Introduction to Generalized Linear Models

Part 4: Poisson Regression Revisited

Recall: Modeling doctor visits.

```
doctor_visits.head()
```

	visits	gender	age	income	illness	reduced
0	1	female	72.0	0.25	4	7
1	0	male	72.0	0.35	0	0
2	0	female	47.0	0.75	1	0
3	0	female	62.0	0.25	0	0
4	4	female	72.0	0.35	4	0

```
poisreg_age.summary()
```

Generalized Linear Model Regression Results

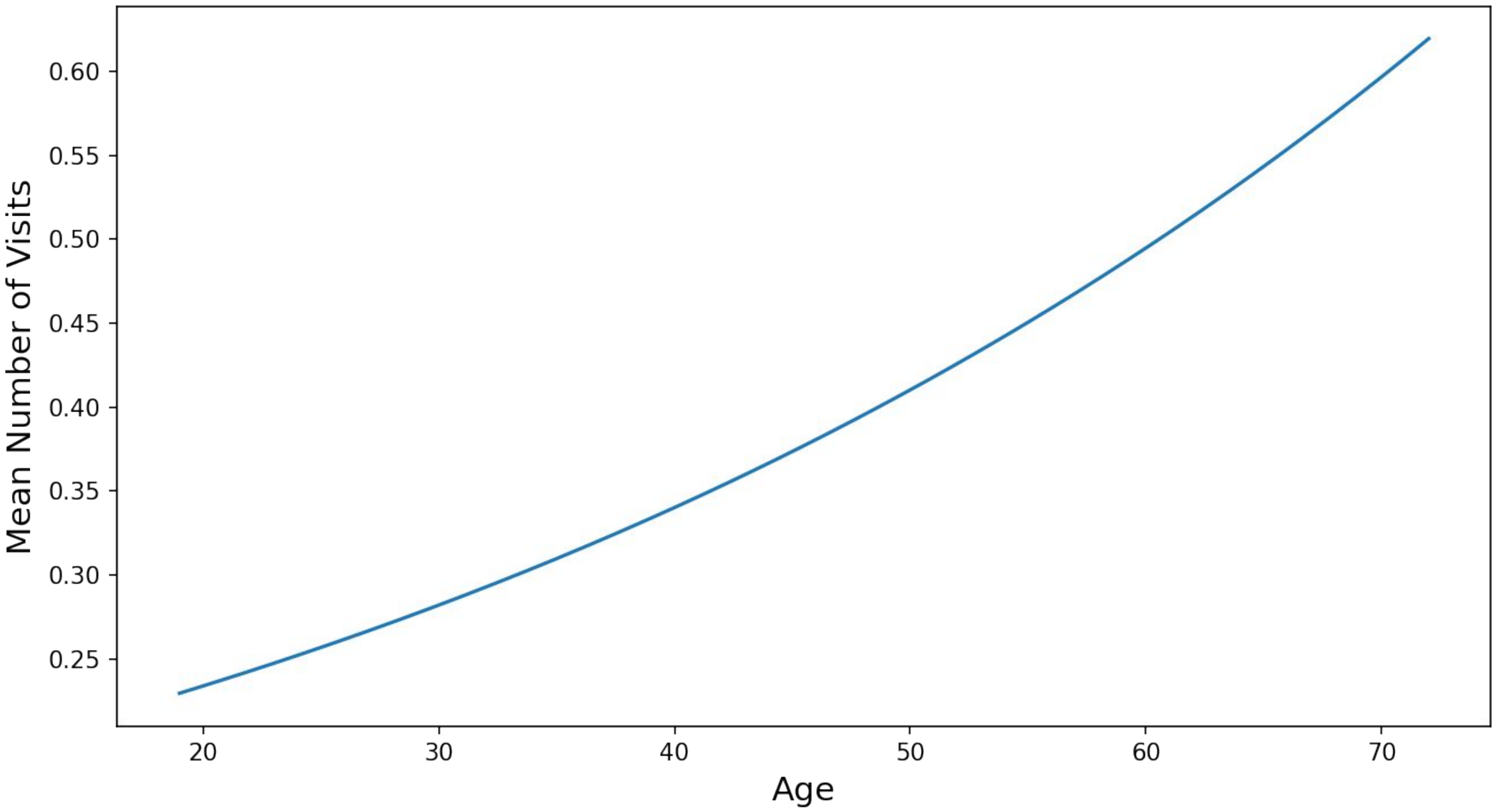
Dep. Variable:	visits	**No. Observations:**	100
Model:	GLM	**Df Residuals:**	98
Model Family:	Poisson	**Df Model:**	1
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-88.646
Date:	Thu, 16 Sep 2021	**Deviance:**	120.45
Time:	11:26:18	**Pearson chi2:**	221.
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8280	0.441	-4.143	0.000	-2.693	-0.963
age	0.0187	0.008	2.396	0.017	0.003	0.034

