5200
6010 - Dr. Kaiser

Generalized Linear Models

flexibility!

Definition – Generalized Linear Model

1. Random Component – specifies the probability distribution of the response variable Y; e.g., normal distribution in the classical regression model, or binomial distribution for in the binary logistic regression model.

2. Systematic Component – specifies the explanatory variables in the model, more specifically, their linear combination; e.g., $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$.

3. Link Function – specifies the link between the random and the systematic components. It indicates how the expected value of the response relates to the linear combination of explanatory variables.

Assumptions – Generalized Linear Model

1. $Y_1, \ldots, Y_n$ are independently distributed and assume a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal, etc.).
2. Linear relationship between the transformed expected response in terms of the link function and the explanatory variables exists; e.g., for binary logistic regression $logit(\pi) = \beta_0 + \beta_1 x_1$.
3. Homogeneity of variance does NOT need to be satisfied.
4. Parameter estimation uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS).

Examples – Generalized Linear Model

Simple Linear Regression

// end lecture 40
05-05-25

$$\underset{E(y_i)}{\underline{\mu_i}} = \beta_0 + \beta_1 x_i$$

traditionally

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

1. Random Component – $Y_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$
2. Systematic Component – x is the explanatory variable (can be continuous or discrete) and is linear in the parameters which can be extended to multiple linear regression with more than 1 explanatory variable; transformations of x are allowed if they are linear fashion in the parameters.
3. Link Function – the identity link, $\eta = g(\mathrm{E(Y)}) = \mathrm{E(Y)}$, is used; this is the simplest link function.

Examples – Generalized Linear Model

Binary Logistic Regression models how the odds of “success” for binary response variable Y depend on a set of explanatory variables

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

1. Random Component – $Y_i \sim \mathcal{B}(n=1,\pi)$, with probability of success equal to $\mathrm{E}(Y) = \pi, i = 1, \ldots, n$
2. Systematic Component – ~~same as above.~~ same as previous example on slide 4
3. Link Function – the log-odds or logit link, $\eta = g(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, is used.

Examples – Generalized Linear Model

Poisson Regression models how the mean of a discrete (count) response variable Y depends on a set of explanatory variables

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

1. Random Component – $Y_i \sim \mathcal{P}oi(\lambda), i = 1, \ldots, n$
2. Systematic Component – – same as ~~above.~~ slide 4
3. Link Function – the log link function

Generalized Linear Model – Why?

1. flexible choices for the distribution of Y.
2. choice of link function is separate from the choice of random component, giving us more flexibility in modeling.
3. utilizes maximum likelihood estimation, so likelihood functions and parameter estimates benefit from asymptotic normal and chi-square distributions.

Generalized Linear Model – Why?

Proc Glimmix for models that include random effects.

1. inference tools and model checking that we will discuss for logistic and Poisson regression models apply for other GLMs too; e.g., Wald and Likelihood ratio tests, deviance, residuals, confidence intervals, and overdispersion.

2. There is often one procedure in a software package to capture all the models listed above, e.g. PROC GENMOD in SAS or glm() in R, etc., with options to vary the three components.

26. A Generalized Linear Model for Bernoulli Response Data

Consider the Gauss-Markov linear model with normal errors:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma \boldsymbol{I}).$$

Another way to write this model is (see slide 4)

$$\forall\; i = 1, \ldots, n, \quad y_i \sim \mathcal{N}(\mu_i, \sigma), \quad \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

and $y_1, \ldots, y_n$ are independent.

This is a special case of what is known as a generalized linear model.

Another special case is:

$$\forall\ i = 1, \ldots, n, \quad y_i \sim \text{Bernoulli}(\pi_i),$$

$E(y_i)$

$$\pi_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})},$$

and $y_1, \ldots, y_n$ are independent.

1. In each example, all responses are independent, and each response is a draw from one type of distribution whose parameters may depend on explanatory variables through a linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$.

