ADA HW8

Liangquan Zhou lz2377

October 30, 2014

Consider the data in Table 1 on mental health.

1. Categorize Mental Health as a binary variable, with values 0, if Normal, and 1, Otherwise; and Education Level with values 0 if No College Degree, and 1 otherwise.

a). Determine whether there is association between Education Level and Mental Health, using logistic regression, without adjusting for Gender. Interpret what the estimated parameters denote.

```
## categorize mental health and education level
dat1 = dat
dat1$Mental_Health = 1*(dat1$Mental_Health != "Normal")
dat1$Education_Level = 1*(dat1$Education_Level != "No_College_Degree")
dat1$Gender = 1*(dat1$Gender == "Male")
fit1 = glm(Mental_Health ~ Education_Level, data = dat1, family = "binomial")
summary(fit1)
```

```
##
## Call:
## glm(formula = Mental_Health ~ Education_Level, family = "binomial",
##     data = dat1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.866  -0.866  -0.858   1.524   1.535
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.8109     0.1900   -4.27    2e-05 ***
## Education_Level   0.0245     0.2603    0.09     0.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 343.19  on 276  degrees of freedom
## Residual deviance: 343.18  on 275  degrees of freedom
## AIC: 347.2
##
## Number of Fisher Scoring iterations: 4
```

The estimated parameter $\beta_0 = -0.81093$, represents the log odds ratio of having a `Depression` or `Severly Depression` for someone who has a `Undergrad Degree` or `Post-grad Degree`.

The estimated parameter $\beta_1 = 0.02445$, represents the log odds ration of having a `Depression` or `Severly Depression` for someone who has `No College Degree`, relative to someone who has a `Undergrad Degree` or `Post-grad Degree`.

b). Repeat (a) adjusting for Gender. Interpret what the estimated parameters denote.

```r
## categorize mental health
fit2 = glm(Mental_Health ~ Education_Level + Gender, data = dat1, family = "binomial")
summary(fit2)
```

```
##
## Call:
## glm(formula = Mental_Health ~ Education_Level + Gender, family = "binomial",
##     data = dat1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.882  -0.871  -0.846   1.505   1.563
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.7739     0.2137   -3.62  0.00029 ***
## Education_Level   0.0309     0.2609    0.12  0.90563
## Gender           -0.0995     0.2655   -0.37  0.70790
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 343.19  on 276  degrees of freedom
## Residual deviance: 343.04  on 274  degrees of freedom
## AIC: 349
##
## Number of Fisher Scoring iterations: 4
```

The estimated parameter $\beta_0 = -0.77389$, represents the log odds ratio of having a `Depression` or `Severly Depression` for someone who has `No College Degree` and with gender `Female`.

The estimated parameter $\beta_1 = 0.03093$, represents the log odds ratio of having a `Depression` or `Severly Depression` for someone who has `No College Degree`, relative to someone who has a `Undergrad Degree` or `Post-grad Degree`, for both `Male` and `Female`.

The estimated parameter $\beta_2 = -0.09946$, represents the log odds ratio of having a `Depression` or `Severly Depression` for someone whose gender is `Male`, relative to someone whose gender is `Female`, for all `Education Level`.

c). Assess whether it is appropriate to pool data across male and female subjects using a suitable logistic regression model.

Use `Hosmer-Lemshow` goodness-of-fit test:

```r
library(ResourceSelection)
hoslem.test(x = fit2$y, y = fitted(fit2), g = 3)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
```

```
##
## data:  fit2$y, fitted(fit2)
## X-squared = 0.0212, df = 1, p-value = 0.8842
```

Since p value is greater than 0.05, we accept H_0 and think it is appropriate to pool data across male and female subjects using a suitable logistic regression model.

2. Repeat 1 (a) - 1 (c) above now using Educational Background as a trichotomous variable, i.e., No College Degree, Undergrad Degree, Post-grad Degree.

a).

Take `No Clooege Degree` as the reference group. Define design variables:

$$D_1 = \begin{cases} 1 & \text{Undergrad Degree} \\ 0 & \text{Otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{Post-grad Degree} \\ 0 & \text{Otherwise} \end{cases}$$

```r
## categorize mental health and education level
dat2 = dat
dat2$Mental_Health = 1*(dat2$Mental_Health != "Normal")
dat2$Gender = 1*(dat2$Gender == "Male")
dat2$D1 = 1*(dat2$Education_Level == "Undergrad_Degree")
dat2$D2 = 1*(dat2$Education_Level == "Post-grad_Degree")

fit1 = glm(Mental_Health ~ D1 + D2, data = dat2, family = "binomial")
summary(fit1)
```

```
##
## Call:
## glm(formula = Mental_Health ~ D1 + D2, family = "binomial", data = dat2)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.874  -0.858  -0.858   1.514   1.538
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.81093    0.19003   -4.27    2e-05 ***
## D1           0.04632    0.30065    0.15     0.88
## D2          -0.00583    0.33466   -0.02     0.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 343.19  on 276  degrees of freedom
## Residual deviance: 343.16  on 274  degrees of freedom
## AIC: 349.2
##
## Number of Fisher Scoring iterations: 4
```

The estimated parameter $\beta_0 = -0.810930$, represents the log odds ratio of having a Depression or Severly Depression for someone who has No College Degree.

The estimated parameter $\beta_1 = 0.046324$, represents the log odds ratio of having a Depression or Severly Depression for someone who has Undergrad Degree, relative to someone who has No College Degree.

