MATH 5910

Logistic Regression and GLM

Logistic Regression

- Consider a regression model with binary response.
- Logistic regression: one possible model.
    - Popular.
    - Other methods possible.
- Independent variables: can be continuous or discrete (or both).
- Close connection to contingency tables.

Logistic Regression

Concentrate on binary response

$$Y = 0 \text{ or } 1$$

Suppose we have one independent variable X (either discrete or continuous).

How to model?

Note: May have more than one independent variables, in general.

Model

Assume that the data $Y_1, \ldots, Y_n$ are iid.

Let

$$p_i = P(Y_i = 1|X_i),$$

the probability of “success.”

Note

$$E(Y_i|X_i) = P(Y_i = 1|X_i) = p_i$$

Model

Then

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

is the logistic model.

We have the logit transform

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

hence the name logistic regression.

Model

Can be seen that

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 X_i}$$

or

$$\begin{aligned} p_i = P(Y_i = 1|X_i) &= \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} \\ &= \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}} \end{aligned}$$

Note the missing error term (why?).

Estmation

Now, both β_0 and β_1 can be estimated based on data (details postponed; can be done in R), so that

$$\hat{p}_i = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X_i)}}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of β_0 and β_1, respectively.

Or

$$\text{logit}(\hat{p}_i) = \log\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Testing

- Individual terms:

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_A : \beta \neq 0$$

- Overall fit: Goodness-of-fit, comparing proposed model against the null model - use χ^2 test.

Example 1

- Data table:

		Y		
		1	0	
X	1	7	3	10
	0	14	82	96
		21	85	106

- Y: 0 = Survive, 1 = Death
- X: 0= No Shock, 1= Shock

Original data: X and Y binary (0 or 1).

Example 1

Analysis - no R for this example.

- We have that $\hat{\beta}_0 = -1.768$ and $\hat{\beta}_1 = 2.615$.
- Then
$$\mathrm{logit}(\hat{p}) = -1.768 + 2.615X$$
- Or
$$\hat{p} = \frac{1}{1 + e^{-(-1.768 + 2.615X)}}$$

Example 1

- For $X = 0$,

$$\hat{p}(0) = \hat{P}(Y = 1|X = 0) = \frac{1}{1 + e^{-(-1.768)}} = 0.146$$

- For $X = 1$,

$$\hat{p}(1) = \hat{P}(Y = 1|X = 1) = \frac{1}{1 + e^{-(-1.768+2.615)}} = 0.7$$

Example 1

Interpretations

- $\hat{P}(Y = 1|X = 0) = 0.146$: Given that no shock was present ($X = 0$), the estimated probability that a patient dies ($Y = 1$) is 0.146.
- $\hat{P}(Y = 1|X = 1) = 0.7$: Given that shock was present ($X = 1$), the estimated probability that a patient dies ($Y = 1$) is 0.7.

Is it that complicated?

Table.

Example 1

Another interpretation:

$\hat{\beta}_1 = 2.615$ is the (estimated) log odds ratio.

How?

Example 1

Recall:

- Can transform probability into odds: odds $= p/(1-p)$
- For convenience, let $p(x) = P(Y = 1|X = x)$, and

  $$\text{odds}(x) = \frac{p(x)}{1-p(x)}$$

  so we can have odds(0) and odds(1)
- OR is the odds ratio, i.e.

  $$\text{OR} = \frac{\text{odds}(1)}{\text{odds}(0)} = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$$

Example 1

Then the log odds ratio is

$$\begin{aligned}
\log \mathrm{OR} &= \log\left(\frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}\right) \\
&= \log\left(\frac{p(1)}{1-p(1)}\right) - \log\left(\frac{p(0)}{1-p(0)}\right) \\
&= \operatorname{logit}(p(1)) - \operatorname{logit}(p(0))
\end{aligned}$$

Example 1

But since we are dealing with estimates,

$$\begin{aligned}
\log \widehat{\mathrm{OR}} &= \mathrm{logit}(\hat{p}(1)) - \mathrm{logit}(\hat{p}(0)) \\
&= \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 - \hat{\beta}_0 - \hat{\beta}_1 \cdot 0 \\
&= \hat{\beta}_1 \\
&= 2.615
\end{aligned}$$

Example 1

To verify (with our set up)

$$\widehat{\mathrm{OR}} = \frac{7 \cdot 82}{3 \cdot 14} = 13.667$$

and that

$$\log \widehat{\mathrm{OR}} = 2.615$$

as was to be seen.

