Zero Inflated Count Regression

Revised: 7/23/2020

Summary

The **Zero Inflated Count Regression** procedure is designed to fit a regression model in which the dependent variable *Y* consists of counts. The fitted regression model relates *Y* to one or more predictor variables *X*, which may be either quantitative or categorical. It is similar to the procedures for *Poisson Regression* and *Negative Binomial Regression* except that it contains an additional component that represents the occurrence of more zeroes than would be expected in those models. Data containing excess zeroes is very common, including such diverse examples as the number of days a student is absent from school, the number of insurance claims within a population where not everyone has insurance, the number of defects in a manufactured item, and wild animal counts.

The calculations are performed by the “pscl” and “mass” packages in R. To run the procedure, R must be installed on your computer together with those packages. For information on downloading and installing R, refer to the document titled “R – Installation and Configuration”.

Sample StatFolio: *Zip.sgp*

Sample Data:

The file *bioChemists.sgd* contains data from a sample of 915 biochemistry graduate students analyzed by Long (1990). A portion of the data is shown below:

articles	*gender*	*status*	*kids*	*department*	*mentor*
0	Male	Married	0	2.52	7
0	Female	Single	0	2.05	6
0	Female	Single	0	3.75	6
0	Male	Married	1	1.18	3
0	Female	Single	0	3.75	26
0	Female	Married	2	3.59	2
0	Female	Single	0	3.19	3
0	Male	Married	2	2.96	4
0	Male	Single	0	4.62	6
0	Female	Married	0	1.25	0
…	…	…	…	…	…

The dependent variable is *articles*, which is the number of articles produced by each Ph.D. student during the last 3 years of their study. The other 5 columns are potential predictor variables including:

- *gender* – the gender of the student
- *status* – the student’s marital status
- *kids* – the number of children aged 5 or younger being raised by each student
- *department* – a measure of the prestige of the student’s Ph.D. department
- *mentor* – the number of articles produced by the student’s Ph.D. mentor during the last 3 years

Statistical Model

This procedure fits 2 basic types of zero-inflated count regression models: a zero-inflated regression model that adds an additional zero-generating component to the normal Poisson or negative binomial regression model, and a hurdle model which consists of separate components for generating zeroes and non-zeroes. The first type of model contains the following components:

1. Count component: a probability distribution such as the Poisson, negative binomial or geometric distribution that describes the generation of counts which may take the values Y=0, 1, 2, 3, …

2. Zero component: an additional binomial distribution that describes the occurrence of excess zeroes.

In contrast, the hurdle model defines the following components:

1. Count component: a zero-truncated probability distribution such as the Poisson, negative binomial or geometric distribution that describes the generation of counts which may take the values Y=1, 2, 3, … (but not zero).

2. Zero component: a second probability distribution such as the binomial or censored Poisson, negative binomial or geometric distribution that describes the generation of zeroes.

The basic difference between the models is that in the standard zero-inflated model, there are 2 sources of zeroes, while there is only one source of zeroes in the hurdle model.

Both the mean of the count component and the probability of 0 in the zero component are in general a function of the independent variables in the model. The relationship between the X’s and the parameters of those models is defined by *link functions*. The link function for the count component is always a log, while the link function for the zero component is usually a logit but may be a probit or other type of function.

A commonly used model for count distributions with excess zeroes is the zero-inflated Poisson process (ZIP). In such a model, the count component takes the form of a Poisson distribution

$$p(y_i) = \frac{\lambda_i e^{-\lambda_i}}{y_i!} \tag{1}$$

where λ_i is the Poisson rate parameter at the settings of the predictor variables corresponding to the *i-th* observation. A log-linear link function connects the rate parameter of the Poisson distribution to the independent variables according to

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

The model for the number of excess zeroes is binomial in which π_i represents the probability that the i-th observation will be an excess zero. Assuming a logit link function, this probability is related to the independent variables according to

$$\pi_i = \frac{1}{1+\exp\left(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i}\right)} \tag{3}$$

The combined probability distribution for the i-th observation is then

$$P(0) = \pi_i + (1-\pi_i)e^{-\lambda_i} \tag{4}$$

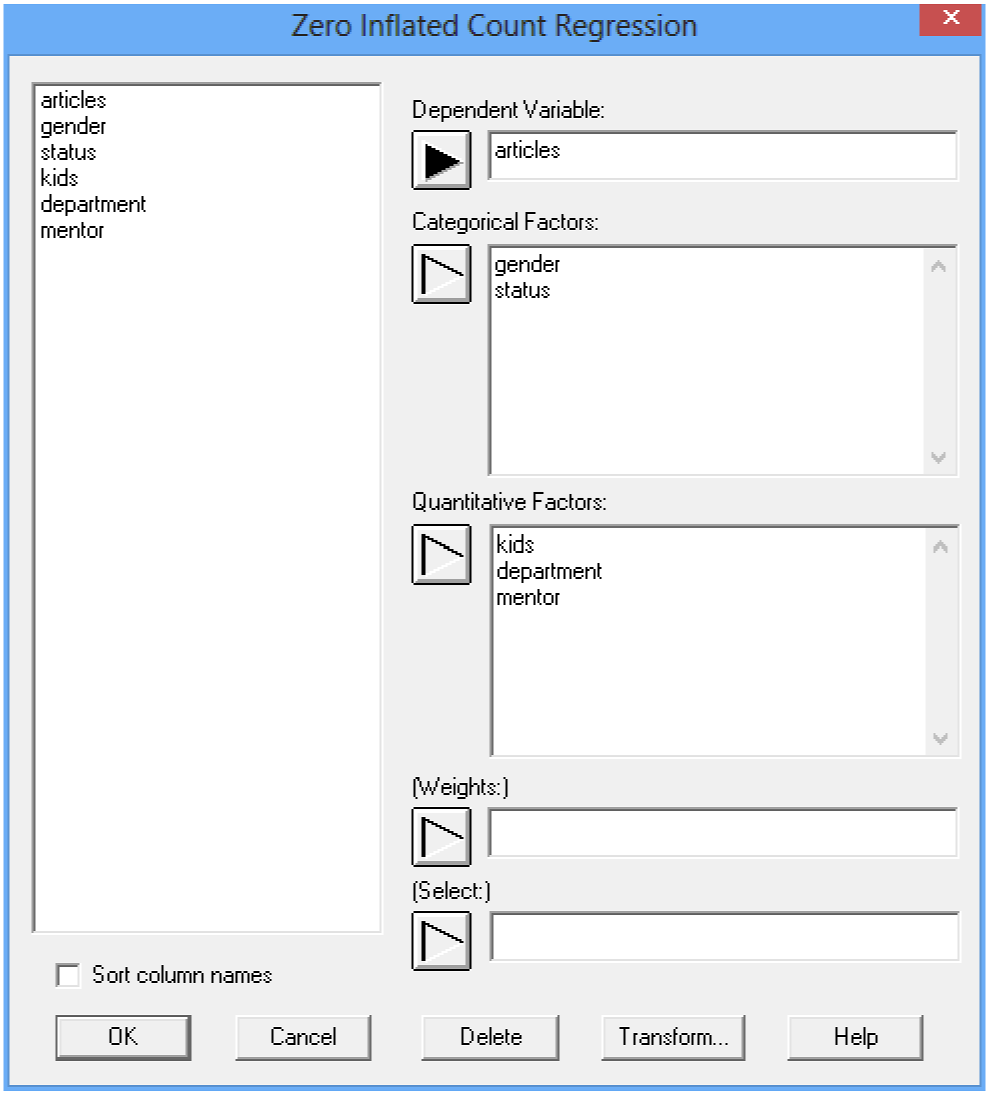
$$P(y_i) = (1-\pi_i)\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \tag{5}$$

The mean of the distribution is $(1-\pi_i)\lambda_i$.

In fitting the model to data, the same set of independent variables may be used in both the count and zero components, or different sets may be employed. As usual care should be taken to not overfit the model. Statgraphics provides tests to compare the fitted model against a model with no zero component.

Data Input

The data input dialog box requests information about the input variables:



articles
gender
status
kids
department
mentor

Dependent Variable: articles

Categorical Factors:
gender
status

Quantitative Factors:
kids
department
mentor

(Weights:)

(Select:)

□ Sort column names

- **Dependent Variable**: a numeric variable containing the *n* values of the dependent variable y_i. *Y* must consist of non-negative integer counts.