2. The second model, for the case of a binary response, is often called a logistic regression model.

3. Binary responses are common (success/failure, survive/die, good customer/bad customer, win/lose, etc.)

4. The logistic regression model can help us understand how explanatory variables are related to the probability of "success."

Example: Disease Outbreak Study

Source: *Applied Linear Statistical Models*, 4th edition, by Neter, Kutner, Nachtsheim, Wasserman (1996)

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease under study.

Response Variable

$$y_i = 0 \text{ (person } i \text{ does not have the disease)}$$

$$y_i = 1 \text{ (person } i \text{ has the disease)}$$

Potential Explanatory Variables

- age in years
- socioeconomic status

$$\begin{aligned} 1 &= \text{upper} \\ 2 &= \text{middle} \\ 3 &= \text{lower} \end{aligned}$$

- sector (1 or 2)

Questions of Interest

The potential explanatory variables and the response were recorded for 196 randomly selected individuals.

Are any of these variables associated with the probability of disease and if so how?

We will demonstrate how to use R to fit a logistic regression model to this dataset.

Before delving more deeply into logistic regression, we will review the basic facts of the Bernoulli distribution.

$y \sim \mathrm{Bernoulli}(\pi)$ has probability mass function

$$\Pr(y = k) = f(k) = \begin{cases} \pi^k(1-\pi)^{1-k} & \text{for } k \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\Pr(y = 0) = f(0) = \pi^0(1-\pi)^{1-0} = 1 - \pi$$

and

$$\Pr(y = 1) = f(1) = \pi^1(1-\pi)^{1-1} = \pi.$$

The variance of y is a function of the mean of y:

- $\mathrm{E}(y) = \sum_{k=0}^{1} k f(k) = 0 \cdot (1-\pi) + 1 \cdot \pi = \pi$
- $\mathrm{E}(y^2) = \sum_{k=0}^{1} k^2 f(k) = 0^2 \cdot (1-\pi) + 1^2 \cdot \pi = \pi$
- $\mathrm{Var}(y) = \mathrm{E}(y^2) - [\mathrm{E}(y)]^2 = \pi - \pi^2 = \pi(1-\pi)$

The Logistic Regression Model

For $i = 1, \ldots, n,\ \ y_i \sim \text{Bernoulli}(\pi_i)$,

where

$$\pi_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

and $y_1, \ldots, y_n$ are independent.

The Logit Function

The function

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

is called the *logit function*.

The logit function maps the interval $(0,1)$ to the real line $(-\infty,\infty)$.

π is a probability, so $\log(\frac{\pi}{1-\pi})$ is the log(odds), where the odds of an event $A \equiv \frac{\Pr(A)}{1-\Pr(A)}$.

Note that

$$\begin{aligned} g(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \log\left[\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})} \Bigg/ \frac{1}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\right] \\ &= \log[\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})] = \boldsymbol{x}_i^\top\boldsymbol{\beta}. \end{aligned}$$

Thus, the logistic regression model says that,

$$y_i \sim \mathrm{Bernoulli}(\pi_i), \text{ where}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

In Generalized Linear Models terminology, the logit is called the link function because it “links” the mean of y_i (i.e., π_i) to the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$.

For Generalized Linear Models, it is not necessary that the mean of y_i be a linear function of $\boldsymbol{\beta}$.

Rather, some function of the mean of y_i is a linear function of $\boldsymbol{\beta}$.

For logistic regression, that function is

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

When the response is Bernoulli or more generally, binomial, the logit link function is one natural choice. However, other link functions can be considered.

Some common choices (that are also available in R) include the following:

1. logit:

$$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

2. probit:

$$\Phi^{-1}(\pi) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

where $\Phi^{-1}(\cdot)$ is the inverse of $\mathcal{N}(0,1)$ CDF.

3. Complementary log-log (cloglog in R):

$$\log(-\log(1-\pi)) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Although any of these link functions (or others) can be used, the logit link has some advantages when it comes to interpreting the results (as we will discuss later).

Thus, the logit link is a good choice if it can provide a good fit to the data.