Example 1

Also,

$$\hat{p}(x) = \hat{P}(Y = 1|X = x) = \frac{1}{1 + e^{-(-1.768+2.615x)}}$$

for this particular example, or

$$\hat{p}(x) = \hat{P}(Y = 1|X = x) = \frac{1}{1 + e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}}$$

in general. Hence, each change in x will affect $\hat{p}(x)$ in the above (nonlinear) fashion.

Compare this against the linear regression model.

Example 2

This time, we will use R.

- Data: Coronary heart disease (CHD, response Y) and Age (X).
- Y is either yes (Y=1) or no (Y=0).
- Goal: investigate the effect of Age on CHD.

Example 2

Read in and inspect data.

```
> ex2data<-read.table('ex2data.txt',header=T)
> ex2data
    Age CHD
1    20   0
2    23   0
3    24   0
4    25   1
:
98   64   0
99   65   1
100  69   1
```

Example 2

For logistic regression, we use `glm()`.

- The syntax is

  ```
  glm(y~x, family=binomial)
  ```

  which is very similar to `lm()`.
- In fact

  ```
  glm(y~x, family=gaussian)
  ```

  and

  ```
  lm(y~x)
  ```

  are the same!
- `glm()` is a flexible function that can handle generalized linear models.

Example 2

Fit the model

```
> ex2.glm<-glm(CHD~Age,family=binomial,data=ex2data)

> summary(ex2.glm)

Call:
glm(formula = CHD ~ Age, family = binomial, data = ex2data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859
```

Example 2

Continued.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
Age          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

Example 2

Some observations.

- Can see that $\hat{\beta}_0 = -5.30945$ and $\hat{\beta}_1 = 0.11092$.
- Both intercept and predictor significant (but with z-tests).
- What about the "deviance" business? AIC?
- Interpretation:

$$\hat{p}(x) = \frac{1}{1 + e^{-(-5.30945 + 0.11092x)}}$$

for example, at Age=50, then the probability of having CHD is $\hat{p}(50) = 0.559$.

Example 2

Prediction in R

```
> predict(ex2.glm, type = "response")
```

Example 2

Or for a particular X (Age) value,

```
> predict(ex2.glm, newdata=data.frame(Age=50),
          type = "response")
        1
0.5588765
```

or arbitrary Age (range)

```
> predict(ex2.glm,newdata=data.frame(Age=seq(1,100)),
          type = "response")
```

Notice the options used.

Example 2

See the help file

```
> ?predict.glm
```

Example 2

Try `anova()`.

```
> anova(ex2.glm)
```

Example 2

Add test option.

```
> anova(ex2.glm, test='Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: CHD

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                   99     136.66
Age   1    29.31       98     107.35 6.168e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example 3

Data

```
>xy.data<-read.table("xy.data.txt",header=T)

> xy.data
   x y
1  1 2
2  2 2
3  1 2
4  1 1
:
24 2 1
```

Example 3

And

```
>  xy.table<-table(xy.data)
```

```
> xy.table
   y
x    1 2
  1 6 7
  2 7 4
```

Example 3

Odds Ratio (OR)

```
> (xy.table[1,1]*xy.table[2,2])/(xy.table[1,2]*xy.table[2,1])
[1] 0.4897959
```

Example 3

Testing

```
> chisq.test(xy.table)

Pearson's Chi-squared test with Yates'
continuity correction

data:  xy.table
X-squared = 0.1983, df = 1, p-value = 0.656
```

Example 3

Now, fit the logistic regression.

