Intro Logistic Regression – Example uses the Class Survey data

Redefining the Response, Y:

Original: 0 or 1 → As probability: [0,1] → Y' ϵ (-∞,∞)

Goal: Use a function, F(Y) that goes from [0,1] interval to the real line

Option 1: A function that does this in reverse, that is, given any real value it produces a probability between 0 and 1 is the cumulative normal distribution, Φ. That is, given any Z-score, Φ(Z) ϵ [0,1]

We can then say that:

Y = Φ(**X**B + e) => Φ^{-1}(Y) = **X**B + e => Y' = **X**B + e

Thus our link function is F(Y) = Φ^{-1}(Y) which is know as the **Probit** (short for "probability unit") link

NOTE: **X**B is the matrix notation for linear model. For ease of understanding, just picture **X**B as representing Bo + B1X1 + B2X2 + etc. for all predictors in your model.

Option 2: An approach based on the odds ratio. If p (some may use π) represents the probability some event occurs, then the odds of that event happening are O(p) = p/(1-p).

- p = 0 → O(p) = 0
- p = ¼ → O(p) = 1/3 meaning odds are 1 to 3 against
- p = ½ → O(p) = 1 which are even odds
- p = ¾ → O(p) = 3 or odds are 3 to 1 in favor
- p = 1 → O(p) = ∞

In redefining Y using the odds, O(Y) ϵ [0, ∞). By taking the log of the odds of Y (i.e. log odds), Y' results in Y' ϵ (-∞,∞). Here, log is referencing the natural log opposed to base-10 logs. The reasoning is that with base-e, in general the slope estimate *B*, can be interpreted as a *B*% increase in Y. For example, if the slope estimate is 0.05 then this can be interpreted as an approximate 5% for a unit change in X.

By using this transformation method our link function, F(Y) = log[O(Y)] = log[y/(1-y)] and is called the **Logit** link. This link is commonly the default link in statistical software.

Recap:

Probit: $\Phi^{-1}(\hat{Y})$ = **X**B Logit: $Log\left(\frac{\hat{Y}}{1-\hat{Y}}\right)$ = **X**B

Solving for Y:

Probit: $\hat{Y}$= Φ(**X**B) or c.d.f. of Z Logit: $\hat{Y} = \frac{e^{XB}}{1+e^{XB}} = \frac{\exp(XB)}{1+\exp(XB)} = \frac{1}{1+e^{-XB}} = \frac{1}{1+\exp(-XB)}$

NOTE: "exp" stands for exponential which in approximate terms is 2.7183

Example using the data from our class survey: Predict Sex of student (1 is Female) based on Height

PROBIT:

```
glm(formula = SexID ~ Height, family = binomial(link = "probit"),
    data = cldt)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.47449  -0.49072  -0.06752   0.47934   2.42312

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  27.2913     7.9047   3.453 0.000555 ***
Height       -0.4071     0.1178  -3.456 0.000547 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The probit regression model is: Φ^{-1}(Y=female) = 27.2913 - 0.4071*Height

Using equation to predict probability a student is Female based on observed height of 67 and 68 inches where $\widehat{Y}$ = Φ(27.2913 - 0.4071*Height);

X = 67: Φ^{-1}(Y=male) = 27.2913 - 0.4071*(67) = 0.0156 and from table P(Z < 0.02) = 0.5080

X = 68: Φ^{-1}(Y=male) = 27.2913 - 0.4071*(68) = -0.3915and from table P(Z < -0.39) = 0.3483

From software we get 0.5052 and 0.3467, respectively.

NOTE that this is slope refers to a change in the Z score and NOT change in probability.

LOGIT:

```
glm(formula = SexID ~ Height, family = binomial, data = cldt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4685  -0.4490  -0.1075   0.4772   2.4579

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  48.6166    16.0271   3.033  0.00242 **
Height       -0.7266     0.2391  -3.039  0.00237 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The logit regression model is: $Log\left(\frac{\hat{Y}}{1-\hat{Y}}\right)$ = 48.6166 – 0.7266*Height

Using equation to predict probability a student is Female based on observed height of 67 and 68 inches where $\hat{Y} = \frac{e^{48.6166 - 0.7266*\text{Height}}}{1+e^{48.6166 - 0.7266*\text{Height}}}$;

X = 67: P(Y=1) = $\frac{e^{48.6166 - 0.7266*(67)}}{1+e^{48.6166 - 0.7266*(67)}}$ = 0.4836

X = 68: P(Y=1) = $\frac{e^{48.6166 - 0.7266*(68)}}{1+e^{48.6166 - 0.7266*(68)}}$ = 0.3117

From software we get 0.4839 and 0.3120, respectively.

