Count data and

Poisson distribution

GENERAL IZED L INEAR MO DELS IN PYTH O N

Ita Cirovic Donev

Data Science Consultant

Count data

Count the number of occurrences in a specied unit of time, distance, area or volume Examples:

Goals in a soccer match

Number of earthquakes

Number of crab satellites

Number of awards won by a person

Number of bike crossings over the bridge

GENERALIZED LINEAR MODELS IN PYTHON

Poisson random variable Events occur independently and randomly Poisson distribution

*P*(*y*) =

*λ* : mean and variance

*y* = 0, 1, 2, 3, ...

Always positive

Discrete (not continuous)

Lower bound at zero, but no upper bound

*λ ey* −*λy*!

GENERALIZED LINEAR MODELS IN PYTHON

Understanding the parameter of the Poisson distribution

GENERALIZED LINEAR MODELS IN PYTHON

Visualizing the response import seaborn as sns

sns.distplot('y')

GENERALIZED LINEAR MODELS IN PYTHON

Poisson regression

Response variable

*y* ∼ *Poisson*(*λ*)

Mean of the response

*E*(*y*) = *λ*

Poisson regression model

*log*(*λ*) = *β*0 + *β*1*x*1

GENERALIZED LINEAR MODELS IN PYTHON

Explanatory variables

Continuous and/or categorical → Poisson regression model Categorical → log-linear model

GENERALIZED LINEAR MODELS IN PYTHON

GLM with Poisson in Python

import statsmodels.api as sm

from statsmodels.formula.api import glm

glm('y ~ x',

data = my\_data,

family = sm.families.Poisson())

GENERALIZED LINEAR MODELS IN PYTHON

Let's practice!

GENERAL IZED L INEAR MO DELS IN PYTH O N

Interpreting model

fit

GENERAL IZED L INEAR MO DELS IN PYTH O N

Ita Cirovic Donev

Data Science Consultant

Parameter estimation Maximum likelihood estimation (MLE) Iteratively reweighted least squares (IRLS)

GENERALIZED LINEAR MODELS IN PYTHON

The response function

Poisson regression model

*log*(*λ*) = *β* + *β x*

0 1 1

The response function:

*λ* = *exp*(*β* + *β x* )

0 1 1

or

*λ* = *exp*(*β* ) × *exp*(*β x* )

0 1 1

GENERALIZED LINEAR MODELS IN PYTHON

The response function

Poisson regression model

*log*(*λ*) = *β* + *β x*

0 1 1

The response function:

*λ* = *exp*(*β* + *β x* )

0 1 1

or

*λ* = *exp*(*β* ) × *exp*(*β x* )

0 1 1

GENERALIZED LINEAR MODELS IN PYTHON

Interpretation of parameters

*exp*(*β* )

0

The eect on the mean *λ* when *x* = 0

*exp*(*β* )

1

The multiplicative eect on the mean *λ* for a 1-unit increase in *x*

GENERALIZED LINEAR MODELS IN PYTHON

Interpreting coefficient effect If *β* > 0

1

*exp*(*β* ) > 1

1

*λ* is *exp*(*β* ) times larger than when

1

*x* = 0

If *β* = 0

1

*exp*(*β* ) = 1

1

*λ* = *exp*(*β* )

0

Multiplicative factor is 1

*y* and *x* are not related

If *β* < 0

*exp*(*β* ) < 1

1

*λ* is *exp*(*β* ) times smaller than when

1

*x* = 0

GENERALIZED LINEAR MODELS IN PYTHON

Example

model = glm('sat ~ weight', data = crab,

family = sm.families.Poisson()).fit()

Generalized Linear Model Regression Results (print cut)

============================================================================= coef std err z P>|z| [0.025 0.975] ----------------------------------------------------------------------------- Intercept -0.4284 0.179 -2.394 0.017 -0.779 -0.078 weight 0.5893 0.065 9.064 0.000 0.462 0.717 =============================================================================

GENERALIZED LINEAR MODELS IN PYTHON

Example - interpretation of beta

Extract model coecients model.params

Intercept -0.428405 weight 0.589304

Compute the eect np.exp(0.589304)

1.803

GENERALIZED LINEAR MODELS IN PYTHON

Confidence interval for ...

*β*

The multiplicative eect on mean

1

print(model.conf\_int())

0 1

Intercept -0.779112 -0.077699 weight 0.461873 0.716735

print(np.exp(crab\_fit.conf\_int()))

0 1

Intercept 0.458813 0.925243 weight 1.587044 2.047737

GENERALIZED LINEAR MODELS IN PYTHON

Let's practice!

GENERAL IZED L INEAR MO DELS IN PYTH O N

The Problem of

Overdispersion

GENERAL IZED L INEAR MO DELS IN PYTH O N

Ita Cirovic Donev

Data Science Consultant

Understanding the data

# mean of y

y\_mean = crab['sat'].mean()

2.919

# variance of y

y\_variance = crab['sat'].var()

9.912

GENERALIZED LINEAR MODELS IN PYTHON

Mean not equal to variance *variance* > *mean* → overdispersion *variance* < *mean* → underdispersion Consequences:

Small standard errors

Small p-value

GENERALIZED LINEAR MODELS IN PYTHON

How to check for overdispersion?

