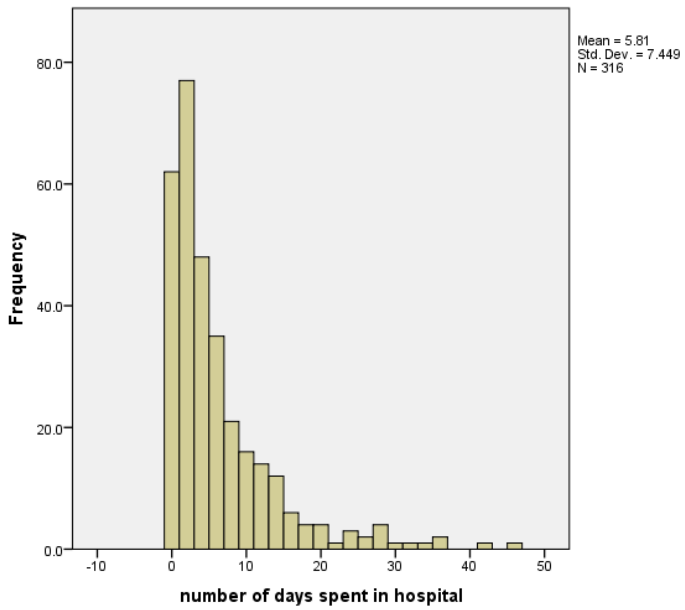
# SPSS Data Analysis Using Poisson Regression For 'Simple' Count Data

Answers to practical questions

**1.**




The data are strongly skewed to the right, breaking the normality assumption of OLS regression. Count data often follow a poisson distribution, so some type of poisson analysis might be appropriate.

**2.**

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Gender | 316 | 1 | 2 | 1.49 | .501 | .251 | .051 | .137 | -2.010 | .273 |
| Ethnicity | 316 | 1 | 6 | 4.08 | .943 | .889 | -.609 | .137 | .247 | .273 |
| hospital 1 or 2 | 316 | 1 | 2 | 1.50 | .501 | .251 | .013 | .137 | -2.013 | .273 |
| FirstTestResult | 316 | 1.01 | 98.99 | 48.7510 | 17.88076 | 319.721 | -.046 | .137 | .630 | .273 |
| SecondTestResult | 316 | 1.01 | 98.99 | 50.0638 | 17.93921 | 321.815 | -.168 | .137 | .567 | .273 |
| number of days spent in hospital | 316 | 0 | 45 | 5.81 | 7.449 | 55.488 | 2.261 | .137 | 6.064 | .273 |
| Valid N (listwise) | 316 | | | | | | | | | |

The variance of **daysInHosp**, our outcome variable, is nearly 10 times larger than the mean. The distribution of **dayInHops** is displaying signs of overdispersion, that is, greater variance than might

be expected in a poisson distribution. The skew and kurtosis statistics offer another way to look at this assumption.

**3.** The first four tables of output can be used to check you have specified the model correctly. The '**Model Information**' table should show poisson as the probability distribution, log as the link function and number of days spent in hospital as the dependent variable.
Remember in descriptive analyses seeing that there were 316 unique ID's? In the '**Case Processing Summary**' table you can see that no participants have any missing data and as a consequence no cases have been excluded from the analysis.
The following two tables give a descriptive summary of the variables included in the model. These are a good place to check that you have specified the correct variables for inclusion in the model in large datasets.

**Model Information**

| | |
|---|---|
| Dependent Variable | number of days spent in hospital |
| Probability Distribution | Poisson |
| Link Function | Log |

**Case Processing Summary**

| | N | Percent |
|---|---|---|
| Included | 316 | 100.0% |
| Excluded | 0 | 0.0% |
| Total | 316 | 100.0% |

**Categorical Variable Information**

| | | N | Percent |
|---|---|---|---|
| Factor gender | female | 162 | 51.3% |
| | male | 154 | 48.7% |
| | Total | 316 | 100.0% |

**Continuous Variable Information**

| | | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| Dependent Variable | number of days spent in hospital | 316 | 0 | 45 | 5.81 | 7.449 |
| Covariate | FirstTestResult | 316 | 1.01 | 98.99 | 48.7510 | 17.88076 |
| | SecondTestResult | 316 | 1.01 | 98.99 | 50.0638 | 17.93921 |