```
> xy.glm<-glm(factor(y)~factor(x),family=binomial,
              data=xy.data)

> summary(xy.glm)

Call:
glm(formula = factor(y) ~ factor(x), family = binomial,
    data = xy.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2435  -1.0240  -0.9508   1.1127   1.4224
```

Example 3

Continued.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1542     0.5563   0.277    0.782
factor(x)2   -0.7138     0.8381  -0.852    0.394

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33.104  on 23  degrees of freedom
Residual deviance: 32.365  on 22  degrees of freedom
AIC: 36.365

Number of Fisher Scoring iterations: 4
```

Example 3

- Note that $\hat{\beta} = -0.7138$.
- Recall that this number can be interpreted as log odds ratio.
- Hence,

  ```
  > exp(-0.7138)
  [1] 0.4897795
  ```

  i.e., $e^{-0.7138} = 0.4897795$, which is very close to 0.4897959 found previously.
- (It turns out that the actual value of $\hat{\beta}$ is -0.7137665, in which case $e^{-0.7137665} = 0.4897959$.)

Example 3

Let us also see the “ANOVA” table.

```
> anova(xy.glm, test='Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: factor(y)

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                         23     33.104
factor(x)  1  0.73878        22     32.365   0.3901
```

Extensions

More than one independent variable?

- For 2 variables X_1, X_2
- We have, suppressing i and writing $p(x_1, x_2) = p_i$,

$$p(x_1, x_2) = P(Y = 1 | X_1 = x_1, X_2 = x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

- Since

$$\log\left(\frac{p(x_1, x_2)}{1 - p(x_1, x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Extensions

- In general, if $\boldsymbol{x} = (x_1, \ldots, x_p)'$
- Then

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Which implies

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

- Obtain $\hat{p} = \hat{p}(\boldsymbol{x})$ by replacing β's with $\hat{\beta}$'s
- Other modifications (polynomial, interaction, etc.) possible

Extensions

Matrix notations:

- If $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_{\boldsymbol{p}})$, then

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}$$

- So that

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x})}}$$

More Extensions

What if Y has more than 2 categories?

- If $Y = 1, 2, \ldots, k$ ordered (i.e., $1 < 2 < \cdots < k$), then use ordinal logistic regression.
- Otherwise, use nominal logistic regression.
- Very complicated...

Example 4

Read in data:

```
> ex4data<-read.table(’ex4data.txt’,header=T)
> ex4data
   Y    X1    X2
1  1  6.77 13.03
2  1 15.03 13.03
3  1  6.43 10.38
4  0  0.10 13.15
5  0  0.10 12.11
:
64 1  1.02 11.41
```

Example 4

Logistic regression: Start with

```
> ex4.glm<-glm(Y~X1*X2,family=binomial,data=ex4data)
> summary(ex4.glm)
> anova(ex4.glm, test='Chisq')
```

Output suppressed. . .

Example 4

Additive Model

```
> summary(glm(Y~X1+X2,family=binomial,data=ex4data))

Call:
glm(formula = Y ~ X1 + X2, family = binomial, data = ex4data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7043  -0.9583   0.1589   1.0026   1.5043
```

Example 4

Continued

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.05339    1.06941   0.985   0.3246
X1           0.21059    0.08729   2.413   0.0158 *
X2          -0.16126    0.09858  -1.636   0.1019
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 88.473  on 63  degrees of freedom
Residual deviance: 73.626  on 61  degrees of freedom
AIC: 79.626