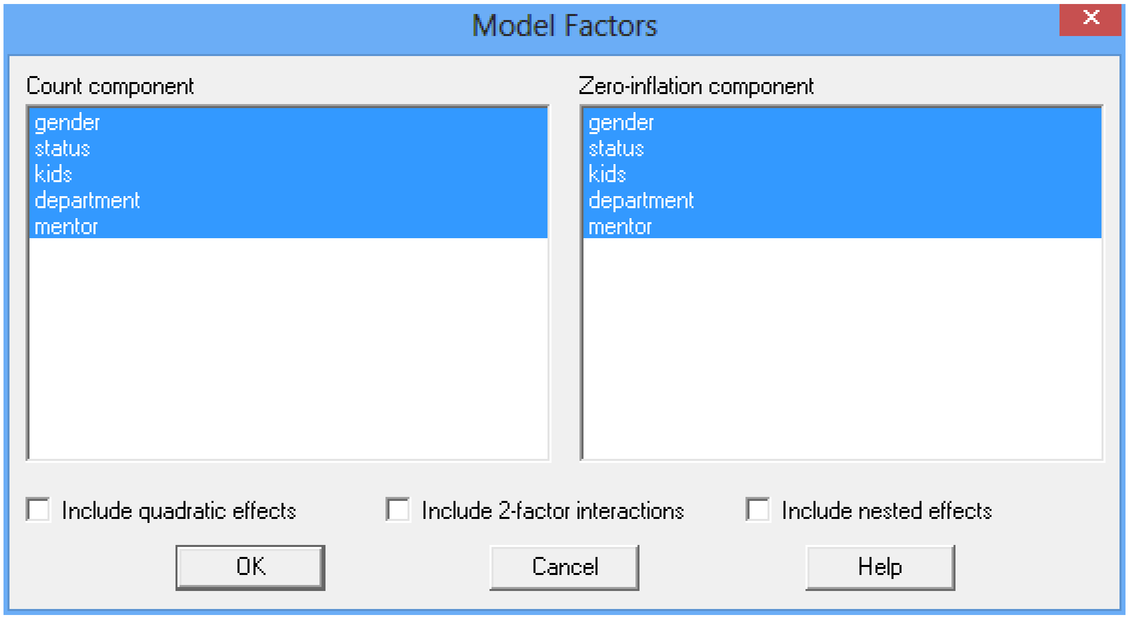
- **Categorical Factors**: numeric or non-numeric columns containing the levels of any categorical factors to be included in the model.

- **Quantitative Factors**: numeric columns containing the values of any quantitative factors to be included in the model.

- **Weights:** optional weights to be applied to each of the *n* observations.

- **Select**: optional subset selection.

After specifying the variables, a second dialog box is displayed:



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

Each specified factor may be included in both the count and zero-inflation components of the model (the default setting) or in only one of the components. To remove a factor from a component, click on its name to remove the highlighting.

In addition to the main effects of each factor, additional terms may be added to the model:

- **Include quadratic effects:** includes a second order term for each quantitative factor. For example, *kids^2* would indicate the values of *kids* squared.

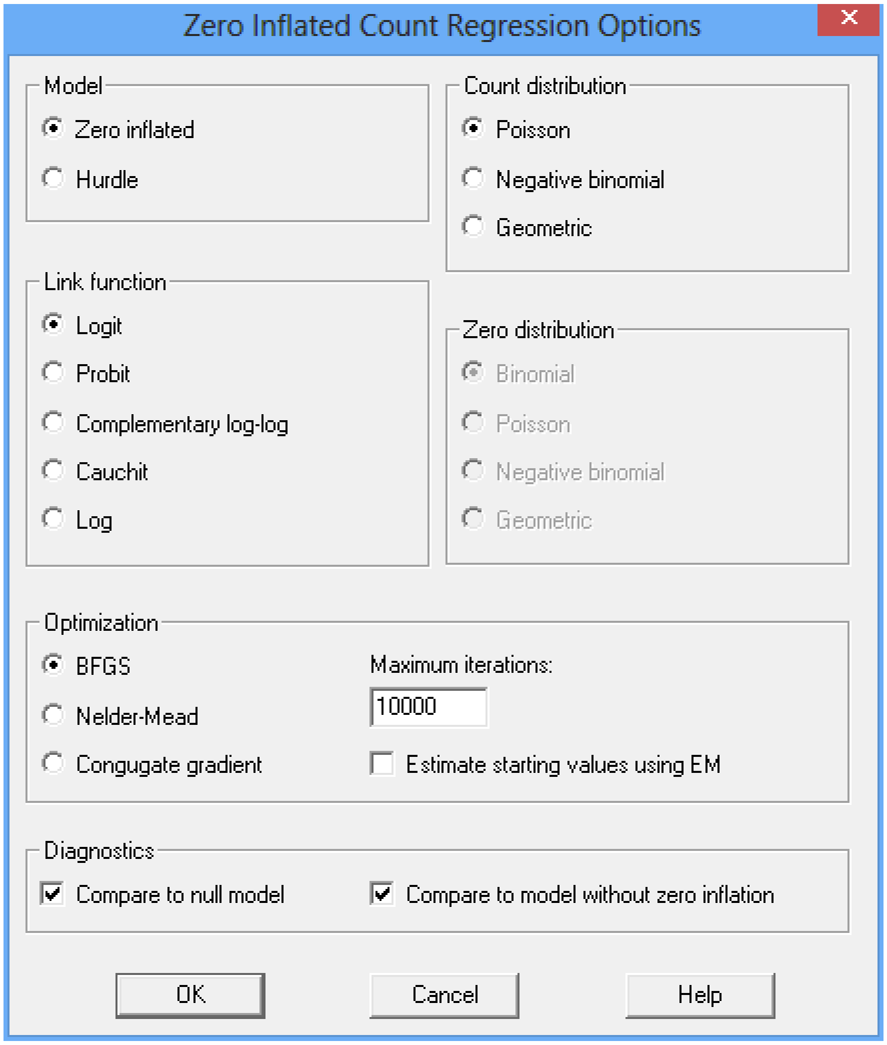
- **Include 2-factor interactions:** includes an interaction term for each pair of factors. For example, *gender*department* would indicate an interaction between *gender* and *department*.

- **Include nested effects:** indicates that one categorical factor is nested within levels of another categorical factor. For example, *status(gender)* would indicate that *status* is nested within levels of *gender*.

It should be noted that not all combinations of these effects may result in an estimable model. R will return an error message if the model cannot be estimated.

Analysis Options

The *Analysis Options* dialog box is used to select the type of model to be fit:



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: 10000
□ Estimate starting values using EM

Diagnostics
☑ Compare to null model
☑ Compare to model without zero inflation

- **Model:** selects between a standard zero-inflated model and a hurdle model. As described earlier, in the case of a hurdle model, only the zero component generates counts equal to 0.

- **Count distribution:** the assumed distribution for counts generated by the count component. The link function is always a log.

- **Zero distribution:** in the case of a hurdle model, the assumed distribution for counts generated by the zero-inflation component. For a standard zero-inflated model, the distribution is always binomial.

- **Link function:** a function that links the independent variables to the parameter of the zero-inflation distribution.

- **Optimization:** the method used to maximize the likelihood function during model estimation.

- **Maximum iterations:** the maximum iterations allowed during optimization of the likelihood function.

- **Estimate starting values using EM:** if checked, starting values for the estimation procedure will be selected using the Expectation Maximization procedure. Otherwise, they will be estimated by the GLM procedure.

- **Compare to null model:** if checked, a likelihood ratio test will be performed to test the significance of the fitted model compared to a model with only a constant.

- **Compare to model without zero inflation**: if checked, the Vuong non-nested test will be used to test the significance of the fitted model compared to a model without a component for excessive zeroes.

