Regression Models
Lecture VIII: Generalised Linear Models II

DT9002: Postgraduate Certificate in Applied Statistics

Dr Joe Condon

School of Mathematical Sciences
Technological University Dublin

©J. Condon 2019

Customised Hypothesis tests

These operate the same way as for linear models via a General Linear Hypothesis (GLH) type set-up. For example we have the following for the titanic data:

```
> summary(fit)

Call:
glm(formula = survived ~ age + factor(sex) + factor(pclass),
    family = binomial(), data = titanic)
...
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.777013   0.401123   9.416  < 2e-16
age             -0.036985   0.007656  -4.831 1.36e-06
factor(sex)male -2.522781   0.207391 -12.164  < 2e-16
factor(pclass)2 -1.309799   0.278066  -4.710 2.47e-06
factor(pclass)3 -2.580625   0.281442  -9.169  < 2e-16
```

Questions:

1. How dose 2nd class compare to 3rd class (p-value and CI)?
2. Compare a woman in 3rd class with a man in second class?
3. Compare a woman in 3rd class with a man in first class?

How dose 2nd class compare to 3rd class (p-value and CI)?

```
> library(multcomp)
> L=cbind(0,0,0,1,-1)
> glh=glht(fit,linfct=L)
> summary(glh) # test
...
Linear Hypotheses:
       Estimate Std. Error z value Pr(>|z|)
1 == 0    1.271      0.244   5.207 1.92e-07 ***
> confint(glh)$confint # CI on linear predicor scale
  Estimate       lwr     upr
1 1.270826 0.7925017 1.74915

> exp(confint(glh)$confint) # CI on OR scale
  Estimate      lwr      upr
1 3.563795 2.208916 5.749716
```

Compare a woman in 3rd class with a man in second class?

```
> L=cbind(0,0,-1,-1,1)
> glh=glht(fit,linfct=L)
> summary(glh) # test
...
Linear Hypotheses:
       Estimate Std. Error z value Pr(>|z|)
1 == 0   1.2520     0.3025   4.139 3.49e-05 ***

> exp(confint(glh)$confint) # CI on OR scale
  Estimate      lwr      upr
1 3.497173 1.93297 6.327164
```

Compare a woman in 3rd class with a man in first class?

```
> L=cbind(0,0,-1,0,1)
> glh=glht(fit,linfct=L)
> summary(glh) # test

Linear Hypotheses:
       Estimate Std. Error z value Pr(>|z|)
1 == 0 -0.05784    0.30533  -0.189     0.85
(Adjusted p values reported -- single-step method)
> exp(confint(glh)$confint) # CI on OR scale
   Estimate       lwr      upr
1 0.9437968 0.5187791 1.717017
```

Model Building and Likelihood Ratio Tests

The same model building issues for GLMs arise as we looked at for linear models. But, the F test is no linger available to use - this is specific to linear models with normally distributed responses.

But we can compare models that are nested using the likelihood ratio test (LRT) statistic:

Definition (LRT):

$$-2\log\left(\frac{\hat{L}_a}{\hat{L}_b}\right) = -2\left(\log \hat{L}_a - 2\log \hat{L}_b\right)$$

where $\hat{L}_a$ is the value of the likelihood at the maximum for model a and $\hat{L}_b$ is the likelihood at the maximum for model b.

This is only useful where a is nested in b, i.e. a subset of the bigger model b.

Nested Models

Two models are nested where the all covariates in one model are also to be found in the other model, e.g.

Model 1:	logit(p_i)	$=$	β_0
Model 2:	logit(p_i)	$=$	$\beta_0 + \beta_1 x_{i1}$
Model 3:	logit(p_i)	$=$	$\beta_0 + \beta_2 x_{i2}$
Model 4:	logit(p_i)	$=$	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

Models 1-3 are nested in model 4.

Model 1 in nested in model 2, 3 and 4.

Model 2 is not nested in model 3 as the x_{i1} covariate in model 2 is not also in model 3.

Model 3 is not nested in model 2 as the x_{i2} covariate in model 3 is not also in model 2.

Nesting also occurs where the model a has the same predictors as model b, but some parameters are constrained to be particular value.

Usually though it is the case that there are less predictors in model a than in model b, i.e. some parameters in model b are constrained in model a to be zero.

For 2 models a and b, where a is nested in b it can be shown that:

$$-2\left(\log \hat{L}_a - 2\log \hat{L}_b\right) \overset{\text{under}}{\underset{H_0:}{\sim}} \chi^2_k$$

where k is the number of parameters constrained under model a.