GENERALIZED LINEAR MODELS IN PYTHON

Compute estimated overdispersion

ratio = crab\_fit.pearson\_chi2 / crab\_fit.df\_resid print(ratio)

3.134

Ratio = 1 → approximately Poisson

Ratio < 1 → underdispersion

Ratio > 1 → overdispersion

GENERALIZED LINEAR MODELS IN PYTHON

Negative Binomial Regression *E*(*y*) = *λ*

2

*V ar*(*y*) = *λ* + *αλ*

*α* - dispersion parameter

GENERALIZED LINEAR MODELS IN PYTHON

GLM negative Binomial in Python

import statsmodels.api as sm

from statsmodels.formula.api import glm

model = glm('y ~ x', data = my\_data,

family = sm.families.NegativeBinomial(alpha = 1)).fit()

GENERALIZED LINEAR MODELS IN PYTHON

Let's practice!

GENERAL IZED L INEAR MO DELS IN PYTH O N

Plotting a regression

model

GENERAL IZED L INEAR MO DELS IN PYTH O N

Ita Cirovic Donev

Data Science Consultant

Import libraries

import seaborn as sns

import matplotlib.pyplot as plt

Crab model 'sat ~ width' is saved as model

GENERALIZED LINEAR MODELS IN PYTHON

Plot data points

# Adjust figure size

plt.subplots(figsize = (8, 5))

# Plot data points

sns.regplot('width','sat', data = crab,

fit\_reg = False)

GENERALIZED LINEAR MODELS IN PYTHON

Add jitter

sns.regplot('width','sat', data = crab,

fit\_reg = False,

y\_jitter = 0.3)

GENERALIZED LINEAR MODELS IN PYTHON

Add linear fit

sns.regplot('width','sat',

data = crab,

y\_jitter = 0.3,

fit\_reg = True,

line\_kws = {'color':'green',

'label':'LM fit'})

GENERALIZED LINEAR MODELS IN PYTHON

Add Poisson GLM estimated values crab['fit\_values'] = model.fittedvalues

sns.scatterplot('width','fit\_values',

data = crab,

color = 'red',

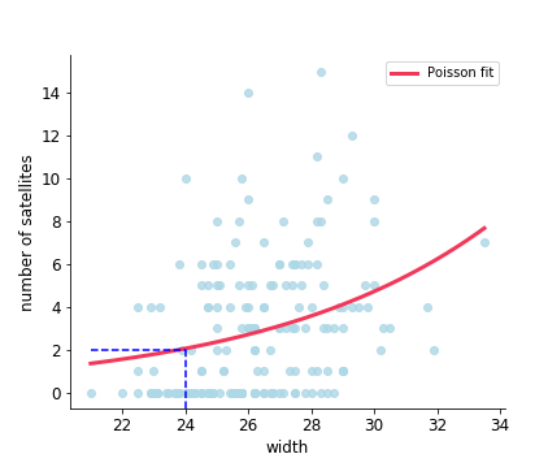
label = 'Poisson')

GENERALIZED LINEAR MODELS IN PYTHON

Predictions

GENERALIZED LINEAR MODELS IN PYTHON

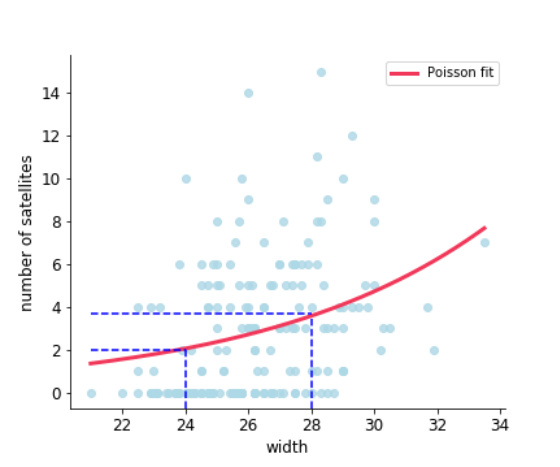
Predictions

new\_data = pd.DataFrame({'width':[24, 28, 32]}) model.predict(new\_data)

0 1.881981

GENERALIZED LINEAR MODELS IN PYTHON

Predictions

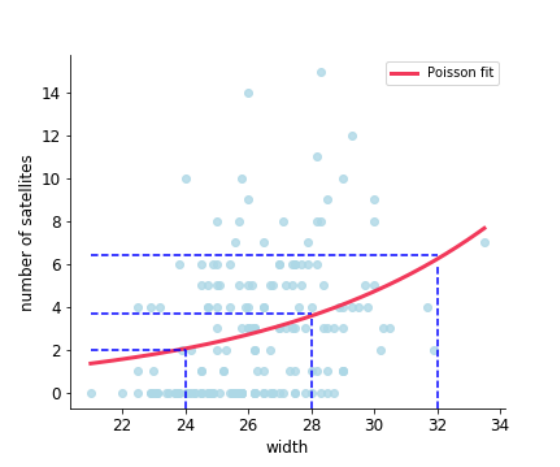
new\_data = pd.DataFrame({'width':[24, 28, 32]}) model.predict(new\_data)

0 1.881981

1 3.627360

GENERALIZED LINEAR MODELS IN PYTHON

Predictions

new\_data = pd.DataFrame({'width':[24, 28, 32]}) model.predict(new\_data)

0 1.881981

1 3.627360

2 6.991433

GENERALIZED LINEAR MODELS IN PYTHON

Let's practice!

GENERAL IZED L INEAR MO DELS IN PYTH O N