Analysis Summary

The *Analysis Summary* displays the output generated by R.

```
model=zeroinfl(articles~gender+status+kids+department+mentor|gender+status+kids+department+mentor,
control=zeroinfl.control(method="BFGS",maxit=10000,EM=FALSE),data=d,dist="poisson",link="logit")
summary(model)

##
## Call:
## zeroinfl(formula = articles ~ gender + status + kids + department +
##     mentor | gender + status + kids + department + mentor, data = d,
##     dist = "poisson", link = "logit", control = zeroinfl.control(method = "BFGS",
##         maxit = 10000, EM = FALSE))
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.3253 -0.8652 -0.2826  0.5404  7.2976
##
## 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
##
## Number of iterations in BFGS optimization: 21
## Log-likelihood: -1605 on 12 Df

# 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
```

The output includes:

- **Call:** the R command used to fit the model.

- **Pearson residuals:** a 5-number summary of the Pearson residuals. The Pearson residuals are a type of standardized residual which should follow approximately a standard normal distribution. Any values in excess of 3 in absolute value are potential outliers.

- **Count model coefficients:** estimates of the coefficients in the count component, z statistics, and approximate P values. Small P-values indicate coefficients that are statistically significantly different than zero. Symbols are used to indicate any coefficients that are significant at the 10%, 5%, 1% and 0.1% significance level.

- **Zero-inflation model coefficients:** estimates of the coefficients in the zero-inflation component, z statistics, and approximate P values.

- **Comparison to model with only a constant:** the results of a likelihood ratio test comparing the fitted model to a model without any independent variables. A small P-value indicates that the independent variables taken together are contributing significantly to the prediction of the dependent variable.

- **Comparison to model without zero inflation:** Vuong tests comparing the fitted model to one that does not account for excess zeroes. Three tests are performed: one with no correction for finite samples and two that penalize models based on the number of parameters they contain. Small P-values indicate that the fitted model is significantly better than a model of similar form that does not account for excessive zeroes.

For the sample data, the P-value for the overall model is 5.352e-27, indicating that the model is significantly better than one with no independent variables. In addition, all 3 of the P-values for the Vuong test are less than 0.01, indicating that the inclusion of a zero component results in a model that is significantly better than a model with only a count component at the 1% significance level.

The count model contains 3 significant predictors: *gender*, *kids* and *mentor*. The listing "genderMale" indicates a dummy variable which takes the value 1 for males and 0 for females. Since the coefficient is positive, males would tend to have larger values of *articles* than females. Note also that the coefficent on *kids* is negative, implying that more *kids* corresponds to fewer *articles*.

The only significant variable in the zero-inflation model is *mentor*. Since the coefficient for *mentor* is negative, it implies that a larger value of *mentor* corresponds to a smaller probability of excess zeroes.

Simplifying the Model

The model as fit is clearly over-parameterized. Before using it to make predictions, it is helpful to simplify it by removing insignificant variables. Taking a backward stepwise approach, variables may be removed one at a time, each time removing the one with the smallest P-value. Taking this approach, the model was simplified as follows:

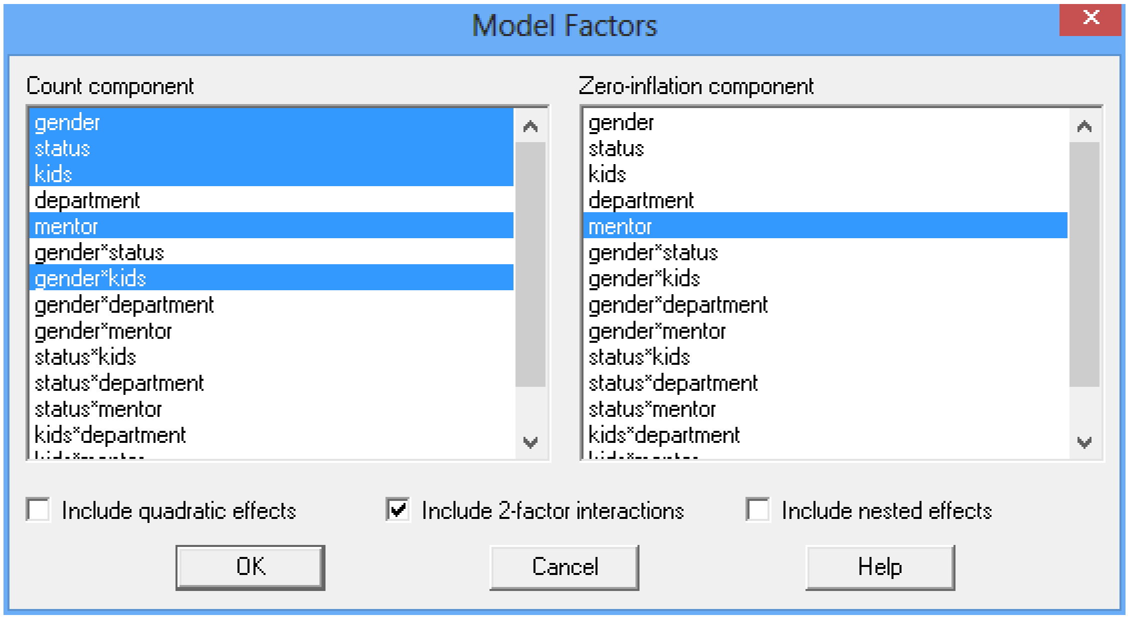
- Step 1: Based on its P-value of 0.994, *department* was removed from the zero-inflation model.
- Step 2: Based on its P-value of 0.827, *department* was removed from the count model.
- Step 3: Based on its P-value of 0.699, *gender* was removed from the zero-inflation model.
- Step 4: Based on its P-value of 0.287, *kids* was removed from the zero-inflation model.
- Step 5: Based on its P-value of 0.385, *status* was removed from the zero-inflation model.

After Step 5, The P-values for all remaining coefficients were less than 0.05. The *Analysis Summary* is shown below:

```
model=zeroinfl(articles~gender+status+kids+mentor|mentor,control=zeroinfl.control(method="BFGS",
maxit=10000,EM=FALSE),data=d,dist="poisson",link="logit")
summary(model)
##
## Call:
## zeroinfl(formula = articles ~ gender + status + kids + mentor |
##     mentor, data = d, dist = "poisson", link = "logit", control = zeroinfl.control
(method = "BFGS", maxit = 10000, EM = FALSE))
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.3222 -0.8700 -0.2588  0.5475  7.1875
##
## 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
##
## Number of iterations in BFGS optimization: 14
## Log-likelihood: -1606 on 7 Df

# Comparison to model with only a constant
mnull <- update(model, . ~ 1)
pchisq(2 * (logLik(model) - logLik(mnull)), df = 10, lower.tail = FALSE)## 'log Lik.' 1.36091e-
26 (df=7)

# 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                       -4.115846 model2 > model1 1.9288e-05
## AIC-corrected             -4.024983 model2 > model1 2.8490e-05
## BIC-corrected             -3.806052 model2 > model1 7.0601e-05
```

It will be noticed that the value of the log-likelihood function is almost identical to that of the initial model.

At this point, it is useful to consider adding interaction terms to the count model, such as an interaction between *gender* and *kids* (thinking that the effect of having many kids might be different for males and females). Such a model is specified as shown below:



Count component

[✓ gender]
[✓ status]
[✓ kids]
department
[✓ mentor]
gender*status
[✓ gender*kids]
gender*department
gender*mentor
status*kids
status*department
status*mentor
kids*department
kids*mentor

Zero-inflation component

gender
status
kids
department
[✓ mentor]
gender*status
gender*kids
gender*department
gender*mentor
status*kids
status*department
status*mentor
kids*department
kids*mentor