Typically k is the number of parameters dropped from model b to make the smaller model a.

This can be used as the basis for model selection.

Use of the LRT

- The LRT are a very convenient way to test the removal of multiple parameters from a model.
- The Wald tests with a GLH type test can also be used.
- The LRT is known to have better properties when applied to finite samples compared to the Wald - i.e. if the two testing methods disagree the LRT is more trustworthy.
- The LRT however have higher requirements to be applied - i.e. a full likelihood specification for the data and nested models being compared.
- For compare non-nested models use the AIC/BIC etc.

Model Building for GLMs

- The All-possible-regressions method can be applied using log likelihood to select the best model in each subset.
- You can apply the forward, backward or stepwise algorithms using hypothesis testing via the LRTs (or Walds).
- Use the AIC (or BIC) to build a model using perhaps forward, backward or stepwise.
- Other penalty based method can be applied also, e.g. Lasso.
- Residuals, dffits, dfbetas, Cook's distance, multicollinearity measures etc. are all available with some computational changes from linear regression.

Example: Neuralgia data

Treatment	Sex	Age	Duration	Pain	Treatment	Sex	Age	Duration	Pain
P	F	68	1	0	P	F	79	20	1
B	M	74	16	0	A	M	70	12	0
P	F	67	30	0	A	F	69	12	0
P	M	66	26	1	B	F	65	14	0
B	F	67	28	0	B	M	70	1	0
B	F	77	16	0	B	M	67	23	0
A	F	71	12	0	A	M	76	25	1
B	F	72	50	0	P	M	78	12	1
B	F	76	9	1	B	M	77	1	1
A	M	71	17	1	B	F	69	24	0
A	F	63	27	0	P	M	66	4	1
A	F	69	18	1	P	F	65	29	0
B	F	66	12	0	P	M	60	26	1
A	M	62	42	0	A	M	78	15	1
P	F	64	1	1	B	M	75	21	1
A	F	64	17	0	A	F	67	11	0
P	M	74	4	0	P	F	72	27	0
A	F	72	25	0	P	F	70	13	1
P	M	70	1	1	A	M	75	6	1
B	M	66	19	0	B	F	65	7	0
B	M	59	29	0	P	F	68	27	1
A	F	64	30	0	P	M	68	11	1
A	M	70	28	0	P	M	67	17	1
A	M	69	1	0	B	M	70	22	0
B	F	78	1	0	A	M	65	15	0
P	M	83	1	1	P	F	67	1	1
B	F	69	42	0	A	M	67	10	0
B	M	75	30	1	P	F	72	11	1
P	M	77	29	1	A	F	74	1	0
A	F	69	3	0	B	M	80	21	1

A study was conducted of the analgesic effects of treatments on elderly patients with neuralgia. Compare two treatments with a placebo, but controlling for age, gender and disease duration.

To fit the full model is R we use the code;

```
  > fit_neuralgia=glm(Pain~factor(Sex)+Age+Duration+factor(
      Treatment),family=binomial(),data=neuralgia)
```

The output from R is;

```
> summary(fit_neuralgia)

Call:
glm(formula = Pain ~ factor(Sex) + Age + Duration + factor(
    Treatment),
    family = binomial(), data = neuralgia)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -20.588282   7.102883  -2.899  0.00375 **
factor(Sex)M         1.832202   0.796206   2.301  0.02138 *
Age                  0.262093   0.097012   2.702  0.00690 **
Duration            -0.005859   0.032992  -0.178  0.85905
factor(Treatment)B  -0.526853   0.937025  -0.562  0.57394
factor(Treatment)P   3.181690   1.016021   3.132  0.00174 **
```

To determine if all the variables included in this model LRTs can be performed.

Model No.	Model	p	log lik	-2 log lik	AIC
1	age	2	-36.53	73.06	77.06
2	sex	2	-37.92	75.84	79.84
3	duration	2	-39.94	79.88	83.88
4	treat	3	-33.74	67.48	73.48
5	age sex	3	-34.45	68.90	74.90
6	age duration	3	-36.24	72.48	78.48
7	age treat	4	-27.52	55.04	63.04
8	sex duration	3	-37.17	74.34	80.34
9	sex treat	4	-29.94	59.88	67.88
10	duration treat	4	-33.34	66.68	74.68
11	age sex duration	4	-34.21	68.42	76.42
12	age sex treat	5	-24.38	48.76	58.76
13	age duration treat	5	-27.52	55.04	65.04
14	sex duration treat	5	-29.61	59.22	69.22
15	age sex duration treat	6	-24.37	48.74	60.74

There are a number of model building strategies - forward, backward and stepwise as before.

