Generalized Linear Models

Binary Data

Demetris Athienitis

UF UNIVERSITY of FLORIDA

GLMs for binary data

- Will spend a substantial amount of time on binary GLMs later (see chapters 4-5 of the book)
- Remember that a binary Bernoulli response Y is defined by the probabilities

$$\mu = P(Y = 1) = \pi \qquad P(Y = 0) = 1 - \pi$$

GLMs for binary data

- Will spend a substantial amount of time on binary GLMs later (see chapters 4-5 of the book)
- Remember that a binary Bernoulli response Y is defined by the probabilities

$$\mu = P(Y = 1) = \pi \qquad P(Y = 0) = 1 - \pi$$

Identity Link

Consider a simple situation, with one predictor, x and identity link:

$$\pi(x) = \alpha + \beta x$$

where the parameter β tells us how x relates to Y.

- $\beta > 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes up
- $\beta < 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes down

Model may predict $\pi(x) < 0$ or $\pi(x) > 1$. As such, other links shall be used, such as *logit* and *probit*.

Identity Link

Consider a simple situation, with one predictor, x and identity link:

$$\pi(x) = \alpha + \beta x$$

where the parameter β tells us how x relates to Y.

- $\beta > 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes up
- $\beta < 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes down

Model may predict $\pi(x) < 0$ or $\pi(x) > 1$. As such, other links shall be used, such as *logit* and *probit*.

Identity Link

Consider a simple situation, with one predictor, x and identity link:

$$\pi(x) = \alpha + \beta x$$

where the parameter β tells us how x relates to Y.

- $\beta > 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes up
- $\beta < 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes down

Model may predict $\pi(x) < 0$ or $\pi(x) > 1$. As such, other links shall be used, such as *logit* and *probit*.

Identity Link

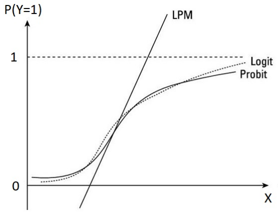
Consider a simple situation, with one predictor, x and identity link:

$$\pi(x) = \alpha + \beta x$$

where the parameter β tells us how x relates to Y.

- $\beta > 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes up
- $\beta < 0$ indicates that as x goes up, $P(Y = 1) = \pi(x)$ goes down

Model may predict $\pi(x) < 0$ or $\pi(x) > 1$. As such, other links shall be used, such as *logit* and *probit*.



Section 2

Logit Link

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

That is,

$$\pi(x) = F_0(\alpha + \beta x) \quad \Rightarrow \quad F_0^{-1}(\pi(x)) = \alpha + \beta x$$

where

$$F_0(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

is the (standard) cdf of the *logistic* distribution.

The link function is the logistic's distribution quantile function (which is also the canonical link)

$$g(\cdot) \equiv F_0^{-1}(\cdot)$$

guaranteeing that $0 \leq \pi(x) \leq 1$

Logit Link

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

That is,

$$\pi(x) = F_0(\alpha + \beta x) \quad \Rightarrow \quad F_0^{-1}\left(\pi(x)\right) = \alpha + \beta x$$

where

$$F_0(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

is the (standard) cdf of the *logistic* distribution.

The link function is the logistic's distribution quantile function (which is also the canonical link)

$$g(\cdot) \equiv F_0^{-1}(\cdot)$$

guaranteeing that $0 \leq \pi(x) \leq 1$

Logit Link

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

That is,

$$\pi(x) = F_0(\alpha + \beta x) \quad \Rightarrow \quad F_0^{-1}(\pi(x)) = \alpha + \beta x$$

where

$$F_0(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

is the (standard) cdf of the *logistic* distribution.

The link function is the logistic's distribution quantile function (which is also the canonical link)

$$g(\cdot) \equiv F_0^{-1}(\cdot)$$

guaranteeing that $0 \leq \pi(x) \leq 1$

Logit Link

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- The parameter β determines the rate of increase or decrease of the curve and the magnitude of β determines how fast the curve increases or decreases
- When $\beta > 0$, $\pi(x)$ increases as x increases
- When $\beta < 0$, $\pi(x)$ decreases as x increases

Logit Link

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- The parameter β determines the rate of increase or decrease of the curve and the magnitude of β determines how fast the curve increases or decreases
- When $\beta > 0$, $\pi(x)$ increases as x increases
- When $\beta < 0$, $\pi(x)$ decreases as x increases

Logit Link

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- The parameter β determines the rate of increase or decrease of the curve and the magnitude of β determines how fast the curve increases or decreases
- When $\beta > 0$, $\pi(x)$ increases as x increases
- When $\beta < 0$, $\pi(x)$ decreases as x increases

Remark

In the next chapters we will see that the $100(1-\alpha)\%$ CI on β is

$$\hat{\beta} \mp z_{1-\alpha/2}\left(s_{\hat{\beta}}\right)$$

where the estimate and standard error are provided by the software.