Number of Fisher Scoring iterations: 5
```

Example 4

Hence,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.05339 + 0.21059X_1 - 0.16126X_2$$

or

$$\hat{p} = \frac{1}{1+e^{-(1.05339+0.21059X_1-0.16126X_2)}}$$

Should be clear from the context how to interpret the results.

Classification

- Two class (Y=0 or 1)
- With input $X_1, \ldots, X_p$
- Predict which class your data belongs to (0 or 1)
  - First, need to model with known Y
  - Then, predict membership of Y, using inputs
  - Can assess performance of your model (error rates)

Classification

- For example
  - Y with 0=normal, 1=disease
  - $X_1 = \text{Age}$, $X_2 = \text{BMI}$, $X_3 = \text{BP}$
  - At which Age, BMI, BP that make the person classified as diseased?
- Many methods
- Logistic regression possible

Example

Recall Example 2

- Data: Coronary heart disease (CHD, response Y) and Age (X).
- Y is either yes (Y=1) or no (Y=0).
- Goal: investigate the effect of Age on CHD.

Example

Where

- Estimates $\hat{\beta}_0 = -5.30945$ and $\hat{\beta}_1 = 0.11092$.
- And
  $$\hat{p}(x) = \frac{1}{1 + e^{-(-5.30945 + 0.11092x)}}$$
  so that, at Age=50, the probability of having CHD is $\hat{p}(50) = 0.559$.
- At Age=40, the probability of having CHD is $\hat{p}(40) = 0.295$.

Example

- Cutoff at 0.5
    - Classify as Diseased (CHD) if $\hat{p}(x) > 0.5$
    - Classify as Normal if $\hat{p}(x) < 0.5$
- At Age=47, $\hat{p}(47) = 0.476$
- At Age=48, $\hat{p}(48) = 0.504$
- If 48 year old or older, then classified as CHD.
- NOT realistic.
- May want more inputs (X)

Several Inputs

- Back to
    - Y with 0=normal, 1=disease
    - X_1 = Age, X_2 = BMI, X_3 = BP
    - At which Age, BMI, BP that make the person classified as diseased?
- Suppose that $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.02$, $\hat{\beta}_2 = 0.06$, $\hat{\beta}_3 = 0.03$
- Then

$$\hat{p} = \hat{p}(x_1, x_2, x_3) = \frac{1}{1 + e^{-(-6+0.02x_1+0.06x_2+0.03x_3)}}$$

Several Inputs

Try different combinations of (Age, BMI, BP)

- If $(30, 20, 120)$ then $\hat{p} = 0.354$
- If $(50, 20, 120)$ then $\hat{p} = 0.450$
- If $(30, 30, 120)$ then $\hat{p} = 0.5$
- If $(50, 30, 120)$ then $\hat{p} = 0.6$
- If $(30, 20, 150)$ then $\hat{p} = 0.574$
- If $(50, 20, 150)$ then $\hat{p} = 0.668$

Many combinations to consider, interactions also possible.

Many Inputs

- More input variables (p) than number of participants (n)
- Modern research, more realistic
- Problem: Logistic regression cannot handle $p > n$
- Specialized methods

Many Inputs

For logistic regression

- Cannot use all $X_1, \ldots, X_p$ inputs if $p > n$
- Work around
- Reduce dimension
  - Principal component analysis (PCA)
  - Regularized methods (Ridge, LASSO, Elastic Net)
- Result with $X_1^*, \ldots, X_k^*$, where $k < n$

More on Logistic Regression

Some unanswered questions

- How are the parameter estimation and testing done?
- Deviance?
- Why is the R function called the `glm()`?
- Best answered in terms of generalized linear models (GLM)

GLM

- Flexible modeling technique that includes many major/popular regression methods.
- Linear regression, logistic regression.
- Unified theory.

GLM

Recall, for the (simple) logistic regression,

$$E(Y_i|X_i) = P(Y_i = 1|X_i) = p_i$$

and

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

where

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

GLM

Whereas for the simple regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

so that we have

$$E(Y_i | X_i) = \beta_0 + \beta_1 X_i$$

(why?)

GLM

More generally, we can use the matrix form to see that

$$E(\boldsymbol{Y}) = \boldsymbol{p}$$

and

$$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{X}\beta$$

where

$$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \left(\log\left(\frac{p_1}{1-p_1}\right), \ldots, \log\left(\frac{p_n}{1-p_n}\right)\right)'$$

for the logistic regression

GLM

And,

$$E(\boldsymbol{Y}) = \boldsymbol{X}\beta$$

for the (normal) linear regression (note that we let $E(\boldsymbol{Y}) \equiv E(\boldsymbol{Y}|\boldsymbol{X})$ for convenience).

Any connections?

GLM Components

The GLM will have 3 main components