One way to go is as follows:

Test for treat; compare model 15 with 11. Conclude?

Test for duration: compare 15 with 12. Conclude?

Test ?: compare 12 with 5. Conclude?

test ?: compare 12 with 7. Conclude?

test ?: compare 12 with 9. Conclude?

Applying LRTs in R

```
> fit1=glm(pain~age+duration+factor(sex)+factor(treatment),
    family=binomial(),data=neuralgia)
> drop1(fit1,test='LRT')
Single term deletions

Model:
pain ~ age + duration + factor(sex) + factor(treatment)
                  Df Deviance    AIC     LRT   Pr(>Chi)
<none>               48.736 60.736
age                1   59.213 69.213 10.4769   0.001209
duration           1   48.767 58.767  0.0317   0.858663
factor(sex)        1   55.036 65.036  6.3000   0.012074
factor(treatment)  2   68.424 76.424 19.6882 5.306e-05
```

```
> fit2=update(fit1,.~.-duration)
> drop1(fit2,test='LRT')
Single term deletions

Model:
pain ~ age + factor(sex) + factor(treatment)
                  Df Deviance    AIC     LRT   Pr(>Chi)
<none>               48.767 58.767
age                1   59.886 67.886 11.1182 0.0008548
factor(sex)        1   55.044 63.044  6.2766 0.0122340
factor(treatment)  2   68.900 74.900 20.1322   4.25e-05

> summary(fit2)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -20.86939    6.94218  -3.006  0.00265
age                  0.26496    0.09591   2.763  0.00573
factor(sex)M         1.82353    0.79195   2.303  0.02130
factor(treatment)B  -0.54741    0.93123  -0.588  0.55664
factor(treatment)P   3.17896    1.01348   3.137  0.00171
```

The Exponential Family

Linear regression is a special case of GLM. It is where the distribution of the data is assumed to be normal.

In general, GLM is concerned with regression type models for a class of probability distributions called the exponential family.

Some other members of the exponential family of distributions are:

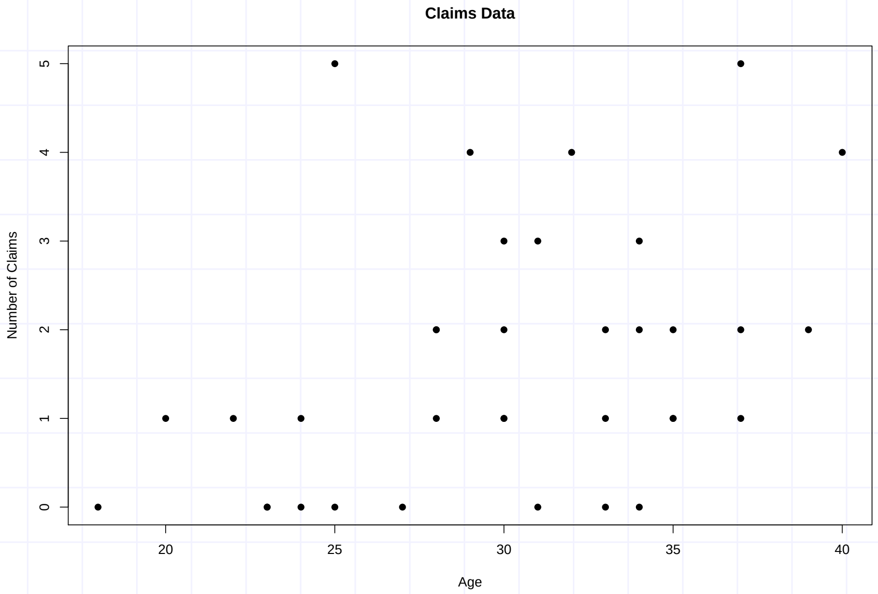
- Poisson (count of events in time, space, volume)
- Exponential/Weibull distributions (time to event/survival analysis)
- Negative binomial (Poisson with extra variance)
- Gamma distribution (positively skewed continuous data with changing variance - e.g. finance and climatology)
- Multinomial (generalises logistic regression to three or more outcomes).
- many others....

Poisson regression

A sample of data on the number of health insurance claims by customers of different ages in give in the table below.

