Logistic Regression

Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning

Logistic Regression Theory

- The linear probability model.

$$\hat{p}_i = B_1 X_i + B_0$$

where

$$\hat{p}_i = \frac{1}{1 + e^{-(B_1 X_i + B_0)}} = \frac{e^{(B_1 X_i + B_0)}}{1 + e^{(B_1 X_i + B_0)}}$$

Expressions of the Logistic Model

- We can determine the second form of the logistic model as

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{(B_1 X_i + B_0)}$$

which is also the equivalent of

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = B_1 X_i + B_0$$

- This means that $B_1 X_i + B_0$ is now in linear form (like the OLS linear model). However, the predicted score has changed form to the *logit* such that

$$logit = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) \text{ or } logit = B_1 X_i + B_0$$

Example from Cohen et al. (2003)

http://faculty.smu.edu/kyler/courses/7312/cohenex.txt

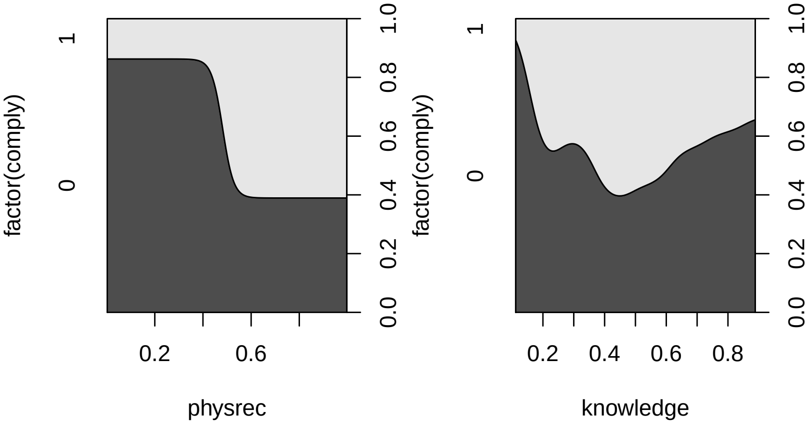
- comply - (1=yes; 0=no) whether or not someone is in compliance with mammography screening
- physrec - whether or not she has received a recommendation from a physician
- knowledge - test of her knowledge of breast cancer screening
- benefits - her perception of mammography screening
- barriers - her perception of the barriers to being screened

```
> mamm <- read.table("cohenex.txt", header = T)
> attach(mamm)
> head(mamm)

  case physrec comply knowledge benefits barriers
1  205       1      1      0.22        4        3
2  210       0      0      0.56        1        1
3  213       1      0      0.44        4        3
4  218       0      0      0.33        3        0
5  229       1      1      0.44        5        0
6  231       0      1      0.56        5        0
```

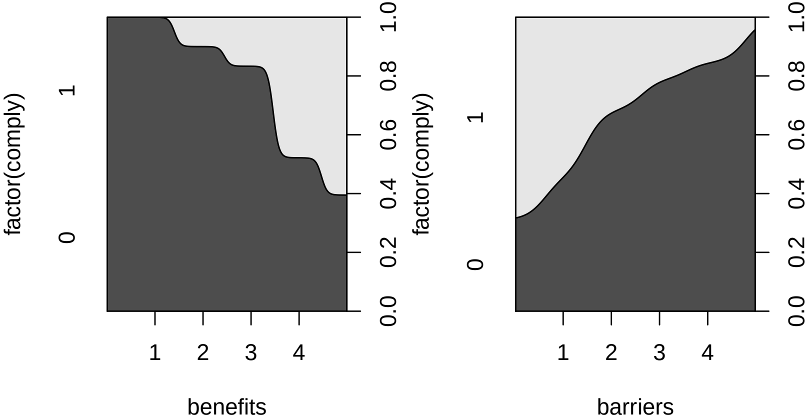
Conditional Density Plots

```
> layout(matrix(1:2, ncol = 2))
> cdplot(factor(comply) ~ physrec)
> cdplot(factor(comply) ~ knowledge)
```



Conditional Density Plots, (cont.)

```
> layout(matrix(1:2, ncol = 2))
> cdplot(factor(comply) ~ benefits)
> cdplot(factor(comply) ~ barriers)
```



Running the Logistic Regression

```
> m1 <- glm(comply ~ physrec, family = binomial(link = "logit"))
> summary(m1)

Call:
glm(formula = comply ~ physrec, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3735  -1.3735  -0.5434   0.9933   1.9929

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8383     0.4069  -4.518 6.26e-06
physrec       2.2882     0.4503   5.081 3.75e-07

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 226.47  on 163  degrees of freedom
Residual deviance: 191.87  on 162  degrees of freedom
AIC: 195.87

Number of Fisher Scoring iterations: 4
```

Interpreting Results

The odds of complying if NOT recommended by physician:

```
> exp(-1.8383)

[1] 0.1590876
```

The odds of complying if recommended by physician:

```
> exp(-1.8383) * exp(2.2882)

[1] 1.568155
```

