

STATISTICS 538, Lecture #4

Modelling Count Data

November 3, 2010

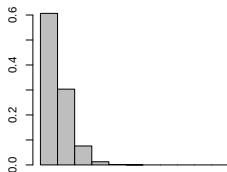
Y_i is the **count** of something for the i -th unit

First modelling thought? $Y_i|X_i \sim \text{Poisson}(\mu_i)$

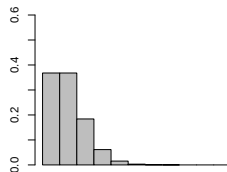
Rationale?

Aside: what do Poisson distributions look like?

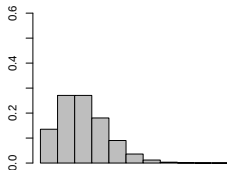
$\mu = 0.5$



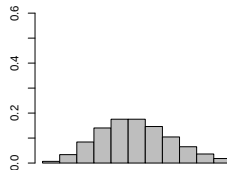
$\mu = 1$



$\mu = 2$



$\mu = 5$



In GLM formulation

Have **variance function** $v(\mu) = \mu$

Need **link function** $g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1}$

Canonical/default:

In business

Quick look at Galapagos data ex.

```
modp <- glm(Species ~ ., family=poisson, gala)
```

```
> summary(modp)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-8.275	-4.497	-0.944	1.917	10.185

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.1548079	0.0517495	60.96	< 2e-16
Area	-0.0005799	0.0000263	-22.07	< 2e-16
Elevation	0.0035406	0.0000874	40.51	< 2e-16
Nearest	0.0088256	0.0018213	4.85	0.0000013
Scruz	-0.0057094	0.0006256	-9.13	< 2e-16
Adjacent	-0.0006630	0.0000293	-22.61	< 2e-16

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3510.73 on 29 degrees of freedom
Residual deviance: 716.85 on 24 degrees of freedom
AIC: 889.7

Number of Fisher Scoring iterations: 5

Galapagos, continued

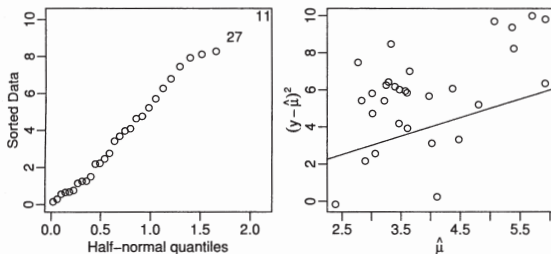


Figure 3.2 Half-normal plot of the residuals of the Poisson model is shown on the left and the relationship between the mean and variance is shown on the right. A line representing mean equal to variance is also shown.

Leisurely look at some data on deaths from coronary heart disease

age	Smokers		Non-smokers	
	deaths	person-years	deaths	person-years
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

```
glm(deaths ~ I(log(pyyears)) + age.grp + smoke, family =  
poisson)
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-21.4739	2.2556	-9.520	< 2e-16	***
I(log(pyyears))	2.4352	0.2270	10.727	< 2e-16	***
age.grp	1.7702	0.1542	11.478	< 2e-16	***
smoke	-1.6991	0.3548	-4.789	1.68e-06	***

Null deviance: 644.269 on 9 degrees of freedom
Residual deviance: 25.576 on 6 degrees of freedom

```
glm(deaths ~ offset(log(pyears)) + age.grp + smoke,  
family = poisson)
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-7.28250	0.12076	-60.304	< 2e-16	***
age.grp	0.83583	0.02904	28.777	< 2e-16	***
smoke	0.40637	0.10720	3.791	0.00015	***

Null deviance: 935.067 on 9 degrees of freedom
Residual deviance: 69.182 on 7 degrees of freedom

```
glm(deaths ~ offset(log(pyears)) + age.grp + I(age.grp^2)
+ smoke, family = poisson)
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-7.93363	0.16395	-48.391	< 2e-16	***
age.grp	1.70594	0.12824	13.303	< 2e-16	***
I(age.grp^2)	-0.19438	0.02715	-7.159	8.14e-13	***
smoke	0.35452	0.10737	3.302	0.00096	***

Null deviance: 935.067 on 9 degrees of freedom
Residual deviance: 12.176 on 6 degrees of freedom

```
glm(deaths ~ offset(log(pyears)) + age.grp + I(age.grp^2)
+ smoke + smoke:age.grp, family = poisson)
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-8.61296	0.29172	-29.524	< 2e-16	***
age.grp	1.98113	0.16025	12.363	< 2e-16	***
I(age.grp^2)	-0.19768	0.02737	-7.223	5.08e-13	***
smoke	1.13342	0.28077	4.037	5.42e-05	***
age.grp:smoke	-0.30755	0.09704	-3.169	0.00153	**

Null deviance: 935.0673 on 9 degrees of freedom
Residual deviance: 1.6354 on 5 degrees of freedom

Fitted Values