For a person whose age is t, the estimated value of the mean is

$$exp(-1.8280 + 0.0187t)$$

$$= e^{(-1.8280 + 0.0187t)}$$



Mean Number of Visits
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
20
30
40
50
60
70
Age

Poisson Regression

$Y|\vec{x}$ follows a $\boxed{\text{Poisson}}$ distribution with mean

$$\mu = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)$$

$$= e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}$$

Example: MLB Data

```
mlb.head()
```

	runs	games	OBP
0	582	139	0.310
1	671	137	0.320
2	566	137	0.304
3	716	141	0.328
4	602	140	0.306

Goal: Estimate the number of runs scored (*runs*) based on the on-base percentage (*OBP*).

Example: MLB Data

```
1 mlb.head()
```

	runs	games	OBP
0	582	139	0.310
1	671	137	0.320
2	566	137	0.304
3	716	141	0.328
4	602	140	0.306

Goal: Estimate the number of runs scored (*runs*) based on the on-base percentage (*OBP*).

Problem: Teams have played a different numbers of games, so the comparison is not really fair.

Example: MLB Data

```
1  mlb.head()
```

	runs	games	OBP
0	582	139	0.310
1	671	137	0.320
2	566	137	0.304
3	716	141	0.328
4	602	140	0.306

Goal: Estimate the number of runs scored (*runs*) based on the on-base percentage (*OBP*).

Problem: Teams have played a different numbers of games, so the comparison is not really fair.

Potential Fix: Estimate the number of runs scored *per game*.

$$\frac{runs}{games} = e^{\beta_0 + \beta_1 (OBP)}$$

$$\frac{runs}{games} = e^{\beta_0 + \beta_1(OBP)}$$

$$\Rightarrow \quad \log\left(\frac{runs}{games}\right) = \beta_0 + \beta_1(OBP)$$

$$\frac{runs}{games} = e^{\beta_0 + \beta_1 (OBP)}$$

$$\Longrightarrow \quad \log\left(\frac{runs}{games}\right) = \beta_0 + \beta_1 (OBP)$$

$$\Longrightarrow \log(runs) - \log(games) = \beta_0 + \beta_1 (OBP)$$

$$\frac{runs}{games} = e^{\beta_0+\beta_1(OBP)}$$

$$\Rightarrow \quad \log\left(\frac{runs}{games}\right) = \beta_0 + \beta_1(OBP)$$

$$\Rightarrow \log(runs) - \log(games) = \beta_0 + \beta_1(OBP)$$

$$\Rightarrow \log(runs) = \beta_0 + \beta_1(OBP) + \log(games)$$

$$\frac{runs}{games} = e^{\beta_0+\beta_1(OBP)}$$

$$\Rightarrow \quad \log\left(\frac{runs}{games}\right) = \beta_0 + \beta_1(OBP)$$

$$\Rightarrow \quad \log(runs) - \log(games) = \beta_0 + \beta_1(OBP)$$

$$\Rightarrow \quad \log(runs) = \beta_0 + \beta_1(OBP) + \boxed{\log(games)}$$

This is called an *offset* term.

```
import statsmodels.api as sm
import numpy as np

mlb_poisson = (sm.GLM(endog = mlb['runs'],
                      exog = sm.add_constant(mlb['OBP']),
                      family = sm.families.Poisson(),
                      offset = np.log(mlb['games']))
               .fit()
              )
```

```
import statsmodels.api as sm
import numpy as np

mlb_poisson = (sm.GLM(endog = mlb['runs'],
                      exog = sm.add_constant(mlb['OBP']),
                      family = sm.families.Poisson(),
                      offset = np.log(mlb['games']))
               .fit()
              )
```

We'll be using the *statsmodels* library and the *numpy* library for the logarithm.

```
import statsmodels.api as sm
import numpy as np

mlb_poisson = (sm.GLM(endog = mlb['runs'],
                      exog = sm.add_constant(mlb['OBP']),
                      family = sm.families.Poisson(),
                      offset = np.log(mlb['games']) )
               .fit()
              )
```

Specify the offset column.

```
mlb_poisson.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	runs	**No. Observations:**	30
Model:	GLM	**Df Residuals:**	28
Model Family:	Poisson	**Df Model:**	1
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-159.09
Date:	Thu, 16 Sep 2021	**Deviance:**	69.995
Time:	12:07:38	**Pearson chi2:**	69.8
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.5456	0.205	-2.663	0.008	-0.947	-0.144
OBP	6.4849	0.646	10.046	0.000	5.220	7.750