The log likelihood function for logistic regression is

$$
\begin{aligned}
\ell(\boldsymbol{\beta} \mid \boldsymbol{y}) &= \sum_{i=1}^{n} \log[\pi_i^{y_i}(1-\pi_i)^{1-y_i}] \\
&= \sum_{i=1}^{n} [\underline{y_i} \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \\
&= \sum_{i=1}^{n} [y_i\{\log(\pi_i) - \log(1-\pi_i)\} + \log(1-\pi_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right] \\
&= \sum_{i=1}^{n} [y_i\, \boldsymbol{x}_i^\top \boldsymbol{\beta} - \log(1+\exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\})]
\end{aligned}
$$

Properties of log

For Generalized Linear Models, Fisher's Scoring Method is typically used to obtain an MLE for $\boldsymbol{\beta}$, denoted as $\widehat{\boldsymbol{\beta}}$.

Fisher's Scoring Method is a variation of the Newton-Raphson algorithm in which the Hessian matrix (matrix of second partial derivatives) is replaced by its expected value (-Fisher Information matrix).

For generalized Linear Models, Fisher's scoring method results in an iterative weighted least squares procedure. in 5200

The algorithm is presented for the general case in Section 2.5 of *Generalized Linear Models* 2nd Edition (1989) by McCullough and Nelder.

For sufficiently large samples, $\widehat{\beta}$ is approximately normal with mean β and a variance-covariance matrix that can be approximated by the estimated inverse of the Fisher Information Matrix; i.e.,

$$\widehat{\boldsymbol{\beta}} \;\dot{\sim}\; \mathcal{N}(\boldsymbol{\beta}, \boxed{\widehat{\boldsymbol{I}}^{-1}(\widehat{\boldsymbol{\beta}})})$$

Inference can be conducted using the Wald approach or via likelihood ratio testing as discussed in our course notes on likelihood-related topics.

For example, a Wald confidence interval for $\boldsymbol{c}^\top\boldsymbol{\beta}$ with approximate coverage probability of 0.95 is given by

$$\boldsymbol{c}^\top\widehat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{c}^\top\widehat{\boldsymbol{I}}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{c}}$$

Interpretation of Logistic Regression Parameters

focus on 1 variable at a time holding all other expl. Variables constant!

Let $\underline{\boldsymbol{x}} = [x_1, x_2, \ldots, x_{j-1}, x_j \quad , x_{j+1}, \ldots, x_p]^\top$.

Let $\widetilde{\boldsymbol{x}} = [x_1, x_2, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p]^\top$.

In other words, $\widetilde{\boldsymbol{x}}$ is the same as $\boldsymbol{x}$ except that the jth explanatory variable has been increased by one unit.

Let $\pi = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top \boldsymbol{\beta})}$ and $\widetilde{\pi} = \frac{\exp(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\beta})}{1+\exp(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\beta})}$.

The Odds Ratio

$$
\begin{aligned}
\frac{\widetilde{\pi}}{1-\widetilde{\pi}} \Bigg/ \frac{\pi}{1-\pi} &= \exp\left\{\log\left(\frac{\widetilde{\pi}}{1-\widetilde{\pi}} \Bigg/ \frac{\pi}{1-\pi}\right)\right\} \\
&= \exp\left\{\log\left(\frac{\widetilde{\pi}}{1-\widetilde{\pi}}\right) - \log\left(\frac{\pi}{1-\pi}\right)\right\} \\
&= \exp\{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\beta} - \boldsymbol{x}^\top \boldsymbol{\beta}\} \\
&= \exp\{(x_j + 1)\beta_j - x_j\beta_j\} \\
&= \exp\{\beta_j\}.
\end{aligned}
$$

Thus,

$$\frac{\widetilde{\pi}}{1-\widetilde{\pi}} = \boxed{\exp(\beta_j)}\frac{\pi}{1-\pi}.$$

All other explanatory variables held constant, the odds of success at $x_j + 1$ are $\exp(\beta_j)$ times the odds of success at x_j.