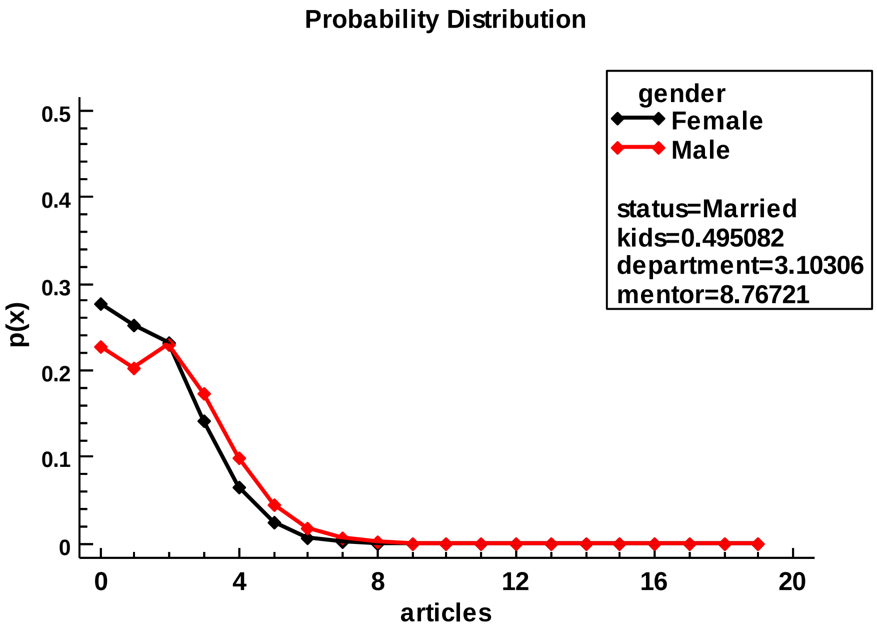
☐ Include quadratic effects ☑ Include 2-factor interactions ☐ Include nested effects

As may be seen below, such an interaction is not significant:

```
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.529784   0.064119   8.263  < 2e-16 ***
## genderMale        0.212201   0.066087   3.211  0.00132 **
## statusSingle     -0.136010   0.066140  -2.056  0.03974 *
## kids             -0.177786   0.086675  -2.051  0.04025 *
## mentor            0.018206   0.002213   8.228  < 2e-16 ***
## genderMale:kids   0.019055   0.095043   0.200  0.84110
```

Probability Distribution (Graph)

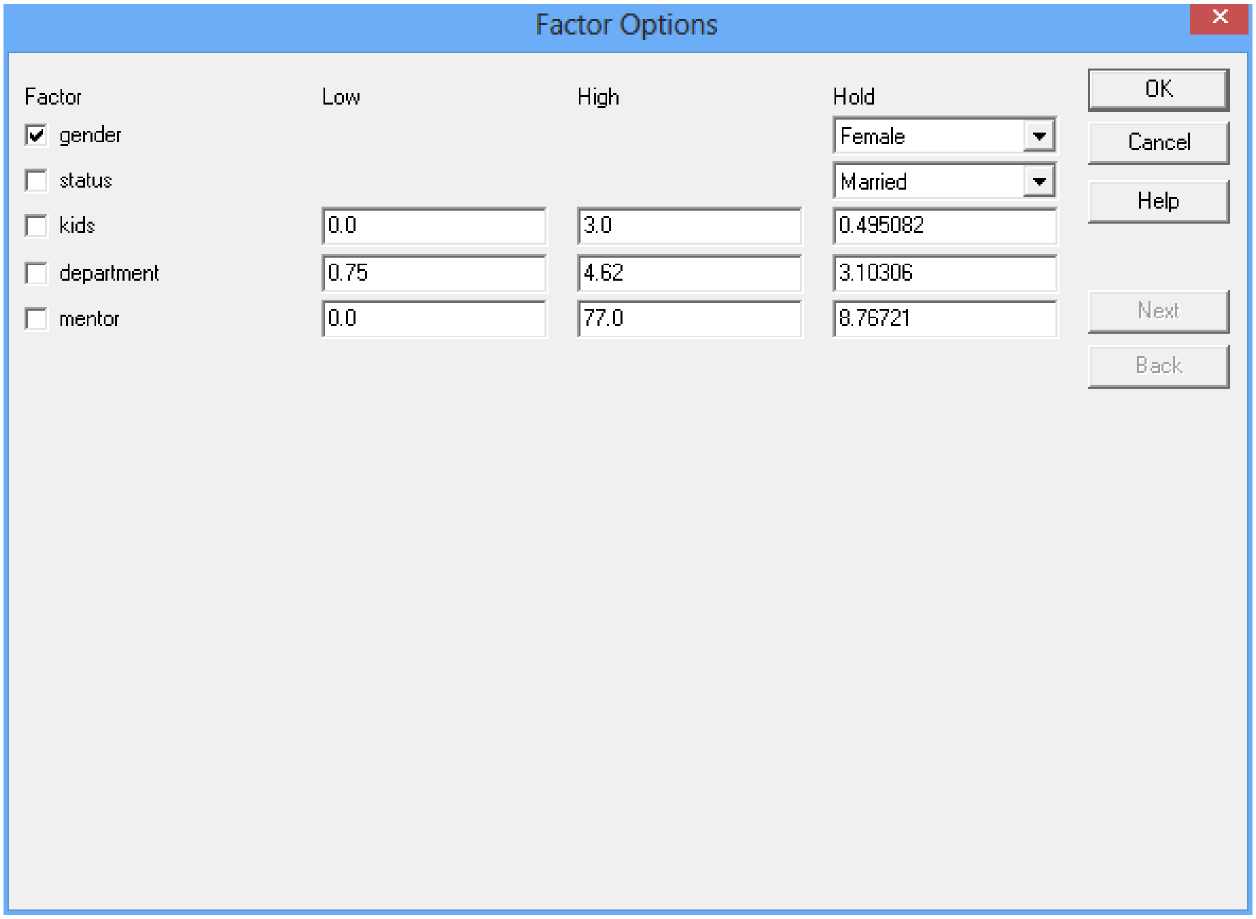
The probability distribution for the dependent variable may be plotted versus any single predictor variable, with the other variables held constant. For example, the plot below shows the estimated probability distribution of articles for females and males, with the other values held constant at the indicated levels:



The probability distribution for males has a higher mean than the distribution for females. You can also see the extra zeroes, particularly for the males since a Poisson distribution normally goes down monitonically on either side of its peak.

Pane Options

By default, the values at which the non-plotted variables are held are set equal to their respective means. This may be changed using *Pane Options*, as may the variable whose levels are changed.



Factor	Low	High	Hold
☑ gender			Female
□ status			Married
□ kids	0.0	3.0	0.495082
□ department	0.75	4.62	3.10306
□ mentor	0.0	77.0	8.76721

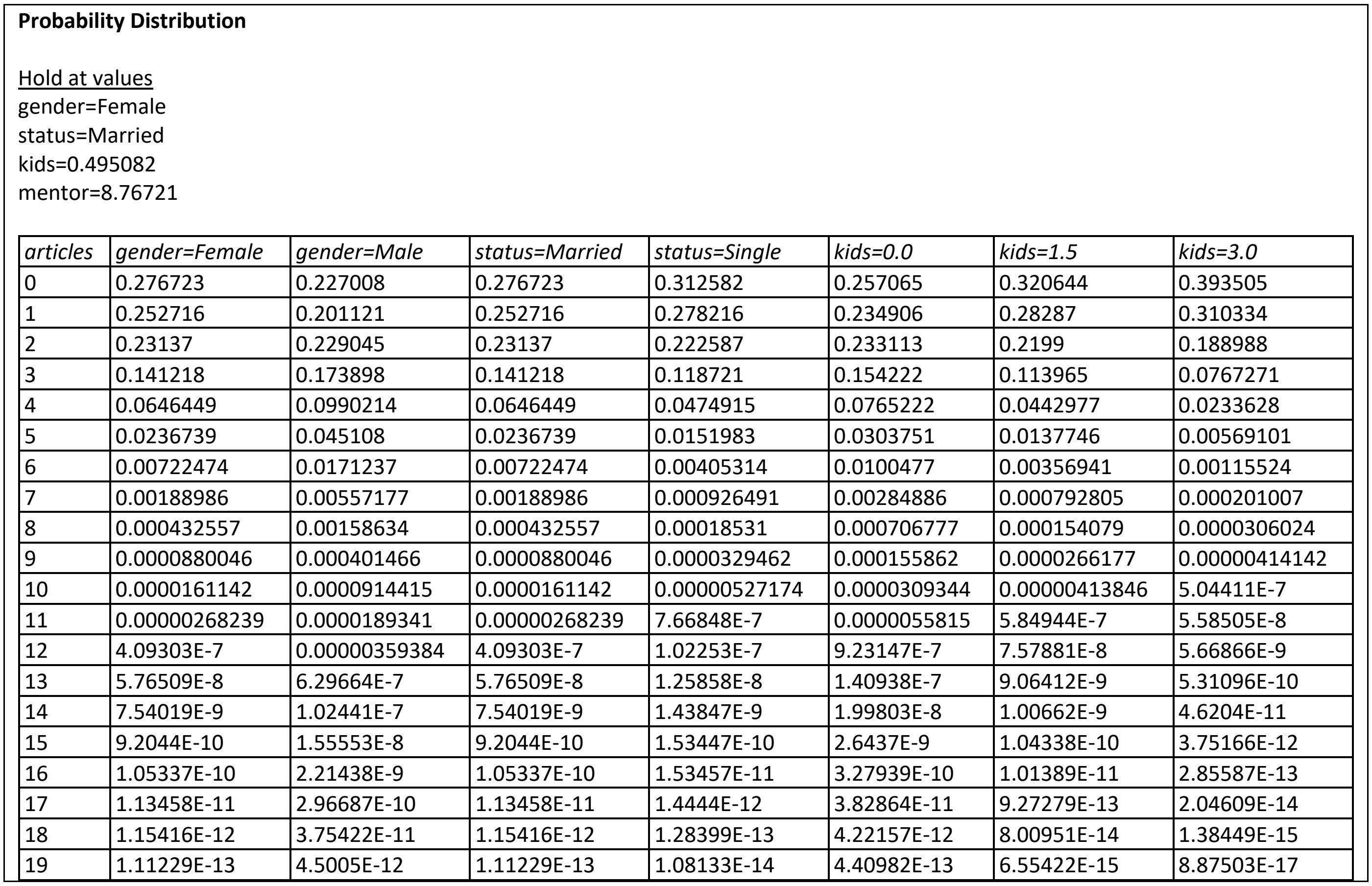
- **Factor:** select the factor to be used to determine the levels at which the probability distribution is displayed. If categorical, the probability distribution will be plotted at each level of that factor. If quantitative, the probability distribution will be plotted at the *low* value indicated, the *high* value indicated, and midway between the *low* and the *high*.

