IDS 702: Module 2.7

Aggregated outcomes; Probit regression

Dr. Olanrewaju Michael Akande

Aggregated binary outcomes

- In the datasets we have seen so far under logistic regression, we observe the binary outcomes for each observation, that is, each $y_i \in \{0, 1\}$.
- This is not always the case. Sometimes, we get an aggregated version, with the outcome summed up by combinations of other variables.
- For example, for individual-level data, suppose we had

response	0	0	1	1	1	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0	0	1	1	0	1
predictor	3	3	2	1	2	3	2	2	2	2	3	1	3	1	1	2	2	2	2	1	3	3	3	1	3

where predictor is a factor with 3 levels: 1,2,3.

- The aggregated version of the same data could look like

predictor	n	successes
1	31	17
2	35	16
3	34	14

IDS 702

Aggregated binary outcomes

- Recall that if $Y \sim \mathrm{Bin}(n, p)$ (that is, Y is a random variable that follows a binomial distribution with parameters n and p), then Y follows a $\mathrm{Bernoulli}(p)$ distribution when $n = 1$.

- Alternatively, we also have that if $Z_1, \ldots, Z_n \sim \mathrm{Bernoulli}(p)$, then $Y = \sum_i^n Z_i \sim \mathrm{Bin}(n, p)$.

- That is, the sum of n "iid" $\mathrm{Bernoulli}(p)$ random variables gives a random variable with the $\mathrm{Bin}(n, p)$ distribution.

- The logistic regression model can be used either for Bernoulli data (as we have done so far) or for data summarized as binomial counts (that is, aggregated counts).

- In the aggregated form, the model is

$$y_i | x_i \sim \mathrm{Bin}(n_i, \pi_i); \quad \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

IDS 702

Bernoulli versus binomial outcomes

Normally, for individual-level data, we would have

```
##   response predictor
## 1        0         3
## 2        0         3
## 3        1         2
## 4        1         1
## 5        1         2
## 6        0         3
```

```
M1 <- glm(response~predictor,data=Data,family=binomial)
summary(M1)
```

```
##
## Call:
## glm(formula = response ~ predictor, family = binomial, data = Data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.261  -1.105  -1.030   1.251   1.332
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1942     0.3609   0.538    0.591
## predictor2   -0.3660     0.4954  -0.739    0.460
## predictor3   -0.5508     0.5017  -1.098    0.272
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.27  on 99  degrees of freedom
## Residual deviance: 137.02  on 97  degrees of freedom
## AIC: 143.02
##
## Number of Fisher Scoring iterations: 4
```

Bernoulli versus binomial outcomes

But we could also do the following with the aggregate level data instead

```
M2 <- glm(cbind(successes,n-successes)~predictor,data=Data_agg,family=binomial)
summary(M2)
```

```
##
## Call:
## glm(formula = cbind(successes, n - successes) ~ predictor, family = binomial,
##     data = Data_agg)
##
## Deviance Residuals:
## [1]  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1942     0.3609   0.538    0.591
## predictor2   -0.3660     0.4954  -0.739    0.460
## predictor3   -0.5508     0.5017  -1.098    0.272
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.2524e+00  on 2  degrees of freedom
## Residual deviance: 1.3323e-14  on 0  degrees of freedom
## AIC: 17.868
##
## Number of Fisher Scoring iterations: 2
```

Same results overall! Deviance and AIC are different because of the different likelihood functions.

Note that some glm functions use n in the formular instead of n-successes.

IDS 702

Probit regression

IDS 702

Probit regression

- Recall the "Bernoulli" logistic regression model:

$$y_i | x_i \sim \text{Bernoulli}(\pi_i); \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

  for $i = 1, \ldots, n$.

- Here the link function is the logit function, which ensures that the probabilities lie between 0 and 1.

- We can also use the probit function Φ^{-1}, which is the quantile function associated with the standard normal distribution $N(0, 1)$, as the link.

IDS 702

Probit regression

- That is, suppose H follows a standard normal distribution, that is, $H \sim N(0,1)$.
- Then Φ is the CDF, that is, $\Pr[H \leq h] = \Phi(h)$.
- Formally, the probit regression model can be written as

$$y_i|x_i \sim \text{Bernoulli}(\pi_i); \quad \Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

- It is then easy to see that

$$\begin{aligned}\Pr[y_i = 1|x_i] = \pi_i &= \Phi\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \\ &= \Pr[H \leq \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}].\end{aligned}$$

IDS 702

Latent variable representation

- It turns out that we can rewrite the probit regression model as

$$\begin{aligned} y_i &= 1[z_i > 0]; \\ z_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i; \quad \epsilon_i \sim N(0,1) \end{aligned}$$

where $y_i = 1[z_i > 0]$ means $y_i = 1$ if $z_i > 0$ and $y_i = 0$ if $z_i < 0$.