This is true regardless of the initial value x_j.

A one unit increase in the jth explanatory variable (with all other explanatory variables held constant) is associated with a multiplicative change in the odds of success by the factor $\boxed{\exp(\beta_j).}$

If (L_j, U_j) is a $100(1-\alpha)\%$ confidence interval for β_j, then

$$(\exp(L_j), \exp(U_j))$$

is a $100(1-\alpha)\%$ confidence interval for $\underline{\exp(\beta_j)}$.

Also, note that

$$\begin{aligned}\pi &= \boxed{\frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}} = \frac{1}{\frac{1}{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}+1} \\ &= \boxed{\frac{1}{1+\exp(-\boldsymbol{x}^\top\boldsymbol{\beta})}.}\end{aligned}$$

Thus, if (L_j,U_j) is a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{x}'\boldsymbol{\beta}$, then a $100(1-\alpha)\%$ confidence interval for π is

$$\left(\frac{1}{1+\exp(-L_j)},\frac{1}{1+\exp(-U_j)}\right).$$

Observational study

```
> d=read.delim("http://dnett.github.io/S510/Disease.txt")
> head(d)
  id age ses sector disease savings
1  1  33   1      1       0       1
2  2  35   1      1       0       1
3  3   6   1      1       0       0
4  4  60   1      1       0       1
5  5  18   3      1       1       0
6  6  26   3      1       0       0
>
> d$ses=factor(d$ses)
> d$sector=factor(d$sector)
```

ignore

```
          y
> o=glm(disease~age+ses+sector,
+         family=binomial(link=logit),
+         data=d)
>
> summary(o)

Call:
glm(formula = disease ~ age + ses + sector,
    family = binomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6576  -0.8295  -0.5652   1.0092   2.0842
```

intercept includes ses1 & sector1

$\hat{\beta}$ vcov(·) $z = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.293933   0.436769  -5.252  1.5e-07 ***
age          0.026991   0.008675   3.111 0.001862 **
ses2         0.044609   0.432490   0.103 0.917849
ses3         0.253433   0.405532   0.625 0.532011
sector2      1.243630   0.352271   3.530 0.000415 ***
---
Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

end lecture 41

05-07-25

```
    Null deviance: 236.33  on 195  degrees of freedom
Residual deviance: 211.22  on 191  degrees of freedom
AIC: 221.22

Number of Fisher Scoring iterations: 3
```

```
> coef(o)
(Intercept)         age        ses2        ses3     sector2
-2.29393347 0.02699100 0.04460863 0.25343316 1.24363036

> round(vcov(o),3)
            (Intercept)    age   ses2   ses3 sector2
(Intercept)       0.191 -0.002 -0.083 -0.102  -0.080
age              -0.002  0.000  0.000  0.000   0.000
ses2             -0.083  0.000  0.187  0.072   0.003
ses3             -0.102  0.000  0.072  0.164   0.039
sector2          -0.080  0.000  0.003  0.039   0.124
```

```
> confint(o)
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -3.19560769 -1.47574975
age          0.01024152  0.04445014
ses2        -0.81499026  0.89014587
ses3        -0.53951033  1.05825383
sector2      0.56319260  1.94992969
```

```
> oreduced=glm(disease~age+sector,
+       family=binomial(link=logit),
+       data=d)
>
> anova(oreduced,o,test="Chisq")
Analysis of Deviance Table

Model 1: disease ~ age + sector
Model 2: disease ~ age + ses + sector
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       193     211.64
2       191     211.22  2   0.4193   0.8109
```

```
> o=oreduced
> anova(o,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: disease

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     195     236.33
age     1   12.013       194     224.32 0.0005283 ***
sector  1   12.677       193     211.64 0.0003702 ***
```

```
> head(model.matrix(o))
  (Intercept) age sector2
1           1  33       0
2           1  35       0
3           1   6       0
4           1  60       0
5           1  18       0
6           1  26       0
>
> b=coef(o)
> b
(Intercept)         age     sector2
-2.15965912  0.02681289  1.18169345
```