1. The random component: The data $\boldsymbol{Y}$, which is random with a distribution, and $E(\boldsymbol{Y}) = \boldsymbol{\mu}$.
2. The systematic component :

$$\boldsymbol{\eta} = \boldsymbol{X}\beta$$

3. The link function: A function $g(\cdot)$ that links $\boldsymbol{\mu}$ with $\boldsymbol{\eta}$,

$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

(technically, $\eta_i = g(\mu_i)$).

GLM Components

For example,

- For logistic regression, $\boldsymbol{Y}$ has the binomial distribution with $E(\boldsymbol{Y}) = \boldsymbol{p}$, and

  $$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = g(\boldsymbol{p}) = \boldsymbol{\eta} = \boldsymbol{X}\beta$$

  so that the link $g(\cdot)$ is the logit function.
- For linear regression, $\boldsymbol{Y}$ is normal with $E(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{X}\beta$, and

  $$\boldsymbol{\mu} = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\beta$$

  so that the link $g(\cdot)$ is the identity function.

GLM Components

- The GLM is the general method of regression that includes many regression models as special cases.
- In the previous two cases, they possess all the components to be a part of the GLM.
- More examples upcoming.

GLM Components

Another component

- Have not discusses the variance of $\boldsymbol{Y}$, yet.
- If the variance of $\boldsymbol{Y}$ can be written in terms of $\boldsymbol{\mu} = E(\boldsymbol{Y})$.
- Then, we may have

$$\text{var}(Y_i) = a(\phi)V(\mu_i)$$

where

  - The $V(\cdot)$ is the variance function (as it relates to μ)
  - The ϕ is the dispersion parameter.

Exponential Family

- Assume that Y has the pdf of the form

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

- If ϕ is known, then Y belong to an exponential family with canonical parameter θ.
- If ϕ is unknown, then ?
- For θ, it is a function of $\mu = E(Y)$, which in turn is a function of β.

Likelihood

- Key to understanding inference of GLM.
- Have the pieces to get started.
- Heavy and messy topic - only essentials covered.

Likelihood

- For individual Y_i,

$$f(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

- Likelihood

$$L = \prod_{i=1}^{n} f(y_i;\theta_i,\phi) = \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

(NOTE: Depending on situation, we may use any one of

$$L,\ L(\boldsymbol{\theta},\phi;\boldsymbol{y}),\ L(\boldsymbol{\mu};\boldsymbol{y}),\ L(\boldsymbol{\beta})$$

which are all equivalent).

Likelihood

- The log-likelihood

$$\ell = \log L = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right)$$

- Same comment above applies for the notation of ℓ.

Estimation

Then

- Solve for β_i from the equation

$$\frac{\partial \ell}{\partial \beta_i} = 0$$

  to obtain $\hat{\beta}_i$ (and $\hat{\boldsymbol{\beta}}$).

- Note the chain rule

$$\frac{\partial \ell}{\partial \beta_i} = \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i}$$

Estimation

- For example, in linear regression,

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- However, most other GLM models won't have a closed form solution.
- Need to solve for $\hat{\beta}$ iteratively.

Estimation

One algorithm

- The score function:

  $$s(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \ell}{\partial \beta_i} \right)$$

  ($(p+1) \times 1$ vector).
- The Fisher information (expected information matrix):

  $$I(\boldsymbol{\beta}) = \left(-E \left(\frac{\partial^2 \ell}{\partial \beta_i \partial \beta_j} \right) \right)$$

  ($(p+1) \times (p+1)$ matrix)

Estimation

Fisher scoring

- The algorithm

  $$\beta^{(t+1)} = \beta^{(t)} I^{-1}(\beta^{(t)}) s(\beta^{(t)})$$

  where $\beta^{(t)}$ is the current estimate of β at the kth step.
- Iterate until convergence, typically very quick.
- Look familiar?
- Other algorithms possible, but all iterative.
- In R, the `glm()` uses another algorithm by default (although very much related to Fisher scoring above).

Estimation

Once we have obtained the $\hat{\boldsymbol{\beta}}$,

- For covariance (assuming that $\phi = 1$), we can write

  $$I(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

  where $\boldsymbol{W}$ is the diagonal matrix with elements