In R

A GLM is fitted using

```
mymodel=glm(formula,family,data)
```

Basic output is provided with `summary(mymodel)` and CI created on the coefficients via `confint(mymodel)`.

For a logistic regression, with Bernoulli/binomial family (and default logit link)

- When the response column is y is 0 or 1,

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

- When there is a column grouping successes and one grouping failures,

```
glm(cbind(Successes,Failures)~x,family=binomial,data=mydata)
```

Please see the `help(glm)` help file.

In R

A GLM is fitted using

```
mymodel=glm(formula,family,data)
```

Basic output is provided with `summary(mymodel)` and CI created on the coefficients via `confint(mymodel)`.

For a logistic regression, with Bernoulli/binomial family (and default logit link)

- When the response column is y is 0 or 1,

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

- When there is a column grouping successes and one grouping failures,

  ```
  glm(cbind(Successes,Failures)~x,family=binomial,data=mydata)
  ```

Please see the `help(glm)` help file.

In R

A GLM is fitted using

```
mymodel=glm(formula,family,data)
```

Basic output is provided with `summary(mymodel)` and CI created on the coefficients via `confint(mymodel)`.

For a logistic regression, with Bernoulli/binomial family (and default logit link)

- When the response column is y is 0 or 1,

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

- When there is a column grouping successes and one grouping failures,

```
glm(cbind(Successes,Failures)~x,family=binomial,data=mydata)
```

Please see the `help(glm)` help file.

In R

A GLM is fitted using

```
mymodel=glm(formula,family,data)
```

Basic output is provided with `summary(mymodel)` and CI created on the coefficients via `confint(mymodel)`.

For a logistic regression, with Bernoulli/binomial family (and default logit link)

- When the response column is y is 0 or 1,

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

- When there is a column grouping successes and one grouping failures,

  ```
  glm(cbind(Successes,Failures)~x,family=binomial,data=mydata)
  ```

Please see the `help(glm)` help file.

In R

A GLM is fitted using

```
mymodel=glm(formula,family,data)
```

Basic output is provided with `summary(mymodel)` and CI created on the coefficients via `confint(mymodel)`.

For a logistic regression, with Bernoulli/binomial family (and default logit link)

- When the response column is y is 0 or 1,

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

- When there is a column grouping successes and one grouping failures,

```
glm(cbind(Successes,Failures)~x,family=binomial,data=mydata)
```

Please see the `help(glm)` help file.

Example (Infant Malformation)

A study was conducted about infant sex organ malformation and pregnant mother's alcohol consumption.

- Y = infant sex organ malformation (1 = present, 0 = absent)
- x = mother's alcohol consumption (avg drinks per day)

Consumption		Malformation	
Measured	Score	Absent	Present
0	0.0	17066	48
< 1	0.5	14464	38
1-2	1.5	788	5
3-5	4.0	126	1
≥ 6	7.0	37	1

Example (continued)

```
> malform.logit=glm(cbind(Present,Absent)~Alcohol,
+  family=binomial(link=logit))
> summary(malform.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
Alcohol       0.3166     0.1254   2.523   0.0116 *
---
    Null deviance: 6.2020  on 4  degrees of freedom
Residual deviance: 1.9487  on 3  degrees of freedom
AIC: 24.576
```

$$\text{logit}\left[\hat{\pi}(x)\right] = -5.9605 + 0.3166(\text{Alcohol Score})$$

Probit Link

Just as the logistic regression model utilized the logistic's distribution quantile function, an alternative is quantile function of the (standard) normal distribution

$$g(\cdot) \equiv \Phi^{-1}(\cdot)$$

which implies

$$\pi(x) = \Phi(\alpha + \beta x)$$

The probit transform maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1 - \pi(x)$, when $\beta < 0$) has the appearance of the normal cdf with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$.