- **Low and high:** for quantitative factors, used to define the levels at which the probability distribution will be displayed.

- **Hold:** used to define the value at which the probability distribution will be displayed for all factors other than the factor selected.

Probability Distribution (Table)

The probability distribution for the dependent variable may be tabulated. It is shown for selected levels of each predictor variable, with the other variables held constant. A portion of the table is shown below.



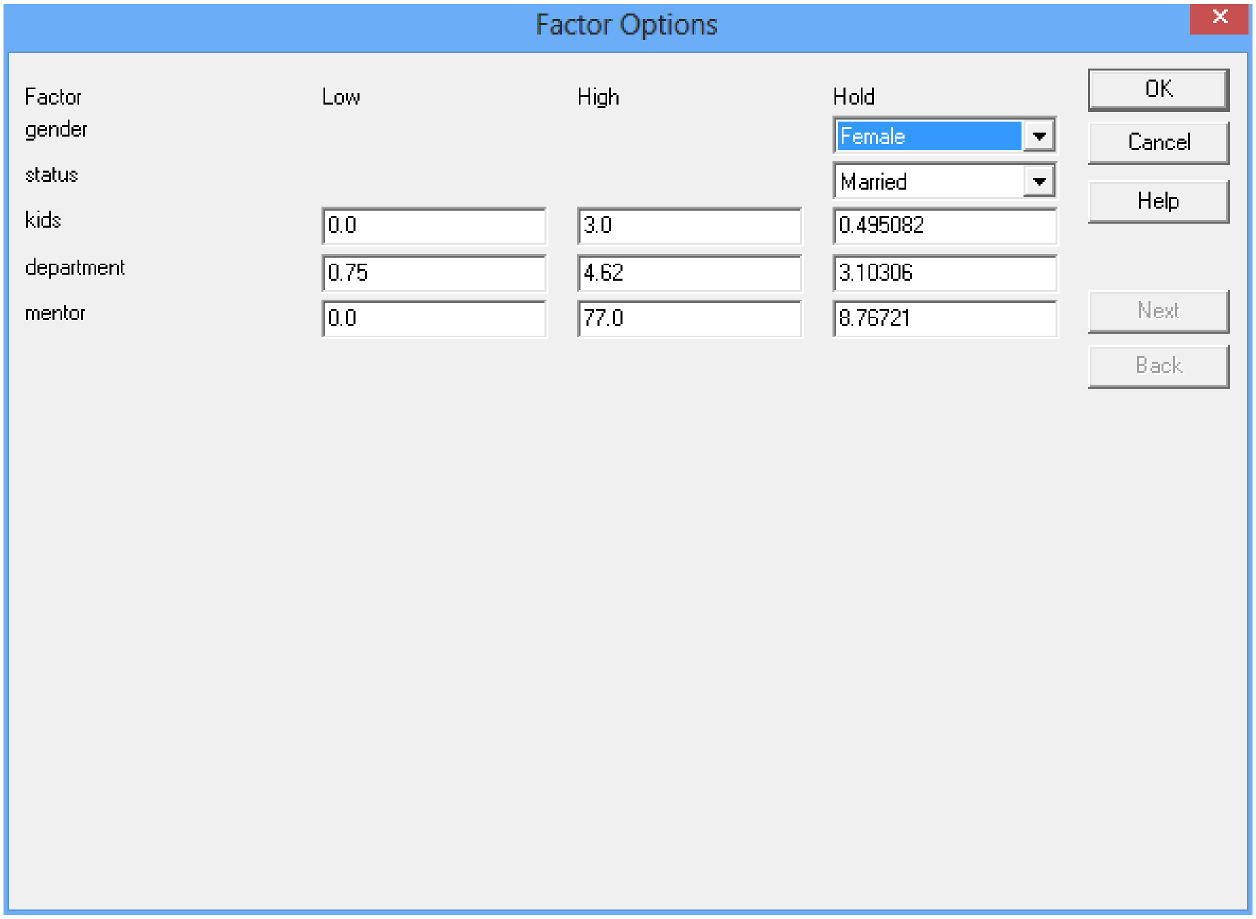
Probability Distribution

Hold at values
gender=Female
status=Married
kids=0.495082
mentor=8.76721

articles	*gender=Female*	*gender=Male*	*status=Married*	*status=Single*	*kids=0.0*	*kids=1.5*	*kids=3.0*
0	0.276723	0.227008	0.276723	0.312582	0.257065	0.320644	0.393505
1	0.252716	0.201121	0.252716	0.278216	0.234906	0.28287	0.310334
2	0.23137	0.229045	0.23137	0.222587	0.233113	0.2199	0.188988
3	0.141218	0.173898	0.141218	0.118721	0.154222	0.113965	0.0767271
4	0.0646449	0.0990214	0.0646449	0.0474915	0.0765222	0.0442977	0.0233628
5	0.0236739	0.045108	0.0236739	0.0151983	0.0303751	0.0137746	0.00569101
6	0.00722474	0.0171237	0.00722474	0.00405314	0.0100477	0.00356941	0.00115524
7	0.00188986	0.00557177	0.00188986	0.000926491	0.00284886	0.000792805	0.000201007
8	0.000432557	0.00158634	0.000432557	0.00018531	0.000706777	0.000154079	0.0000306024
9	0.0000880046	0.000401466	0.0000880046	0.0000329462	0.000155862	0.0000266177	0.00000414142
10	0.0000161142	0.0000914415	0.0000161142	0.00000527174	0.0000309344	0.00000413846	5.04411E-7
11	0.00000268239	0.0000189341	0.00000268239	7.66848E-7	0.0000055815	5.84944E-7	5.58505E-8
12	4.09303E-7	0.00000359384	4.09303E-7	1.02253E-7	9.23147E-7	7.57881E-8	5.66866E-9
13	5.76509E-8	6.29664E-7	5.76509E-8	1.25858E-8	1.40938E-7	9.06412E-9	5.31096E-10
14	7.54019E-9	1.02441E-7	7.54019E-9	1.43847E-9	1.99803E-8	1.00662E-9	4.6204E-11
15	9.2044E-10	1.55553E-8	9.2044E-10	1.53447E-10	2.6437E-9	1.04338E-10	3.75166E-12
16	1.05337E-10	2.21438E-9	1.05337E-10	1.53457E-11	3.27939E-10	1.01389E-11	2.85587E-13
17	1.13458E-11	2.96687E-10	1.13458E-11	1.4444E-12	3.82864E-11	9.27279E-13	2.04609E-14
18	1.15416E-12	3.75422E-11	1.15416E-12	1.28399E-13	4.22157E-12	8.00951E-14	1.38449E-15
19	1.11229E-13	4.5005E-12	1.11229E-13	1.08133E-14	4.40982E-13	6.55422E-15	8.87503E-17