The Model output begins with the **Goodness of Fit table** and the **Ominubus table**. Both list various statistics indicating overall model fit.
***Information criteria*** can be compared between models, with other non-nested models, to assess their comparative fit. A lower number indicates a better fitting (more parsimonious model). NB: it is better to assess the corrected AICC when the dataset is small.
***Deviance*** should not be much higher than 1, here it is quite high indicating a poor fit.
The goodness-of-fit chi-squared test can be utilised. We evaluate the deviance (2235) as Chi-square distributed with the model degrees of freedom (312). This is not a test of the model coefficients (as specified in the header information), but a test of the model form: Does the poisson model form fit our data? We conclude that the model fits reasonably well because the ***goodness-of-fit chi-squared test*** is not statistically significant (p =8.892). If the test had been statistically significant, it would indicate that the data do not fit the model well.

In a situation where model fit is very poor we can try to determine why by checking if there are omitted predictor variables of importance, if our linearity assumption holds and/or if there is an issue of over-dispersion (which we already suspect in this case).

The ***omnibus test*** is a test of the model fit as a whole the p-value indicates that the overall model is significant so we can go ahead and interpret the different pieces of the model.

**Goodness of Fit[a]**

| | Value | df | Value/df |
|---|---|---|---|
| Deviance | 2234.546 | 312 | 7.162 |
| Scaled Deviance | 2234.546 | 312 | |
| Pearson Chi-Square | 2774.414 | 312 | 8.892 |
| Scaled Pearson Chi-Square | 2774.414 | 312 | |
| Log Likelihood[b] | -1547.971 | | |
| Akaike's Information Criterion (AIC) | 3103.942 | | |
| Finite Sample Corrected AIC (AICC) | 3104.071 | | |
| Bayesian Information Criterion (BIC) | 3118.965 | | |
| Consistent AIC (CAIC) | 3122.965 | | |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult
a. Information criteria are in small-is-better form.
b. The full log likelihood function is displayed and used in computing information criteria.

**Omnibus Test[a]**

| Likelihood Ratio Chi-Square | Df | Sig. |
|---|---|---|
| 175.274 | 3 | .000 |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult
a. Compares the fitted model against the intercept-only model.

Next is the Tests of Model Effects. This evaluates each of the model variables with the appropriate degrees of freedom. All variables are significant or bordering on significance so should remain in the model.

**Tests of Model Effects**

| Source | Type III | | |
|---|---|---|---|
| | Wald Chi-Square | df | Sig. |
| (Intercept) | 1374.727 | 1 | .000 |
| Gender | 68.582 | 1 | .000 |
| FirstTestResult | 3.742 | 1 | .053 |
| SecondTestResult | 43.865 | 1 | .000 |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult

The last table shows the Parameter Estimates. This includes the regression coefficients (B) for each of the variables along with standard errors, p-values and 95% confidence intervals for the coefficients. In SPSS you need to indicate that the Exp(B) is to be reported, once done, it is given alongside 95% CI representing an incidence rate ratio.

The exponentiated coefficient for gender indicates a positive association between days in hospital and being female; the rate of days in hospital for women is 49% higher than it is for men (holding all other variables constant).

There is weak evidence (p-value borders significance) to suggest that for a one unit increase in 'First test result' the rate of 'days in hospital' decrease by 1% (holding all other variables constant).
The exponentiated coefficient for 'Second test result' indicates a negative association between days in hospital and second test result. The rate of days in hospital for those with a unit increase in Second test result decreases by 1.2%(holding all other variables constant).

**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | Df | Sig. | | Lower | Upper |
| (Intercept) | 2.287 | .0700 | 2.150 | 2.424 | 1068.590 | 1 | .000 | 9.843 | 8.582 | 11.289 |
| [gender=1] | .401 | .0484 | .306 | .496 | 68.582 | 1 | .000 | 1.493 | 1.358 | 1.642 |
| [gender=2] | 0[a] | . | . | . | . | . | . | 1 | . | . |
| FirstTestResult | -.004 | .0018 | -.007 | 4.656E-005 | 3.742 | 1 | .053 | .996 | .993 | 1.000 |
| SecondTestResult | -.012 | .0018 | -.016 | -.009 | 43.865 | 1 | .000 | .988 | .984 | .991 |
| (Scale) | 1[b] | | | | | | | | | |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.