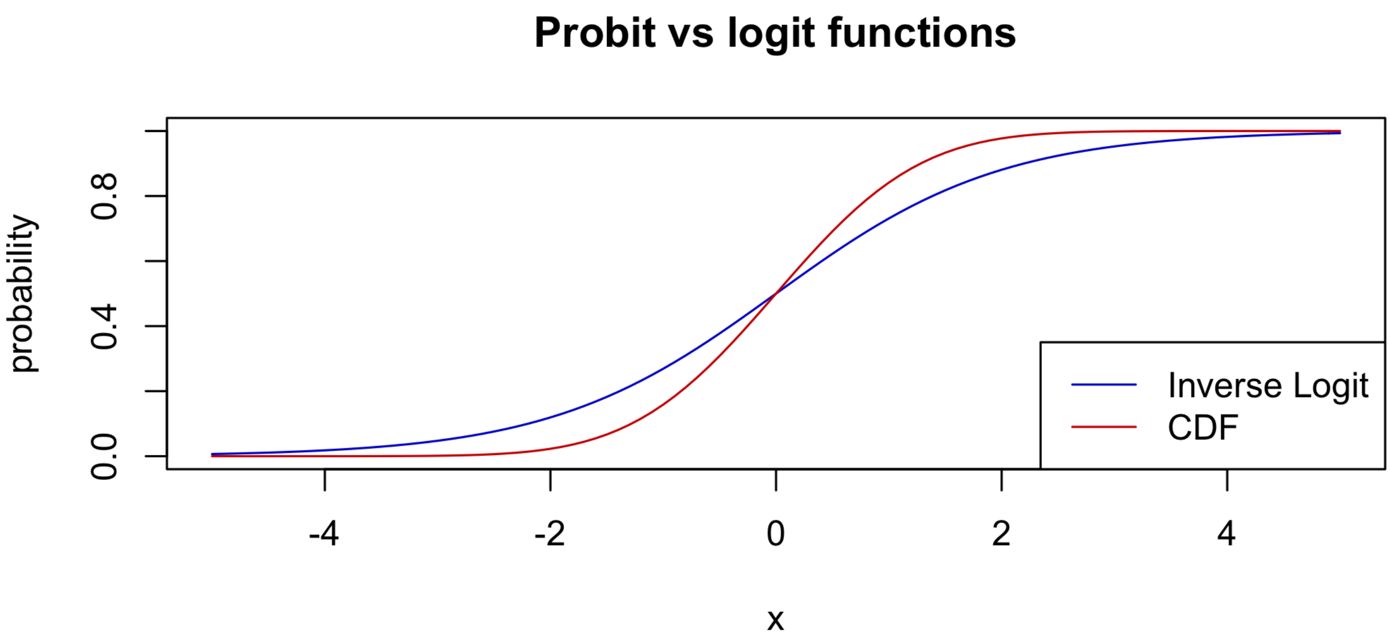
- To see that the two representations are equivalent, note that

$$\begin{aligned} \Pr[y_i = 1 | x_i] &= \Pr[z_i > 0] \\ &= \Pr[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i > 0] \\ &= \Pr[\epsilon_i > -(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})] \\ &= \Pr[\epsilon_i < (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})] \quad [\text{since} \quad \epsilon_i \sim N(0,1)] \\ &= \Phi\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) = \pi_i \end{aligned}$$

- Clearly, we do not observe $Z = (z_1, z_2, \ldots, z_n)$ and it is thus referred to as an auxiliary variable.

Probit vs logit functions?

- The plots below compares the inverse logit function $\pi_i = \dfrac{e^x}{1+e^x}$ and the CDF function (inverse probit) $\pi_i = \Phi(x)$.



- Notice that they are similar, but the CDF of the standard normal distribution has fatter tails (the inverse logit has thinner tails).

IDS 702

Probit or logistic regression?

- In practice, the decision to use one or the other is often based on preference: the overall conclusions from both are usually quite similar.
- The results based on logistic regression (using odds and odds ratio) can be more interpretable than those based on Probit regression.
- In some applications, interpreting the z_i's may be meaningful but that is not always the case.
- For example, suppose y_i is a binary variable for whether or not person i chooses to buy the new iPhone, then z_i can be thought of as person i's "utility" in a way.
- Works in this example, but does not always work across different domains.
- In R, use the `glm` command but set the option `family="binomial(link=probit)` instead of `family="binomial(link=logit)`.

IDS 702

What's next?

Move on to the readings for the next module!