```
mlb_poisson.summary()
```

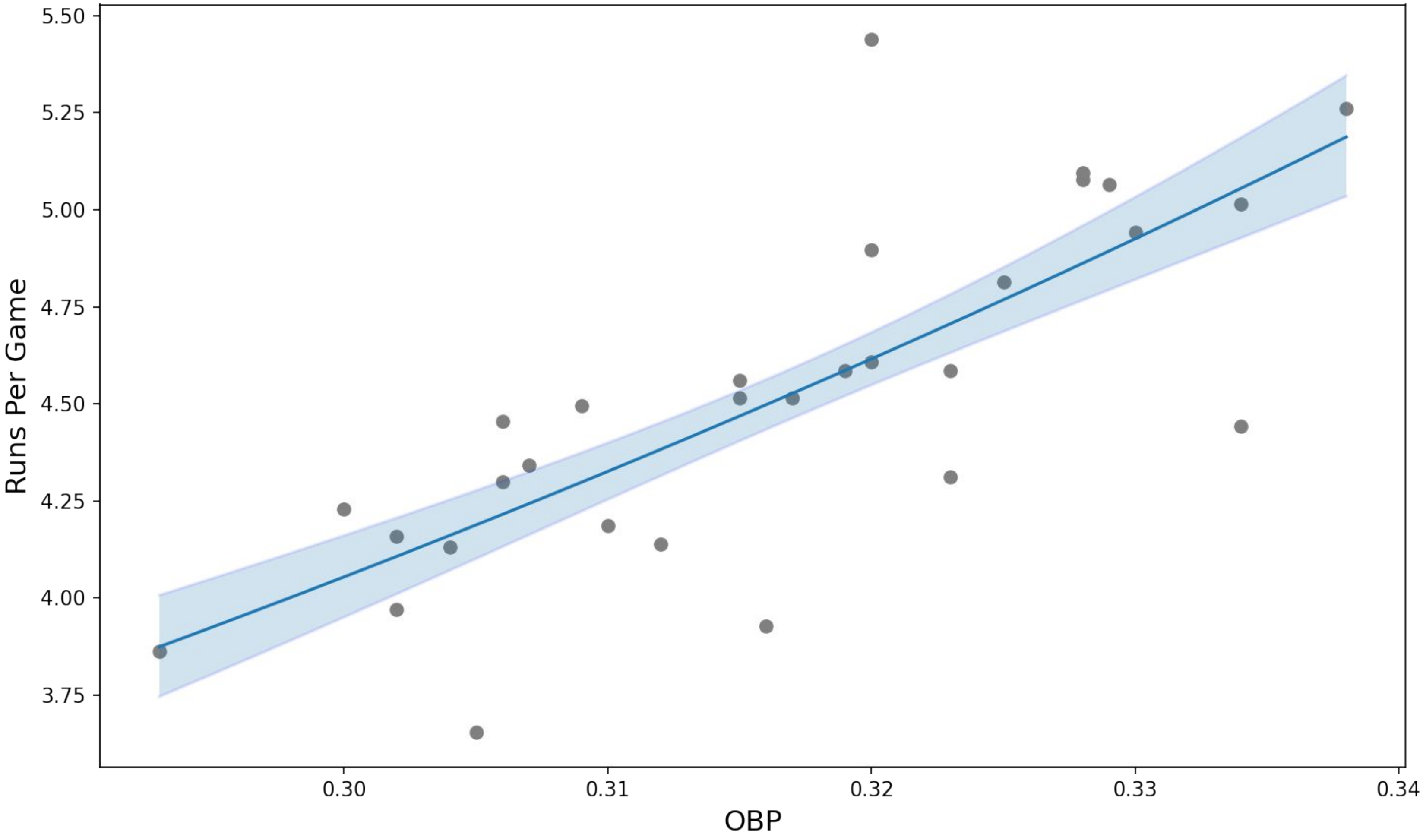
Generalized Linear Model Regression Results

Dep. Variable:	runs	**No. Observations:**	30
Model:	GLM	**Df Residuals:**	28
Model Family:	Poisson	**Df Model:**	1
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-159.09
Date:	Thu, 16 Sep 2021	**Deviance:**	69.995
Time:	12:07:38	**Pearson chi2:**	69.8
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.5456	0.205	-2.663	0.008	-0.947	-0.144
OBP	6.4849	0.646	10.046	0.000	5.220	7.750

Given the OBP of a team, the model estimates the mean runs per game as

$= e^{(-0.5456 + 6.4849(OBP))}$



5.50
5.25
5.00
4.75
4.50
4.25
4.00
3.75
Runs Per Game
0.30
0.31
0.32
0.33
0.34
OBP