Probit Link

Just as the logistic regression model utilized the logistic's distribution quantile function, an alternative is quantile function of the (standard) normal distribution

$$g(\cdot) \equiv \Phi^{-1}(\cdot)$$

which implies

$$\pi(x) = \Phi(\alpha + \beta x)$$

The probit transform maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1 - \pi(x)$, when $\beta < 0$) has the appearance of the normal cdf with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$.

Probit Link

Just as the logistic regression model utilized the logistic's distribution quantile function, an alternative is quantile function of the (standard) normal distribution

$$g(\cdot) \equiv \Phi^{-1}(\cdot)$$

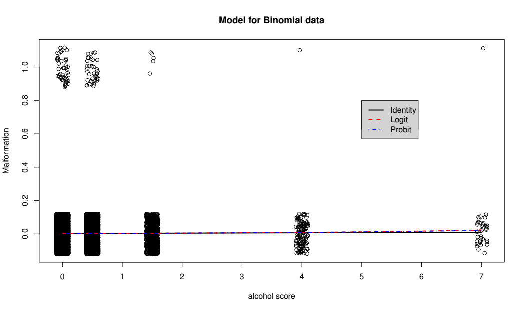
which implies

$$\pi(x) = \Phi(\alpha + \beta x)$$

The probit transform maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1 - \pi(x)$, when $\beta < 0$) has the appearance of the normal cdf with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$.

Example (continued)

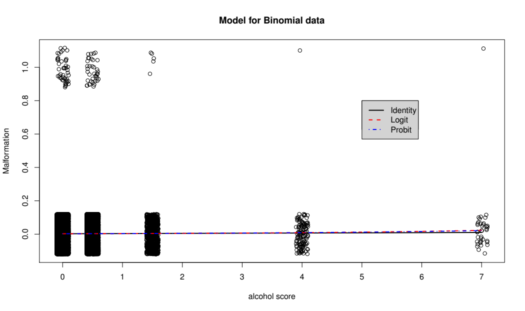
```
> malform.probit=glm(cbind(Present,Absent)~Alcohol,
+  family=binomial(link=probit))
```



In this case all three links seem to be adequate and in general when probit is adequate so will logit.

Example (continued)

```
> malform.probit=glm(cbind(Present,Absent)~Alcohol,
+  family=binomial(link=probit))
```



In this case all three links seem to be adequate and in general when probit is adequate so will logit.

Example (Challenger disaster)

For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the data shows the temperature at the time of flight and whether at least one primary O-ring suffered thermal distress.

Flight	Temp	Failure
1	66	0
2	70	1
⋮	⋮	⋮
22	76	0
23	58	1

Example (continued)

```
> preC.logit=glm(Failure~Temp,family=binomial(link=logit),data=
> summary(preC.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temp         -0.2322     0.1082  -2.145   0.0320 *
---
    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

> confint(preC.logit)
                 2.5 %      97.5 %
(Intercept)  3.3305848 34.34215133
Temp        -0.5154718 -0.06082076
```

Example (continued)

According to the report, the air temperature at the time of launch, 11:38 a.m. EST, was 36 degrees. This temperature was 15 degrees colder than any previous launch and the O-ring suffered catastrophic failure.

Example (continued)

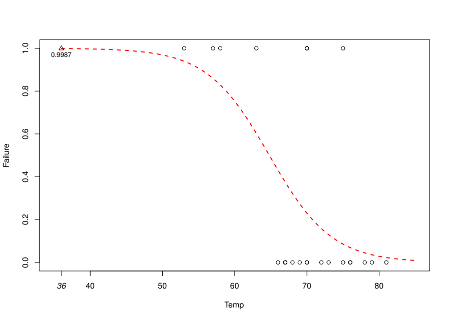
According to the report, the air temperature at the time of launch, 11:38 a.m. EST, was 36 degrees. This temperature was 15 degrees colder than any previous launch and the O-ring suffered catastrophic failure.

```
> predict.glm(preC.logit,newdata=data.frame(Temp=36),
+  type="response")
0.9987521
```

Example (continued)

According to the report, the air temperature at the time of launch, 11:38 a.m. EST, was 36 degrees. This temperature was 15 degrees colder than any previous launch and the O-ring suffered catastrophic failure.

```
> predict.glm(preC.logit,newdata=data.frame(Temp=36),
+  type="response")
0.9987521
```



We learned

For Bernoulli/binomial data we use

- Logit Link
- Probit Link

and comprehend (slightly) the impact of the coefficient β on the probability of "succeess"