  $$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{\text{var}(Y_i)}$$

- Then we have that

  $$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$$

  where $\hat{\boldsymbol{W}}$ is $\boldsymbol{W}$ evaluated at $\hat{\boldsymbol{\beta}}$.

Testing

We can then test for

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_A : \beta_i \neq 0$$

by using the test statistic

$$z_i = \frac{\hat{\beta}_i}{\widehat{\text{se}}(\hat{\boldsymbol{\beta}})} = \frac{\hat{\beta}_i}{(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1/2}}$$

which approximately follows standard normal under H_0. (Note that linear regression still follows the t-test, as before).

Deviance

Recall that R outputs had something called the deviance.

- What is it?
  - Somewhat similar to model selection.
  - Deviance: Compare a proposed model versus the “saturated” model.
  - Then: Compare the proposed model deviance against the “null” model deviance.
- How to compute it?

Deviance

Saturated model

- When all observation has a parameter each - perfect fit.
- In other words, $\boldsymbol{\mu} = \boldsymbol{y}$
- Compare against a proposed model to get a deviance.

Deviance

Then

$$-2\log\frac{\text{ML for the (proposed) model}}{\text{ML for the saturated model}} = -2[\ell(\hat{\boldsymbol{\mu}};\boldsymbol{y}) - \ell(\boldsymbol{y};\boldsymbol{y})]$$

where

- $\ell(\hat{\boldsymbol{\mu}};\boldsymbol{y})$ is the maximized log-likelihood for the proposed model.
- $\ell(\boldsymbol{y};\boldsymbol{y})$ is the maximized log-likelihood for the saturated model, where $\boldsymbol{\mu} = \boldsymbol{y}$ is substituted in $\ell(\boldsymbol{\mu};\boldsymbol{y})$

Deviance

Now,

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2\phi[\ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - \ell(\boldsymbol{y}; \boldsymbol{y})]$$

is called the (scaled) deviance. If $\phi = 1$, it is simply the deviance.

- The deviance $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ has the (approximate) χ^2 distribution with $n - (p + 1)$ degrees of freedom.
- The greater the deviance, the poorer the model fit.
- However. . .

Null Model

- We usually compare the proposed model against the null model
- Null model: A model without parameters (except may be intercept)
- Fit this model and obtain $\hat{\boldsymbol{\mu}}_0$.
- We can then obtain the null deviance

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) = -2\phi[\ell(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}) - \ell(\boldsymbol{y}; \boldsymbol{y})]$$

Null Model

The R output displays both

- Null deviance

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0)$$

- Residual (proposed model) deviance

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$$

AIC

Another output is concerned with AIC, which is defined as

$$AIC = -2\ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) + 2p$$

- Used in model selection (like variable selection in regression)
- Smaller the AIC, the better the model.
- Again, not directly used.

GLM ANOVA

Finally,

- When we run `anova` on a `glm` object, we also get deviance and residual deviances, with their differences as a chi-squared statistic.
- Where

  $$-2[\ell(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y})] = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$$

  is the difference of the two deviances, which also follows the χ^2 distribution.
- The degrees of freedom is the difference in the number of parameters of the proposed and the null model.
- Hence the χ^2-test is to see if the proposed model is favored over the null model.

Residuals

- R outputs deviance residuals.
- Used for diagnostics.
- Not covered here.

Binary Model

- Random component: Binomial
- We may have link functions other than logit.

Binary Model

Link (suppose that $\mu = p$)

- Logit (logistic regression)

$$g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

- Probit

$$g(\mu) = \Phi^{-1}(\mu)$$

  where $\Phi(\cdot)$ is the cdf of $N(0,1)$.

- Complementary log-log

$$g(\mu) = \log(-\log(1-\mu))$$

All available in R.

Example 5

Continue with Example 4

