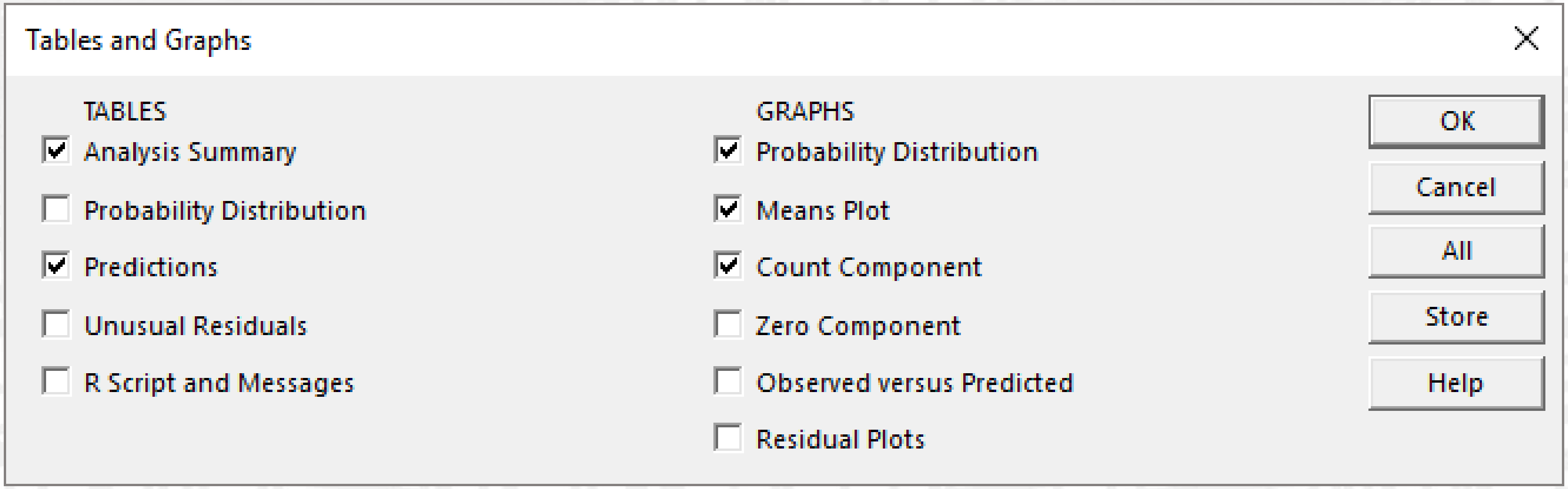
Model Specification



Analysis Summary



Tables and Graphs



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
```

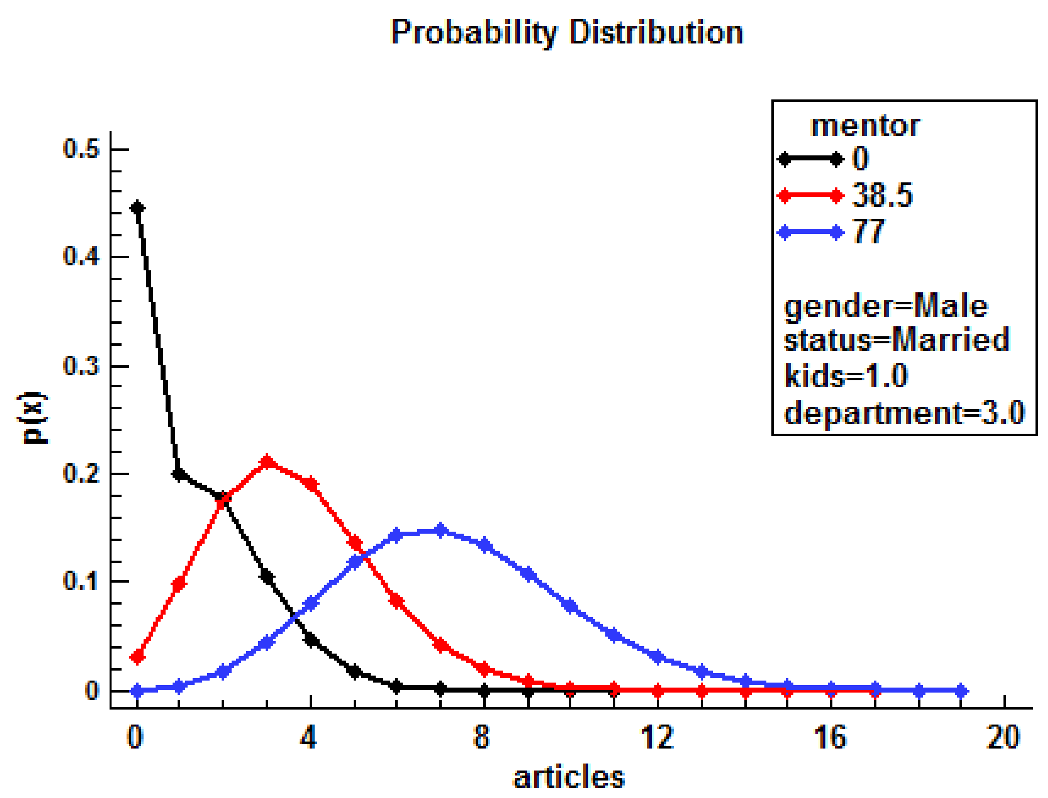
Model Comparisons