The estimated parameter $\beta_2 = -0.005831$, represents the log odds ratio of having a Depression or Severly Depression for someone who has Post-grad Degree, relative to someone who has No College Degree.

b).

```r
## categorize mental health
fit2 = glm(Mental_Health ~ D1 + D2 + Gender, data = dat2, family = "binomial")
summary(fit2)
```

```
##
## Call:
## glm(formula = Mental_Health ~ D1 + D2 + Gender, family = "binomial",
##     data = dat2)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.889  -0.871  -0.837   1.517   1.563
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.77454    0.21377   -3.62  0.00029 ***
## D1           0.04974    0.30087    0.17  0.86869
## D2           0.00459    0.33594    0.01  0.98910
## Gender      -0.09771    0.26582   -0.37  0.71319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 343.19  on 276  degrees of freedom
## Residual deviance: 343.02  on 273  degrees of freedom
## AIC: 351
##
## Number of Fisher Scoring iterations: 4
```

The estimated parameter $\beta_0 = -0.774535$, represents the log odds ratio of having a Depression or Severly Depression for someone who has No College Degree and with gender Female.

The estimated parameter $\beta_1 = 0.049738$, represents the log odds ratio of having a Depression or Severly Depression for someone who has Undergrad Degree, relative to someone who has No College Degree, for both Male and Female.

The estimated parameter $\beta_2 = 0.004591$, represents the log odds ratio of having a Depression or Severly Depression for someone who has Post-grad Degree, relative to someone who has No College Degree, for both Male and Female.

The estimated parameter $\beta_3 = -0.097709$, represents the log odds ratio of having a Depression or Severly Depression for someone whose gender is Male, relative to someone whose gender is Female, for all Education Level.

c).

Use `Hosmer-Lemshow` goodness-of-fit test:

```r
hoslem.test(x = fit2$y, y = fitted(fit2), g = 4)
```

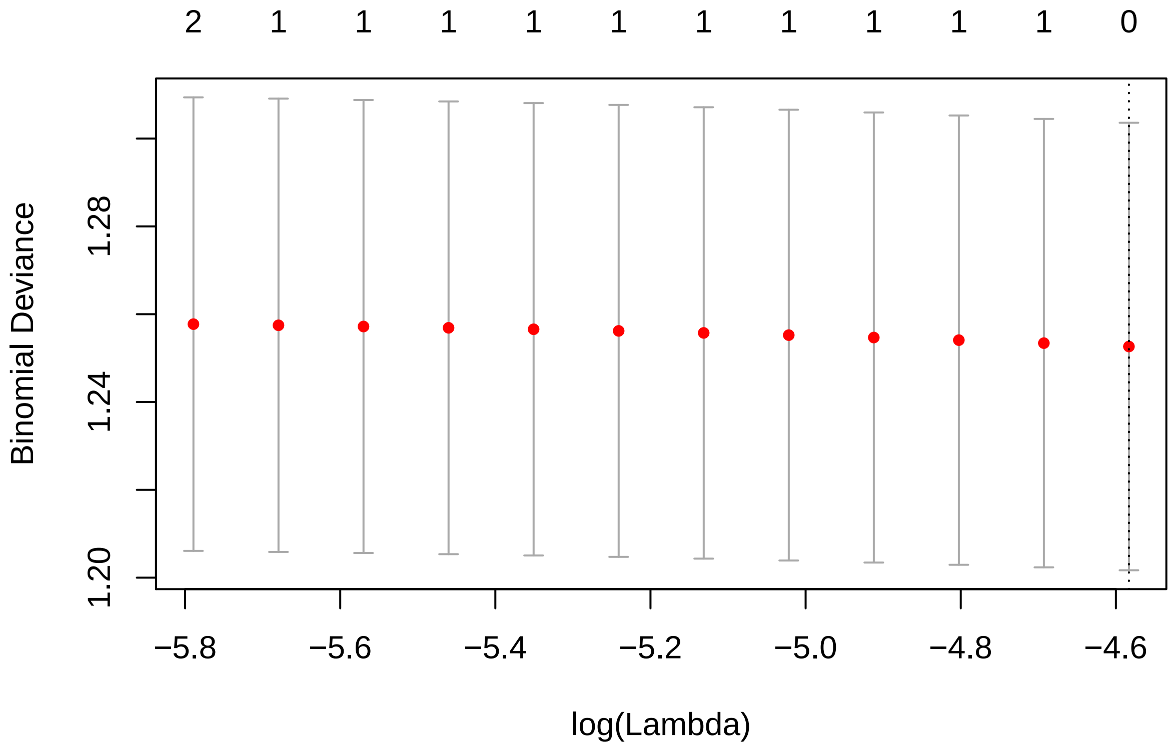
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  fit2$y, fitted(fit2)
## X-squared = 0.0409, df = 2, p-value = 0.9798
```

Since p value is greater than 0.05, we accept H_0 and think it is appropriate to pool data across male and female subjects using a suitable logistic regression model.

3. Repeat 1 (b) using the lasso.

b).

```r
## categorize mental health
library(glmnet)
X = data.matrix(subset(dat1, select = c(Education_Level, Gender)))
y = dat1[,3]
fit <- glmnet(X,y, family = "binomial")
cv.fit <- cv.glmnet(X,y, family = "binomial",nlambda = 85)
plot(cv.fit)
```



```
## best lambda
cv.fit$lambda.min
```

```
## [1] 0.01022
```

```
## best model
model.final <- cv.fit$glmnet.fit
# the best model's coefficients
model.coef <- coef(cv.fit$glmnet.fit, s = cv.fit$lambda.min)
model.coef
```

```
## 3 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)     -0.7979
## Education_Level  .
## Gender           .
```

The lasso regression model is a constant: $logit(p_{DepressedorSeverelyDrepressed|EducationLevel,Gender}) = -0.7979261$