## 4.

***Information criteria*** can be compared between models, with other non-nested models, to assess their comparative fit. A lower number indicates a better fitting (more parsimonious model). NB: it is better to assess the corrected AICC when the dataset is small. Here we see that the information criteria have lowered slightly, indicating that the model fits better when all variables are included.
***Deviance*** should not be much higher than 1, it remains quite high indicating that the model still has quite a poor fit.
The goodness-of-fit chi-squared; we evaluate the deviance (1938) as Chi-square distributed with the model degrees of freedom (306). This is not a test of the model coefficients (as specified in the header information), but a test of the model form: Does the poisson model form with further variables included fit our data? We conclude that although the goodness of fit is quite poor that the model fits reasonably well because the ***goodness-of-fit chi-squared test*** is not statistically significant (p =8.892).
Given the model fit is still quite poor we can try to determine why by checking if there are omitted predictor variables of importance, if our linearity assumption holds and/or if there is an issue of over-dispersion (which we already suspect in this case).

**Goodness of Fit**[a]

| | Value | Df | Value/df |
|---|---|---|---|
| Deviance | 1937.753 | 306 | 6.333 |
| Scaled Deviance | 1937.753 | 306 | |
| Pearson Chi-Square | 2337.494 | 306 | 7.639 |
| Scaled Pearson Chi-Square | 2337.494 | 306 | |
| Log Likelihood[b] | -1399.575 | | |
| Akaike's Information Criterion (AIC) | 2819.149 | | |
| Finite Sample Corrected AIC (AICC) | 2819.870 | | |
| Bayesian Information Criterion (BIC) | 2856.707 | | |
| Consistent AIC (CAIC) | 2866.707 | | |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult, ethnic, hospital
a. Information criteria are in small-is-better form.
b. The full log likelihood function is displayed and used in computing information criteria.

As before the ***omnibus test*** shows the model fit as a whole the p-value indicates that the overall model is significant so we again can go ahead and interpret the different pieces of the model.

**Omnibus Test**[a]

| Likelihood Ratio Chi-Square | Df | Sig. |
|---|---|---|
| 472.067 | 9 | .000 |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult, ethnic, hospital
a. Compares the fitted model against the intercept-only model.

Next is the Tests of Model Effects. This evaluates each of the model variables with the appropriate degrees of freedom. Most variables remain significant so should remain in the model, except from First Test Result, however, we may have causal or clinical reasons for keeping this variable forced in the model.

**Tests of Model Effects**

| Source | Type III | | |
|---|---|---|---|
| | Wald Chi-Square | df | Sig. |
| (Intercept) | 325.624 | 1 | .000 |
| Gender | 71.298 | 1 | .000 |
| FirstTestResult | .693 | 1 | .405 |
| SecondTestResult | 4.522 | 1 | .033 |
| Ethnic | 68.735 | 5 | .000 |
| Hospital | 69.802 | 1 | .000 |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult, ethnic, hospital

Comparing goodness of fit data shows that the more complicated model is more parsimonious, however, dependent on causal (or perhaps clinical) reasoning we may want to remove '**FirstTestResult'** from the model. In this case, as it is part of our causal theory I would leave it in.
Again we have the regression coefficients (B) for each of the variables along with standard errors, p-values and 95% confidence intervals for the coefficients. Don't forget that in SPSS you need to indicate that the Exp(B) is to be reported, once done, it is given alongside 95% CI representing an incidence rate ratio.
The exponentiated coefficient for gender still indicates a positive association between days in hospital and being female; the effect size has increased slightly. The rate of days in hospital for women is 51% higher than it is for men (holding all other variables constant).
This model shows no strong evidence to suggest an association between 'First test result' and the rate of 'days in hospital'.
The exponentiated coefficient for 'Second test result' shows weak evidence of a negative association between days in hospital and second test result. The effect estimate has more than halved, rate of days in hospital for those with a unit increase in Second test result decreases by 0.5%(holding all other variables constant).
The exponentiated coefficient for 'Hospital' type shows evidence of a fairly strong association. A more formal way to report the incidence rate ratio for this is as follows: In type one hospitals comparative to type two hospitals the incidence rate ratio for days in hospital is 1.798 (95% CI 1.57-2.06, p-value = <0.001)
All ethnicity variables, except Filipino which borders on significance, show a relationship with days in hospital comparative to Pacific Islanders. For categorical variables you may wish to change the held variable this can be easily done when setting the model parameters.