```
> #ex4.glm<-glm(Y~X1*X2,family=binomial,data=ex4data)
> summary(ex4.glm)
> anova(ex4.glm, test='Chisq')
```

The default link is logit, so that the above is equal to

```
> ex4.glm1<-glm(Y~X1*X2,family=binomial(link="logit"),data=ex4data)
> summary(ex4.glm1)
> anova(ex4.glm1, test='Chisq')
```

Example 5

Probit link:

```
> ex4.glm2<-glm(Y~X1*X2,family=binomial(link="probit"),
                 data=ex4data)
> summary(ex4.glm2)
> anova(ex4.glm2, test='Chisq')
```

Complementary log-log link:

```
> ex4.glm3<-glm(Y~X1*X2,family=binomial(link="cloglog"),
                 data=ex4data)
> summary(ex4.glm3)
> anova(ex4.glm3, test='Chisq')
```

Log Linear Model

A GLM with

- Poisson random component

$$Y \sim Poisson(\mu)$$

- Link: log

$$g(\mu) = \log(\mu)$$

- We have that, with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$,

$$\log(\boldsymbol{\mu}) = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\beta$$

  where

$$\log(\boldsymbol{\mu}) = (\log(\mu_1,) \ldots, \log(\mu_n))'$$

Log Linear Model

- Data Y: Counts
- Independent variable(s)
  - May be discrete or continuous, or both
  - If independent variable discrete: Can see relationship with contingency table

Very similar to the logistic regression in the set up.

Example 6

From `help(glm)`

```
> counts <- c(18,17,15,20,10,20,25,13,12)
> outcome <- gl(3,1,9)
> treatment <- gl(3,3)
> data.frame(treatment, outcome, counts))
  treatment outcome counts
1         1       1     18
2         1       2     17
3         1       3     15
4         2       1     20
5         2       2     10
6         2       3     20
7         3       1     25
8         3       2     13
9         3       3     12
```

Example 6

Run the log linear model (with `family=poisson`)

```
> glm.D93 <- glm(counts ~ outcome + treatment,
                 family = poisson())
> anova(glm.D93)
Analysis of Deviance Table

Model: poisson, link: log
Response: counts
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                          8    10.5814
outcome    2   5.4523         6     5.1291
treatment  2   0.0000         4     5.1291
```

Example 6

Running `summary()`

```
> summary(glm.D93)

Call:
glm(formula = counts ~ outcome + treatment,
    family = poisson())

Deviance Residuals:
       1         2         3         4         5         6         7         8
-0.67125   0.96272  -0.16965  -0.21999  -0.95552   1.04939   0.84715  -0.09167
```

Example 6

Continued

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.045e+00  1.709e-01  17.815   <2e-16 ***
outcome2    -4.543e-01  2.022e-01  -2.247   0.0246 *  
outcome3    -2.930e-01  1.927e-01  -1.520   0.1285    
treatment2   1.338e-15  2.000e-01   0.000   1.0000    
treatment3   1.421e-15  2.000e-01   0.000   1.0000    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10.5814  on 8  degrees of freedom
Residual deviance:  5.1291  on 4  degrees of freedom
AIC: 56.761

Number of Fisher Scoring iterations: 4
```

Example 6

- Relationship with tables?
- What if we ran ordinary (normal) ANOVA?

Example 6

Log linear model with Poisson family:

- Also called a Poisson regression

Example 7

Galapagos Dataset

```
library(alr4)

galapagos
?galapagos

summary(galapagos)
plot(galapagos)
```

Which predictors/factors contribute to the number of species?

Example 7

Fit a Poisson regression/log linear model (on select variables)

```
gala.a.poi<-glm(NS~Area, family=poisson, data=galapagos)
summary(gala.a.poi)

gala.e.poi<-glm(NS~Elevation, family=poisson, data=galapagos)
summary(gala.e.poi)

gala.ae.poi<-glm(NS~Area+Elevation, family=poisson, data=galapagos)
summary(gala.ae.poi)

gala.ane.poi<-glm(NS~Area+Anear+Elevation, family=poisson, data=galapagos)
summary(gala.ane.poi)

gala.ade.poi<-glm(NS~Area+Dist+Elevation, family=poisson, data=galapagos)
summary(gala.ade.poi)
```

Example 7

Possible to do the following (bad practice)

```
gala.all.poi<-glm(NS˜., family=poisson, data=galapagos)
summary(gala.all.poi)
step(gala.all.poi)

gala.poi<-glm(NS˜., family=poisson, data=galapagos[complete.cases(galapagos),])
summary(gala.poi)
gala.poi.step<-step(gala.poi)
summary(gala.poi.step)
anova(gala.poi.step)
```

Other Models

- Other models possible in GLM with different random component (distribution) and link.
- For example, Gamma random component with inverse link.
- Not covered here, but if interested…

More?

- Yes, there is much more to it than presented here.
- Omitted topics: plots, diagnostics, model selection, other GLM models, etc.
  - Some topics covered in future lectures.