```
# 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 indistinguishible)
## -------------------------------------------------------------
##               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
```

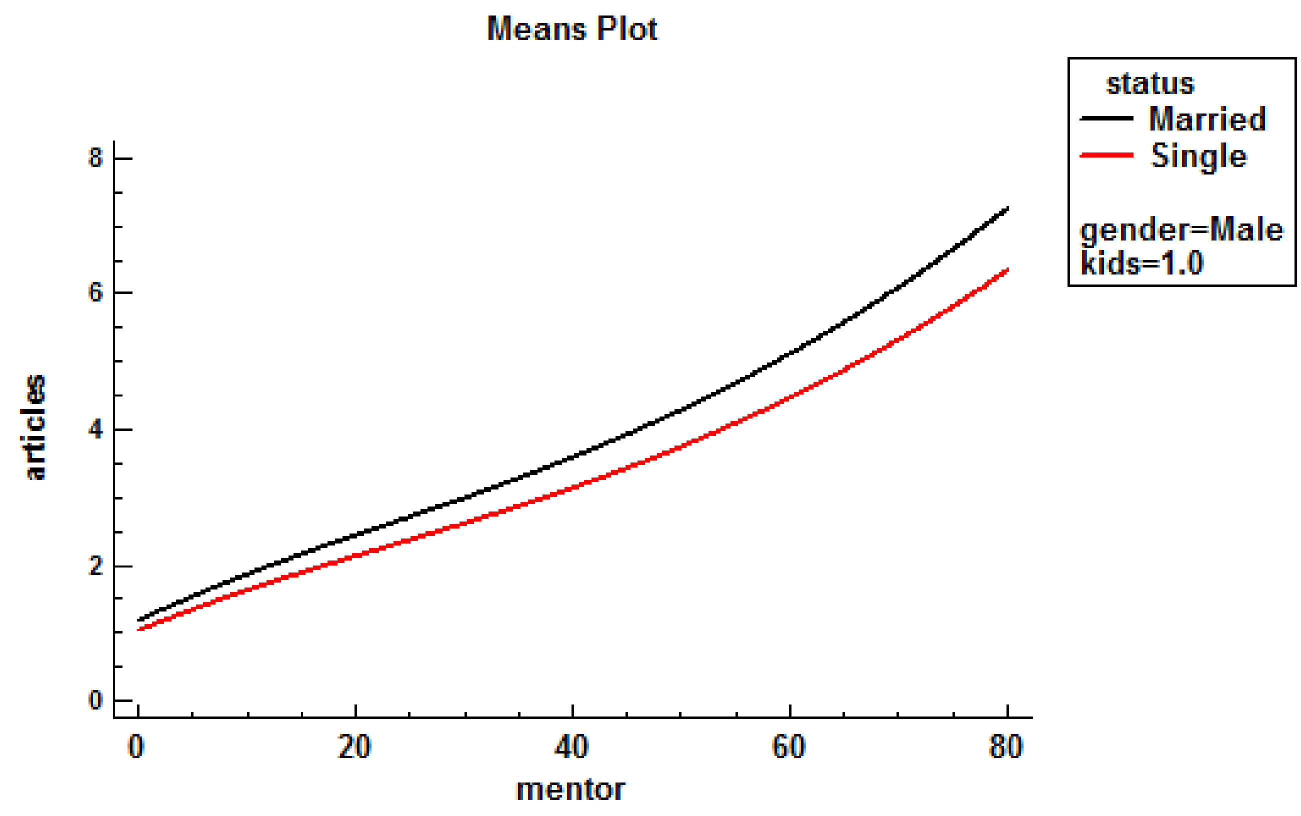
Simplifying the Model