Pane Options

By default, the values at which the other variables are held are set equal to their respective means (if quantitative) or their first level (if categorical). This may be changed using *Pane Options*:



Factor Options

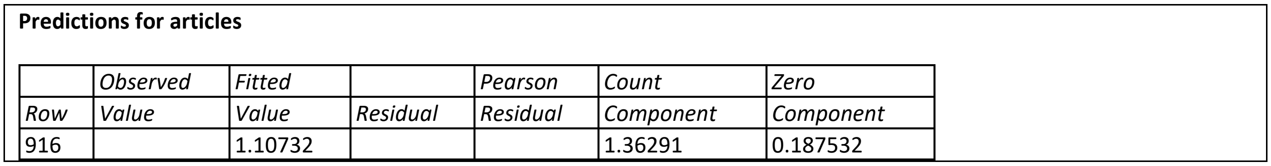
Factor	Low	High	Hold
gender			Female
status			Married
kids	0.0	3.0	0.495082
department	0.75	4.62	3.10306
mentor	0.0	77.0	8.76721

- **Low and high:** for quantitative factors, used to define the levels at which the probability distribution will be displayed. It will be shown at the *Low* value, at the *High* value, and midway between.

- **Hold:** used to define the value at which the probability distribution will be displayed for all additional factors.

Predictions

The fitted regression model may be used to predict the outcome of new samples whose predictor variables are given. For example, suppose a prediction is desired for a PhD candidate with *gender* = Female, *status* = Married, *kids* = 2, *department* = 3.50, and *mentor* = 6. A new row could be added to the datasheet with these values for the predictor variables, but the entry for *articles* would be left blank. The *Predictions* pane would then display:

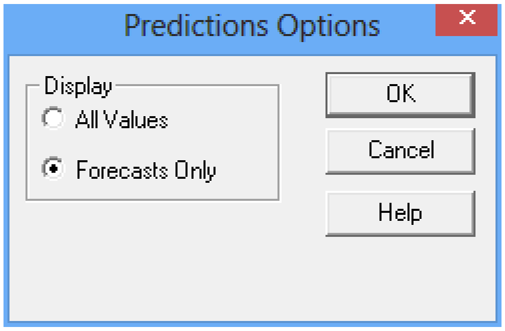


Predictions for articles

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

The *Fitted Value* shows the predicted mean number of articles produced by a candidate with the specified characteristics. In addition, the value of the count component and the zero component are displayed.

Pane Options



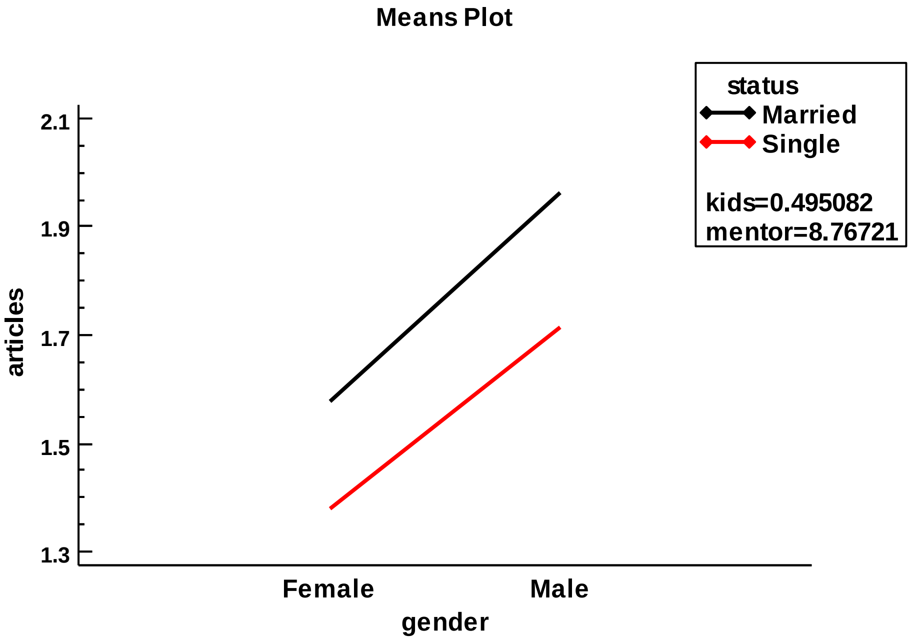
Predictions Options

Display
○ All Values
● Forecasts Only

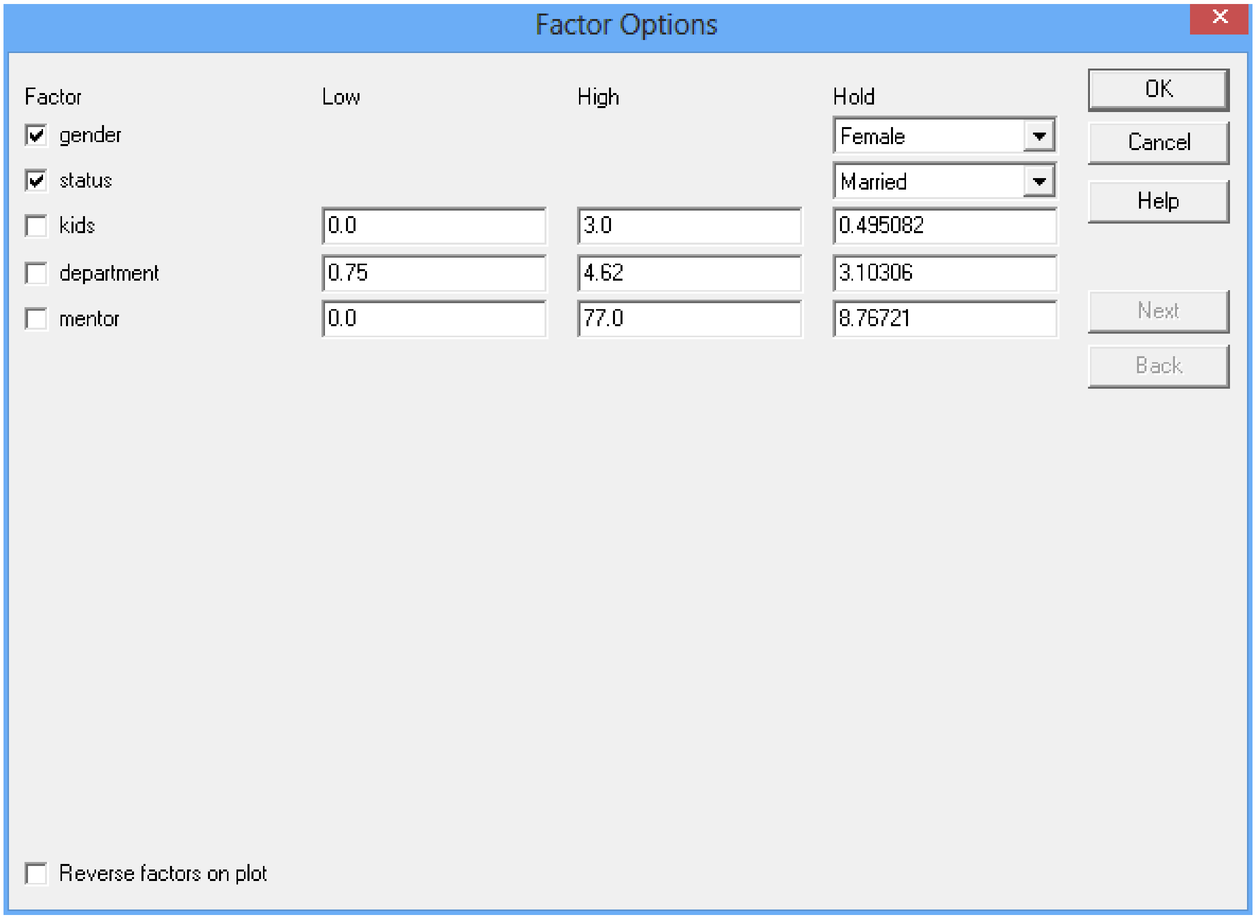
- **Display**: display *All Values* (predictions for all rows in the datasheet), or *Forecasts Only* (predictions for rows with missing values for the dependent variable).