NOTE that this is slope refers to a change in estimated log odds of Y =1.

e^b = odds ratio e.g. $e^{-0.7266}$ = 0.484 The interpretation of the odds ratio is that for every increase of 1 unit in Height, the estimated odds of student being female are multiplied by about 0.5

At Height = 68, the estimate odds are exp(48.616 *- 0.7266*68) = 0.4528

At Height = 67, the estimate odds are exp(48.616 *- 0.7266*67) = 0.9365

The resulting odds ratio is 0.4528/0.9365 = 0.484

See that 0.48 * 0.9365 = 0.453

Confidence Intervals for Slope Coefficients:

In simplest terms, the profile CI requires calculating the profile likelihood for different values of X where the profile likelihood requires maximizing the likelihood function and is time-intensive. See the online notes for explanation of this likelihood function.

```
> confint(lprobit)
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) 14.2350043 45.9164919
Height      -0.6831923 -0.2131476

> confint(llogit)
Waiting for profiling to be done...
               2.5 %     97.5 %
(Intercept) 23.817501 89.8821669
Height      -1.341628 -0.3567943

> #Alternative - fit based on asymptotic normality
> confint.default(lprobit)
                2.5 %     97.5 %
(Intercept) 11.7983912 42.7841261
Height      -0.6380021 -0.1762756
> confint.default(llogit)
               2.5 %     97.5 %
(Intercept) 17.203991 80.0291631
Height      -1.195119 -0.2580396
```

```
cldt = read.table('ClassData.csv', sep=',', header=T)
#NOTE: use SexID as response must be 0,1 not text

#fit probit regression model
lprobit <- glm(SexID ~ Height, family = binomial(link = "probit"), data = cldt)
summary(lprobit)

#fit logit regression model
#Note the default link is logit-doesn't need to be named
llogit <- glm(SexID ~ Height, family = binomial, data = cldt)
summary(llogit)

#CI for slope estimates
#NOTE: these are profiled confidence intervals by default...
#...created by profiling the likelihood function and may not be symmetric
confint(lprobit)
confint(llogit)

#Alternative - fit based on asymptotic normality
confint.default(lprobit)
confint.default(llogit)

#Predicted probabilities for new observations
#Can use the SE to construct CI for observations
predict(lprobit, newdata=data.frame(Height=c(67,68)), type="response",
    se.fit=TRUE)#uses probit link
predict(llogit, newdata=data.frame(Height=c(67,68)), type="response",
    se.fit=TRUE)#uses logit link

#Fitted plot - probit link
plot(cldt$Height, cldt$SexID, pch = 16, xlab = "Height (inches)", ylab = "Probability Female")
curve(predict(lprobit,data.frame(Height=x),type="resp"),add=TRUE, col="blue")

#Fitted plot - logit link
plot(cldt$Height, cldt$SexID, pch = 16, xlab = "Height (inches)", ylab = "Probability Female")
curve(predict(llogit,data.frame(Height=x),type="resp"),add=TRUE, col="red")

#Fitted plot with both links in one graph
plot(cldt$Height, cldt$SexID, pch = 16, xlab = "Height (inches)", ylab = "Probability Female")
curve(predict(lprobit,data.frame(Height=x),type="resp"),add=TRUE, col="blue")
curve(predict(llogit,data.frame(Height=x),type="resp"),add=TRUE, col="red")
```

```
#Some diagnostic graphs - see online notes for formulas
#Best results are no patterns or residual values > |2|
plot(residuals(lprobit, type="pearson"), type="b", main="Pearson Res - Probit")
plot(residuals(lprobit, type="deviance"), type="b", main="Deviance Res - Probit")
plot(residuals(llogit, type="pearson"), type="b", main="Pearson Res - Logit")
plot(residuals(llogit, type="deviance"), type="b", main="Deviance Res - Logit")
```