```
## 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

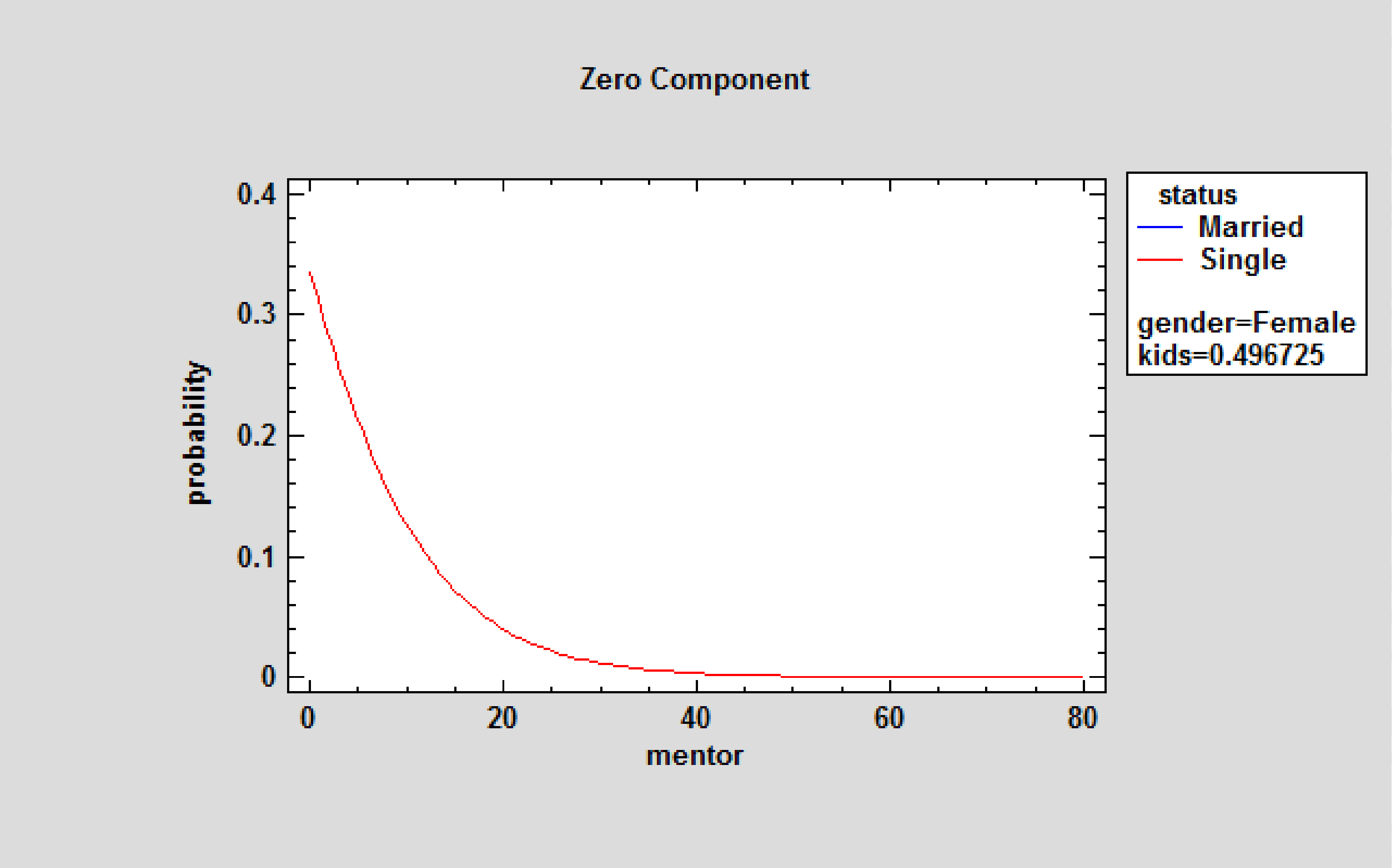


Means Plot



Means Plot
status
Married
Single
gender=Male
kids=1.0
articles
0
2
4
6
8
0
20
40
60
80
mentor

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

Also check our website for upcoming webinars.