Means Plot

The *Means Plot* displays the predicted value for the mean of the dependent variable as a function of the independent variables. The levels of one factor are plotted on the horizontal axis. The levels of a second factor are used to determine how many lines will be displayed on the plot. All other factors are held constant at specific values. For example, the plot below displays the predicted mean number of articles written by females and males who were both married and single with *kids* and *mentor* set at the indicated values.



Pane Options



Factor	Low	High	Hold
☑ gender			Female
☑ status			Married
□ kids	0.0	3.0	0.495082
□ department	0.75	4.62	3.10306
□ mentor	0.0	77.0	8.76721

□ Reverse factors on plot

- **Factor:** select 2 factors, one to be plotted on the horizontal axis and a second to define the levels at which the means will be plotted.

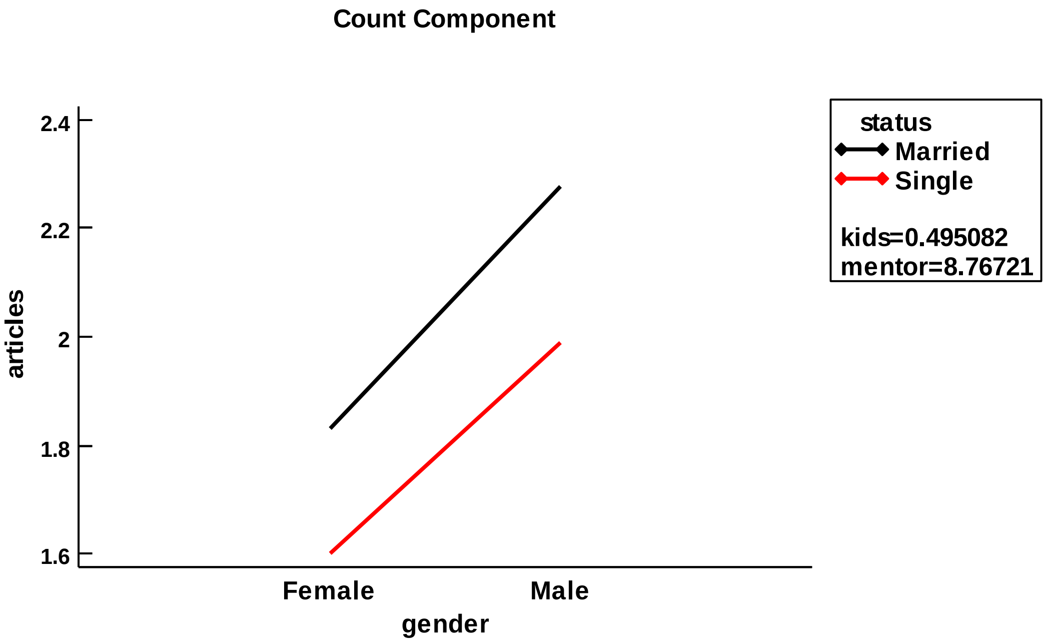
- **Low and high:** for quantitative factors, defines the range plotted on the horizontal axis when selected as the first factor and the levels at which the means are plotted when selected as the second factor.

- **Hold:** levels at which the factors are fixed when not selected.

- **Reverse factors on plot:** if checked, the second factor selected is plotted on the horizontal axis. Otherwise, the first factor selected is plotted on the horizontal axis.

Count Component

The *Count Component* displays the count component of the fitted model as a function of the independent variables.

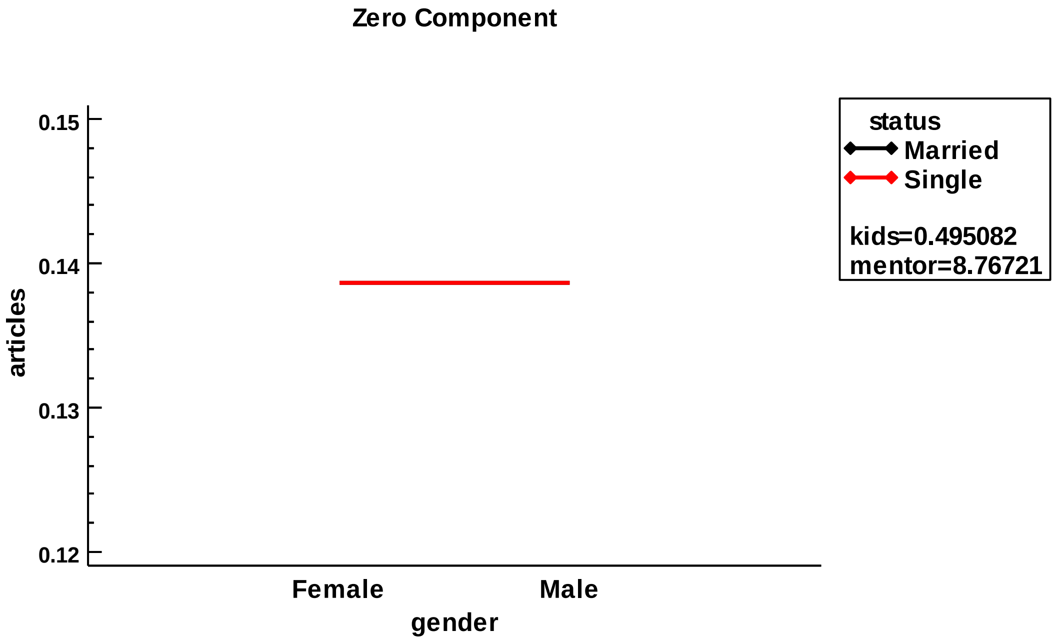


Pane Options

The options are the same as for the *Means Plot*.

Zero Component

The *Zero Component* displays the zero-inflation component of the fitted model as a function of the independent variables.



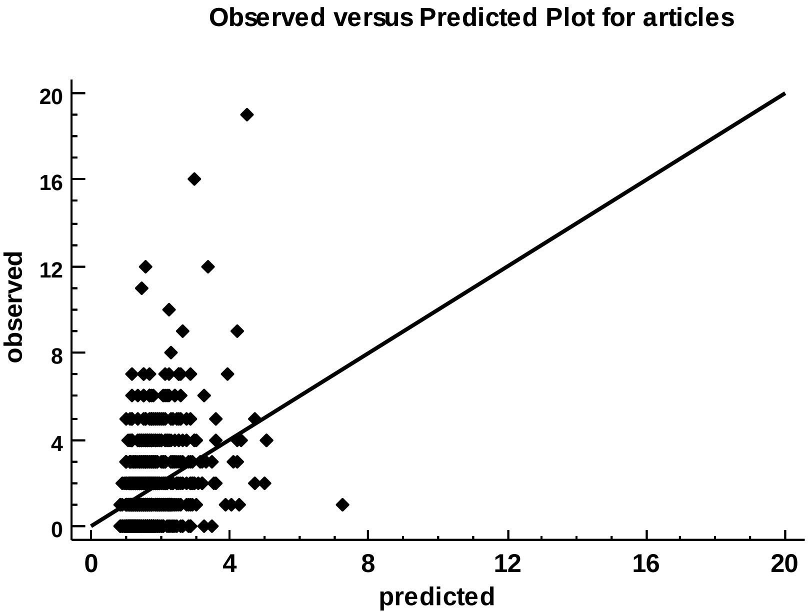
Since *gender* and *status* are not contained in the zero component model, the value is the same for all combinations of those factors.