**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | Df | Sig. | | Lower | Upper |
| (Intercept) | .722 | .2174 | .296 | 1.148 | 11.033 | 1 | .001 | 2.059 | 1.345 | 3.153 |
| [gender=1] | .411 | .0487 | .316 | .507 | 71.298 | 1 | .000 | 1.509 | 1.371 | 1.660 |
| [gender=2] | 0[a] | . | . | . | . | . | . | 1 | . | . |
| FirstTestResult | -.002 | .0018 | -.005 | .002 | .693 | 1 | .405 | .998 | .995 | 1.002 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SecondTestResult | -.004 | .0020 | -.008 | .000 | 4.522 | 1 | .033 | .996 | .992 | 1.000 |
| [ethnic=1] | 1.343 | .4559 | .450 | 2.237 | 8.681 | 1 | .003 | 3.832 | 1.568 | 9.365 |
| [ethnic=2] | 1.052 | .2043 | .652 | 1.453 | 26.507 | 1 | .000 | 2.864 | 1.919 | 4.274 |
| [ethnic=3] | .880 | .1984 | .491 | 1.268 | 19.653 | 1 | .000 | 2.410 | 1.633 | 3.555 |
| [ethnic=4] | .809 | .1916 | .434 | 1.185 | 17.841 | 1 | .000 | 2.247 | 1.543 | 3.271 |
| [ethnic=5] | .344 | .2075 | -.063 | .750 | 2.741 | 1 | .098 | 1.410 | .939 | 2.118 |
| [ethnic=6] | 0[a] | . | . | . | . | . | . | 1 | . | . |
| [hospital=1] | .586 | .0702 | .449 | .724 | 69.802 | 1 | .000 | 1.798 | 1.566 | 2.063 |
| [hospital=2] | 0[a] | . | . | . | . | . | . | 1 | . | . |
| (Scale) | 1[b] | | | | | | | | | |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult, ethnic, hospital
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.

## 6.

Comparing the goodness of fit data for this model with the others shows that the more complicated negative binomial model is more parsimonious, however, dependent on causal (or perhaps clinical) reasoning we may want to remove '**FirstTestResult'** from the model. In this case, as it is part of our causal theory I would leave it in.

The regression coefficients (B) for each of the variables along with standard errors, p-values and 95% confidence intervals for the coefficients is displayed in the parameter table alongside the indicated Exp(B) with 95% CI representing an incidence rate ratio.

The exponentiated coefficient for gender still indicates a positive association between days in hospital and being female; the effect size has again slightly increased. The rate of days in hospital for women is 54% higher than it is for men (holding all other variables constant).

This model shows no strong evidence to suggest an association between 'First test result' and the rate of 'days in hospital'.

The exponentiated coefficient for 'Second test result' shows weak evidence of a negative association between days in hospital and second test result. The effect estimate is similar to that in the first model, the rate of days in hospital for those with a unit increase in Second test result decreases by 1.4%(holding all other variables constant).

**Goodness of Fit[a]**

| | Value | Df | Value/df |
|---|---|---|---|
| Deviance | 425.459 | 312 | 1.364 |
| Scaled Deviance | 425.459 | 312 | |
| Pearson Chi-Square | 418.790 | 312 | 1.342 |
| Scaled Pearson Chi-Square | 418.790 | 312 | |
| Log Likelihood[b] | -884.423 | | |
| Akaike's Information Criterion (AIC) | 1776.846 | | |

| | | | |
|---|---|---|---|
| Finite Sample Corrected AIC (AICC) | 1776.975 | | |
| Bayesian Information Criterion (BIC) | 1791.869 | | |
| Consistent AIC (CAIC) | 1795.869 | | |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult
a. Information criteria are in small-is-better form.
b. The full log likelihood function is displayed and used in computing information criteria.