Customer ID	Age	Number of Claims	Customer ID	Age	Number of Claims
1	18	0	19	31	0
2	20	1	20	31	3
3	22	1	21	32	4
4	23	0	22	33	2
5	23	0	23	33	0
6	24	0	24	33	1
7	24	1	25	34	2
8	25	0	26	34	3
9	25	5	27	34	0
10	27	0	28	35	1
11	28	1	29	35	2
12	28	2	30	35	1
13	28	2	31	37	2
14	29	4	32	37	5
15	30	2	33	37	1
16	30	1	34	39	2
17	30	3	35	40	4
18	30	1			

* Data from Pawitan, 2001



NB. (1) The responses are integers, (2) the variance seems to increase with the mean and, (3) the number of claims has no natural limit.

An obvious choice for distribution with these characteristics in the Poisson.

$$p(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad \lambda > 0, \qquad y \in \{0, 1, 2, 3, \ldots\}$$

The $E[Y] = \lambda$ and $Var[Y] = \lambda$

Poisson regression Model

We want to relate age to the mean response. The mean response must be positive so a natural choice is the following:

$$E[Y_i] = \lambda_i = e^{\beta_0+\beta_1(x_{i1}+\beta_2 x_{i2}+\ldots)} = e^{\eta_i} \qquad \Rightarrow \log \lambda_i = \eta_i$$

Which leads to the following likelihood (assuming independence between observations):

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\ \text{where}\lambda_i &= e^{\eta_i} \end{aligned}$$

For the claims data we might try:

$$\lambda_i = e^{\beta_0+\beta_1(age_i)}$$

Using R we get:

```
> hclaims = read.table("hclaims.txt",header=T)
> fit_hclaims=glm(claims~age,data=hclaims,family=poisson)
> summary(fit_hclaims)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.56988    0.84788  -1.852   0.0641
age          0.06626    0.02621   2.528   0.0115
```

Interpret both $\hat{\beta}_1$ and $\hat{\beta}_0$?

Hypothesis testing can proceed as for the logistic regression case, i.e. using Wald or (better) LR based tests.

Model building can be based on hypothesis testing using stage-wise algorithms, and/or AIC/AICc etc.

Poisson regression with Offset

Epilepsy data

Subject	treatment	Weeks	Attacks
1	active	12	3
2	active	5	2
3	active	7	4
4	active	14	3
5	active	10	5
6	active	10	2
7	active	12	1
8	active	8	3
9	active	11	3
10	active	8	3
11	placebo	11	4
12	placebo	11	7
13	placebo	8	6
14	placebo	16	8
15	placebo	11	11
16	placebo	7	8
17	placebo	15	7
18	placebo	9	7
19	placebo	7	4
20	placebo	4	2
21	placebo	6	6
22	placebo	4	1

- The variable follow-up time is a crucial feature of these data
- The Poisson mean is a rate in unit time - however the unit is defined.
- Therefore we need to use an **offset** term in the Poisson model to account for this.
- Such an offset term is used whenever the number of events observed is recorded from observations with different exposure to the event.

Model including Offset

$$\lambda_i = t_i e^{\eta_i} = e^{\log t_i + \eta_i}$$

where t_i is the exposure for observation i.

This leads to the following likelihood:

$$L(\beta) \quad = \quad \prod_{i=1}^{n} \frac{[e^{\log t_i + \eta_i}]^{y_i} e^{-e^{\log t_i + \eta_i}}}{y_i!}$$

This log likelihood is maximised as before, WRT to the β's - NB. not with respect to exposure.

In essence, the exposure term acts as a weight for each observation in the maximisation and enters the linear predictor with a 'known' coefficient of 1.

We implement in the software by specifying the exposure variable as a predictor with a known coefficient of one.

```
> epilepsy = read.csv("epilepsy.csv",header=T)
> epilepsy$treatment=factor(epilepsy$treatment)
> fit_epilepsy=glm(Attacks~treatment+offset(I(log(Time))),
    data=epilepsy,family=poisson)
> summary(fit_epilepsy)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.2074     0.1857  -6.502 7.92e-11
treatmentplacebo   0.7787     0.2204   3.534  0.00041
```

We can now interpret the parameters as modelling the rate of events per unit of exposure - i.e. number of attacks per week for this example.

Other topics in Regression?

Some of these are...

- Time to event analysis with incomplete data - aka survival analysis.
- Non-linear models for normally distributed data.
- Classification models with regression type structure (e.g. LDA, Cart/Rpart type models).
- Regression models with dependence between observations, e.g. medical data including observations on members of the same family (generalised least squares, linear mixed models, generalised linear mixed models, general estimating equations).
- Bayesian analysis within the regression framework.