Pane Options

The options are the same as for the *Means Plot.*

Observed Versus Predicted

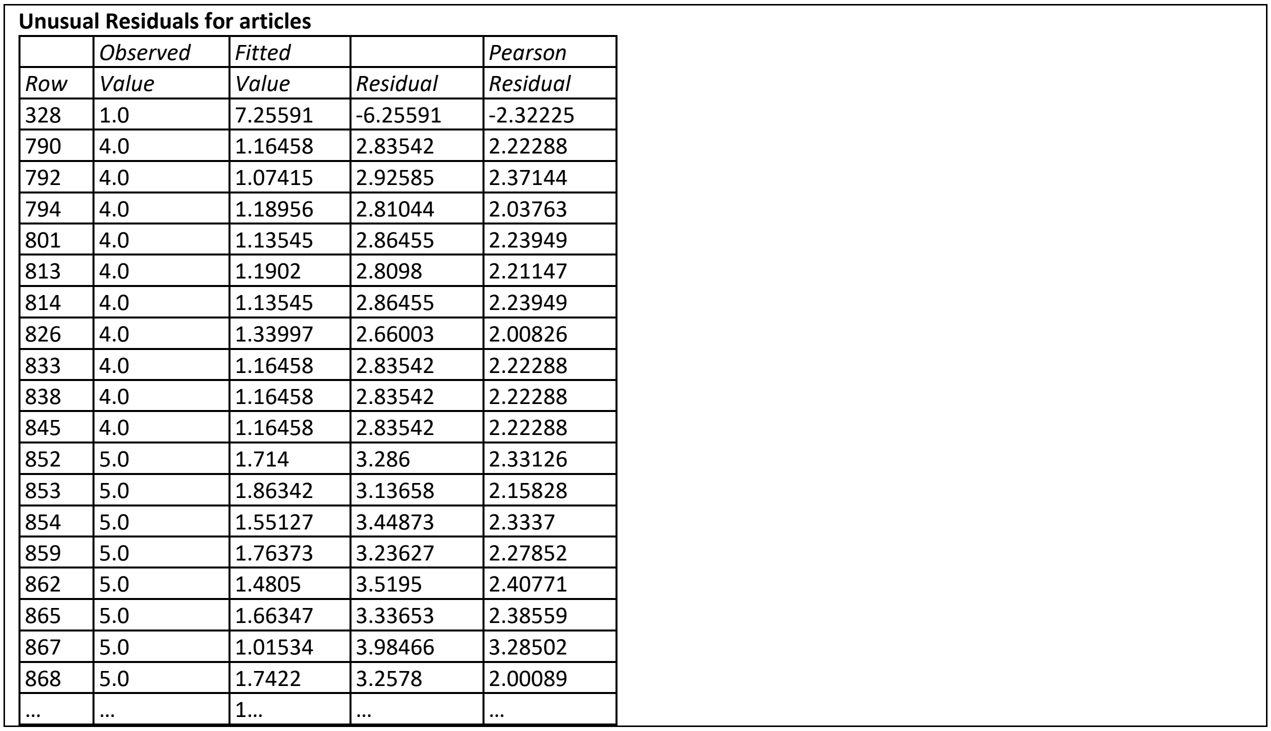
The *Observed versus Predicted* plot shows the observed values of the dependent variable on the vertical axis and the predicted values on the horizontal axis.



If the model fits well, the points should be randomly scattered around the diagonal line.

Unusual Residuals

Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have unusually large residuals. A portion of the table for the sample data is shown below:



Unusual Residuals for articles

	Observed	*Fitted*		*Pearson*
Row	*Value*	*Value*	*Residual*	*Residual*
328	1.0	7.25591	-6.25591	-2.32225
790	4.0	1.16458	2.83542	2.22288
792	4.0	1.07415	2.92585	2.37144
794	4.0	1.18956	2.81044	2.03763
801	4.0	1.13545	2.86455	2.23949
813	4.0	1.1902	2.8098	2.21147
814	4.0	1.13545	2.86455	2.23949
826	4.0	1.33997	2.66003	2.00826
833	4.0	1.16458	2.83542	2.22288
838	4.0	1.16458	2.83542	2.22288
845	4.0	1.16458	2.83542	2.22288
852	5.0	1.714	3.286	2.33126
853	5.0	1.86342	3.13658	2.15828
854	5.0	1.55127	3.44873	2.3337
859	5.0	1.76373	3.23627	2.27852
862	5.0	1.4805	3.5195	2.40771
865	5.0	1.66347	3.33653	2.38559
867	5.0	1.01534	3.98466	3.28502
868	5.0	1.7422	3.2578	2.00089
…	…	1…	…	…

The table displays:

- **Row** – the row number in the datasheet.

- **Observed Value** – the observed value of the dependent variable.

- **Predicted Value** – the value predicted by the fitted model.

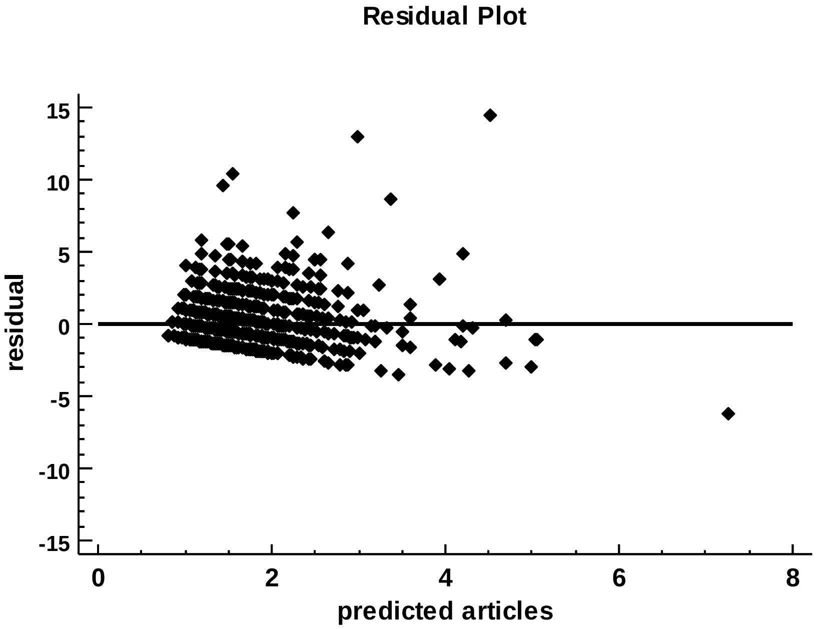
- **Residual** – the difference between the observed and predicted values.

- **Pearson Residual** – a standardized residual in which each residual is divided by an estimate of its standard error.

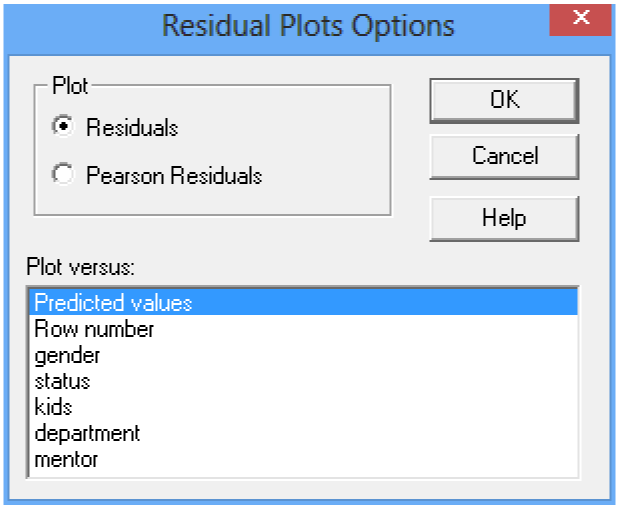
The table includes all rows for which the absolute value of the Pearson residual is greater than 2.0. Absolute values in excess of 3 should be examined carefully to determine whether they correspond to outliers.

Residual Plots

This graph plots the residuals from the fitted model versus predicted values. It is helpful in visualizing whether the variability of the residuals is constant or depends on the predicted value.



Pane Options



Plot

● Residuals

○ Pearson Residuals

Plot versus:

Predicted values
Row number
gender
status
kids
department
mentor

- **Plot:** selects the type of residuals to display.

- **Plot versus:** selects the item to be plotted on the horizontal axis.

Save Results

The following results may be saved to the datasheet:

1. *Fitted values* – the fitted values corresponding to each row of the datasheet.
2. *Residuals* – the ordinary residuals.
3. *Pearson Residuals* – the standardized Pearson residuals.
4. *Count component* – the count component for each row.
5. *Zero component* – the zero component for each row.

Calculations

The calculations are performed by R using the PSCL package.

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

Vuong, Q.H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 57:307-333.