**Omnibus Test[a]**

| Likelihood Ratio Chi-Square | df | Sig. |
|---|---|---|
| 26.733 | 3 | .000 |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult
a. Compares the fitted model against the intercept-only model.

**Tests of Model Effects**

| Source | Type III | | |
|---|---|---|---|
| | Wald Chi-Square | df | Sig. |
| (Intercept) | 176.233 | 1 | .000 |
| Gender | 11.807 | 1 | .001 |
| FirstTestResult | .145 | 1 | .704 |
| SecondTestResult | 8.159 | 1 | .004 |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult

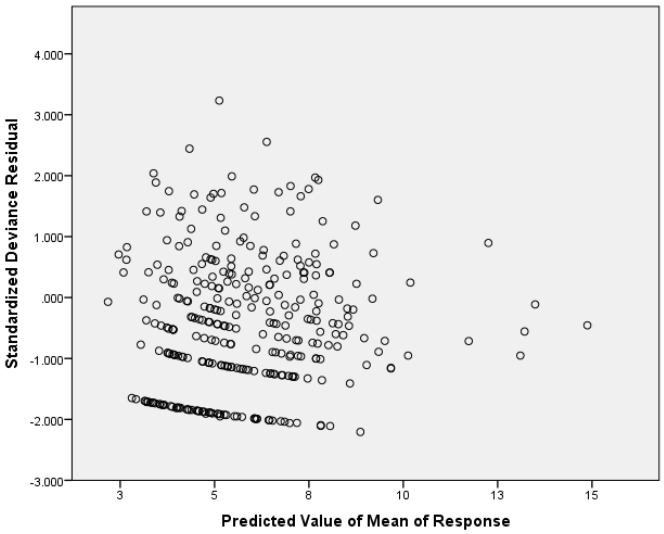
**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | 2.285 | .1881 | 1.917 | 2.654 | 147.532 | 1 | .000 | 9.828 | 6.797 | 14.211 |
| [gender=1] | .431 | .1253 | .185 | .676 | 11.807 | 1 | .001 | 1.538 | 1.203 | 1.966 |
| [gender=2] | 0[a] | . | . | . | . | . | . | 1 | . | . |
| FirstTestResult | -.002 | .0044 | -.010 | .007 | .145 | 1 | .704 | .998 | .990 | 1.007 |
| SecondTestResult | -.014 | .0050 | -.024 | -.004 | 8.159 | 1 | .004 | .986 | .976 | .996 |
| (Scale) | 1[b] | | | | | | | | | |
| (Negative binomial) | 1[b] | | | | | | | | | |

Dependent Variable: number of days spent in hospital
Model: (Intercept), gender, FirstTestResult, SecondTestResult
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.

Things you may not have looked at today!

- Plotting the standard deviation residuals against the predicted values of the mean response in order to judge model fit. This can be done by selecting 'predicted value of mean response' and 'standardized deviance residual' on the save tab when running regression. The generated variables can then be used to create a scatter plot with deviance on the y axis and mean values on the x. This is what we would have seen for the negative binomial regression model. We are looking to have a 'cloud' of residuals with deviance values between -3.3 and 3.3 i.e. centred around 0. Here we see no significant outliers which can compromise model fit. However, we do have a slight tendency for more variability at the lower values suggesting there is some heteroscedasity in the model.




- Using the shapiro-wilk or/and the kolmogorov-smirnov test as a formal assessment of normality (these are often somewhat underpowered and normality can be tested easily in a non-formal way (e.g. histogram or box plot).
  Analyze -> descriptive statistics -> Explore -> Choose dependent var and then click plots and choose plots with normality statistics

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| number of days spent in hospital | .218 | 316 | .000 | .742 | 316 | .000 |

a. Lilliefors Significance Correction

Both test are significant indicating that our outcome data is not normally distributed.

- Using Q-Q plots as test for normality, when normality exists data will fit closely to trend lines. In skewed count data we will expect to see our observed values resting on the line of normality until you reach the higher values.

Analyze -> descriptive statistics -> Q-Q plots