The probability of complying if NOT recommended by physician:

```
> exp(-1.8383)/(1 + exp(-1.8383))

[1] 0.1372525
```

The probability of complying if recommended by physician:

```
> (exp(-1.8383) * exp(2.2882))/(1 + exp(-1.8383) *
+     exp(2.2882))

[1] 0.6106155

> -1.8383 + 2.2882

[1] 0.4499

> exp(0.45)/(1 + exp(0.45))

[1] 0.6106392
```

Huberty I Index

- The Huberty I Index is a measure of the correct classification of individuals given the model.

```
> correct.m1 <- ifelse(m1$fitted < 0.5, 0, 1)
> table(comply, correct.m1)

      correct.m1
comply  0  1
     0 44 44
     1  7 69

> cbind(physrec, comply, logit = m1$linear, prob = m1$fitted)[1:
+     ]

  physrec comply      logit      prob
1       1      1  0.4499169 0.6106195
2       0      0 -1.8382795 0.1372549
3       1      0  0.4499169 0.6106195
4       0      0 -1.8382795 0.1372549
5       1      1  0.4499169 0.6106195
6       0      1 -1.8382795 0.1372549
```

Adding Other Variables to the Model

```
> m2 <- glm(comply ~ knowledge, family = binomial(link = "logit"
> summary(m2)

Call:
glm(formula = comply ~ knowledge, family = binomial(link = "logit"))

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.276  -1.099  -1.032    1.223    1.330

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3109     0.5707   0.545    0.586
knowledge    -0.7451     0.8945  -0.833    0.405

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 226.47  on 163  degrees of freedom
Residual deviance: 225.78  on 162  degrees of freedom
AIC: 229.78

Number of Fisher Scoring iterations: 3
```

Knowledge Variable

Odds Ratio for Knowledge

```
> exp(-0.745)

[1] 0.4747343
```

So the probability of being in compliance for someone with a knowledge score of 50% (or 0.50).

```
> 0.3109 + (-0.745 * 0.5)

[1] -0.0616

> exp(-0.0616)/(1 + exp(-0.0616))

[1] 0.4846049
```

Huberty I Index

```
> correct.m2 <- ifelse(m2$fitted < 0.5, 0, 1)
> table(comply, correct.m2)

      correct.m2
comply  0  1
     0 77 11
     1 70  6
```

Running Models with Multiple Variables

```
> m3 <- glm(comply ~ physrec + knowledge, family = binomial(link
> summary(m3)

Call:
glm(formula = comply ~ physrec + knowledge, family = binomial(link = "l

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4687  -1.3197  -0.5275   0.9854   2.0408

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5679     0.7447  -2.105   0.0352
physrec       2.2779     0.4509   5.052 4.37e-07
knowledge    -0.4286     0.9955  -0.430   0.6668

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 226.47  on 163  degrees of freedom
Residual deviance: 191.68  on 161  degrees of freedom
AIC: 197.68
```

Odds Ratios, I, and Probability

Odds ratio for `physrec`

```
> exp(2.278)

[1] 9.757147
```

Odds ratio for `knowledge`

```
> exp(-0.429)

[1] 0.6511599
```

Probability of compliance for someone who received a physicians recommendation and had a score of 70% (0.70) on the knowledge test.

```
> -1.5679 + 2.2779 + (-0.4286 * 0.7)

[1] 0.40998

> exp(0.40998)/(1 + exp(0.40998))

[1] 0.6010831

> table(comply, ifelse(m3$fitted < 0.5, 0, 1))

comply  0  1
     0 44 44
     1  7 69
```

Running Models with Multiple Variables (cont.)

```
> m4 <- glm(comply ~ benefits + barriers, family = binomial(link
> summary(m4)

Call:
glm(formula = comply ~ benefits + barriers, family = binomial(link = "l

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6946  -1.0741  -0.3495   0.9504   2.3366

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3664     1.0187  -2.323  0.02018
benefits      0.7061     0.2217   3.185  0.00145
barriers     -0.6036     0.1540  -3.920 8.86e-05

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 226.47  on 163  degrees of freedom
Residual deviance: 184.55  on 161  degrees of freedom
AIC: 190.55
```

Odds Ratios, I, and Probability

Odds ratio for `physrec`

```
> table(comply, ifelse(m4$fitted < 0.5, 0, 1))

comply  0  1
     0 71 17
     1 25 51
```

Probability of compliance for someone who ranked a "3" on benefits and a "4" on barriers.

```
> -2.3664 + (0.7061 * 3) + (-0.6036 * 4)

[1] -2.6625

> exp(-2.6625)/(1 + exp(-2.6625))

[1] 0.06522275
```

Probability of compliance for someone who ranked a "5" on benefits and a "1" on barriers.

```
> -2.3664 + (0.7061 * 5) + (-0.6036 * 1)

[1] 0.5605

> exp(0.5605)/(1 + exp(0.5605))

[1] 0.6365682
```

Probability of being in compliance as a function of the perceived benefits and barriers

```
> plot(benefits ~ barriers)
> symbols(barriers, benefits, circles = predict(m4,
+       type = "response"), add = T)
```