Introduction to Generalized Linear Models

Part 3: Poisson Regression

Linear Regression - Continuous Target

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

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Linear Regression - Continuous Target

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

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Logistic Regression - Binary Target

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

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

Logistic Regression - Binary Target

$Y|\vec{x}$ follows a Bernoulli distribution with mean

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

Logistic Regression - Binary Target

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

$$\mu = \text{logistic}(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)$$

Logistic Regression - Binary Target

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

$$\mu = \text{logistic}(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

What if our target variable is a count?

Example: Here, our target is the number of times an individual visited the doctor in the last two weeks.

```
1 doctor_visits.head()
```

	visits	gender	age	income	illness	reduced	healtl
0	1	female	72.0	0.25	4	7	3
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	(

Example: Here, our target is the number of times an individual visited the doctor in the last two weeks.

```
doctor_visits.head()
```

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

Since this is count data, we can try using a **Poisson distribution** to model the target variable.

Example: Here, our target is the number of times an individual visited the doctor in the last two weeks.

```
doctor_visits.head()
```

	visits	gender	age	income	illness	reduced	healt
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	

Since this is count data, we can try using a **Poisson distribution** to model the target variable.

We just need to estimate the mean of this distribution.

Example: Here, our target is the number of times an individual visited the doctor in the last two weeks.

```
1  doctor_visits.head()
```

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

We'll estimate the number of visits based on the *age* variable.

Example: Estimate using the *age* variable.

```
poisreg_age = (sm.GLM(endog = doctor_visits['visits'],
                      exog = sm.add_constant(doctor_visits[['age']]),
                      family = sm.families.Poisson())
               .fit()
              )
```

Example: Estimate using the *age* variable.

```
poisreg_age = (sm.GLM(endog = doctor_visits['visits'],
                      exog = sm.add_constant(doctor_visits[['age']]),
                       family = sm.families.Poisson())
               .fit()
              )
```

We are modeling the target variable as a following a Poisson distribution.

Example: Estimate using the *age* variable.

```
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

Example: Estimate using the *age* variable.

```
poisreg_age.summary()
```

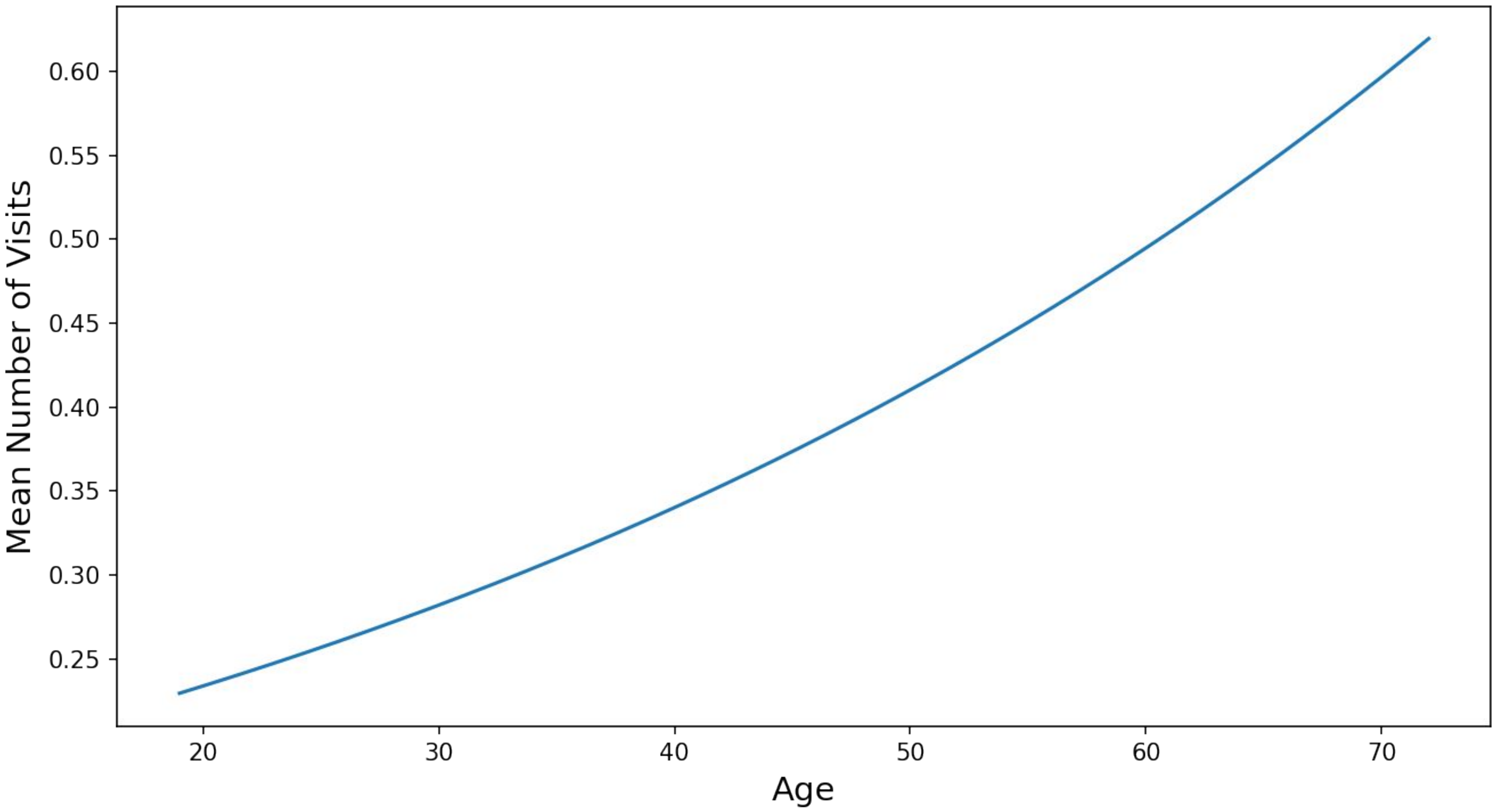
Generalized Linear Model Regression Results

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

Summary - Linear, Logistic, and Poisson Regression

Linear Regression

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

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Logistic Regression

$Y|\vec{x}$ follows a Bernoulli distribution with mean

$$\mu = \text{logistic}(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)$$

Logistic Regression

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

$$\mu = \text{logistic}(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Poisson Regression

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

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

Poisson Regression

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

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

Poisson Regression

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

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

Poisson Regression

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

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

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

To Be Continued