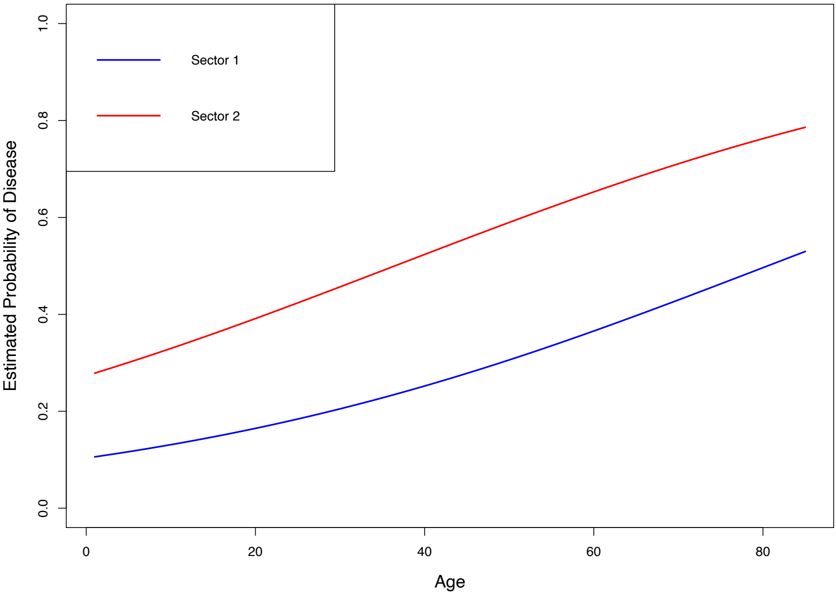
```
> ci=confint(o)
Waiting for profiling to be done...
> ci
                 2.5 %      97.5 %
(Intercept) -2.86990940 -1.51605906
age          0.01010532  0.04421365
sector2      0.52854584  1.85407936
```

```
> #How should we interpret our estimate of
> #the slope coefficient on age?
> exp(b[2])
     age
1.027176
> #All else equal, the odds of disease are about 1.027
> #times greater for someone age x+1 than for someone
> #age x. An increase of one year in age is associated
> #with an increase in the odds of disease by about 2.7%.
> #A 95% confidence interval for the multiplicative
> #increase factor is
> exp(ci[2,])
   2.5 %   97.5 %
1.010157 1.045206
```

```
> #How should we interpret our estimate of
> #the slope coefficient on sector?
> exp(b[3])
sector2
3.25989
> #All else equal, the odds of disease are about 3.26
> #times greater for someone living in sector 2 than for
> #someone living in sector one.
> #A 95% confidence interval for the multiplicative
> #increase factor is
> exp(ci[3,])
   2.5 %   97.5 %
1.696464 6.385816
```

```
> #Estimate the probability that a randomly
> #selected 40-year-old living in sector 2
> #has the disease.
> x=c(1,40,1)
> 1/(1+exp(-t(x)%*%b))
          [,1]
[1,] 0.5236198
> #Approximate 95% confidence interval
> #for the probability in question.
> sexb=sqrt(t(x)%*%vcov(o)%*%x)
> cixb=c(t(x)%*%b-2*sexb,t(x)%*%b+2*sexb)
> 1/(1+exp(-cixb))
[1] 0.3965921 0.6476635
```

```
> #Plot estimated probabilities as a function
> #of age for each sector.
>
> x=1:85
plot(x,1/(1+exp(-(b[1]+b[2]*x))),ylim=c(0,1),
     type="l",col=4,lwd=2,xlab="Age",
     ylab="Estimated Probability of Disease", cex.lab=1.3)
lines(x,1/(1+exp(-(b[1]+b[2]*x+b[3]))),col=2,lwd=2)
legend("topleft", legend=c("Sector 1","Sector 2"),
       col=c(4,2),lwd=2)
```


Sector 1
Sector 2
Estimated Probability of Disease
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Age