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 $\hat{\beta}_0 = -0.81093$, represents the log odds ($log(\frac{p_{depressed}}{1-p_{depressed}})$) of having a `Depression` or `Severly Depression` for someone who has `No College Degree`.

The estimated parameter $\hat{\beta}_1 = 0.02445$, represents the log odds ratio of having a `Depression` or `Severly Depression` for someone who has a `Undergrad Degree` or `Post-grad Degree`, relative to someone who has `No College Degree`.

Since $\hat{\beta}_1$ is very close to zero and the p value for $\hat{\beta}_2$ greater than 0.05(not significantly small), we conclude that there is no asscociation between `Education Level` and `Mental Health`.

If we see the odds ratio corresponding to `Education Level` and it 95 percent confidence intervel;

```
exp(coef(fit1))
```

```
##     (Intercept) Education_Level
##          0.4444          1.0248
```

```
exp(confint.default(fit1))
```

```
##                   2.5 % 97.5 %
## (Intercept)     0.3062  0.645
## Education_Level 0.6153  1.707
```

The odds ratio is colse to 0 and 1 belongs to the 95 percent confidence intervel.

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

```
## 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 $\hat{\beta}_0 = -0.77389$, represents the log odds of having a `Depression` or `Severly Depression` for someone who has `No College Degree` and with gender `Female`.

The estimated parameter $\hat{\beta}_1 = 0.03093$, represents the log odds ratio of having a `Depression` or `Severly Depression` for someone who who has a `Undergrad Degree` or `Post-grad Degree`, relative to someone who has `No College Degree`, for both `Male` and `Female`.

The estimated parameter $\hat{\beta}_2 = -0.09946$, 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`.

Since $\hat{\beta}_1$ is very close to zero and $\hat{\beta}_2$ is somewhat not clse to zero, we think there is association between `Gender` and `Mental Health`, but not sure about the association between `Education Level` and `Mental Health`.

See odds ratios and confidence intervals:

```
exp(coef(fit2))
```

```
##     (Intercept) Education_Level          Gender
##          0.4612          1.0314          0.9053
```

```
exp(confint.default(fit2))
```

```
##                   2.5 % 97.5 %
## (Intercept)     0.3034 0.7011
## Education_Level 0.6185 1.7200
## Gender          0.5381 1.5232
```

We can see that the odds ratios for `Education Level` and `Gender` are all close to 1, and both of their 95 percent confidence interval contains 1. Then we conclude that there are no association between `Education Level` and `Mentel Health`, as well as `Gender` and `Mental Healtl`.

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:

```
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:

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

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

```
## 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 $\hat{\beta}_0 = -0.810930$, represents the log odds of having a `Depression` or `Severly Depression` for someone who has `No College Degree`.

The estimated parameter $\hat{\beta}_1 = 0.046324$, 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 $\hat{\beta}_2 = -0.005831$, 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`.

From the value of $\hat{\beta}_1$ and $\hat{\beta}_2$ we can find that D_1 is has a stronger association with `Mental Health` than D_2, so we think that D_1 has a association with `Mental Health` but D_2 does not.

See odds ratios and confidence intervals:

```
exp(coef(fit1))
```

```
## (Intercept)          D1          D2
##      0.4444      1.0474      0.9942
```

```
exp(confint.default(fit1))
```

```
##              2.5 % 97.5 %
## (Intercept) 0.3062  0.645
## D1          0.5810  1.888
## D2          0.5159  1.916
```

Then we conclude that both D_1 and D_2 are not associated with `Mental Health`.

b).

```
## 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 $\hat{\beta}_0 = -0.774535$, represents the log odds of having a `Depression` or `Severly Depression` for someone who has `No College Degree` and with gender `Female`.

The estimated parameter $\hat{\beta}_1 = 0.049738$, 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 $\hat{\beta}_2 = 0.004591$, 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 $\hat{\beta}_3 = -0.097709$, 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`.

Same as previous questions, see odds ratios and confidence intervals:

```
exp(coef(fit2))
```

```
## (Intercept)          D1          D2      Gender
##      0.4609      1.0510      1.0046      0.9069
```

```
exp(confint.default(fit2))
```

```
##               2.5 % 97.5 %
## (Intercept) 0.3032 0.7008
## D1          0.5828 1.8954
## D2          0.5200 1.9406
## Gender      0.5386 1.5270
```

Then we conclude that D_1, D_2, `Gender` are not associated with `Mental Health`.

c).

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

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

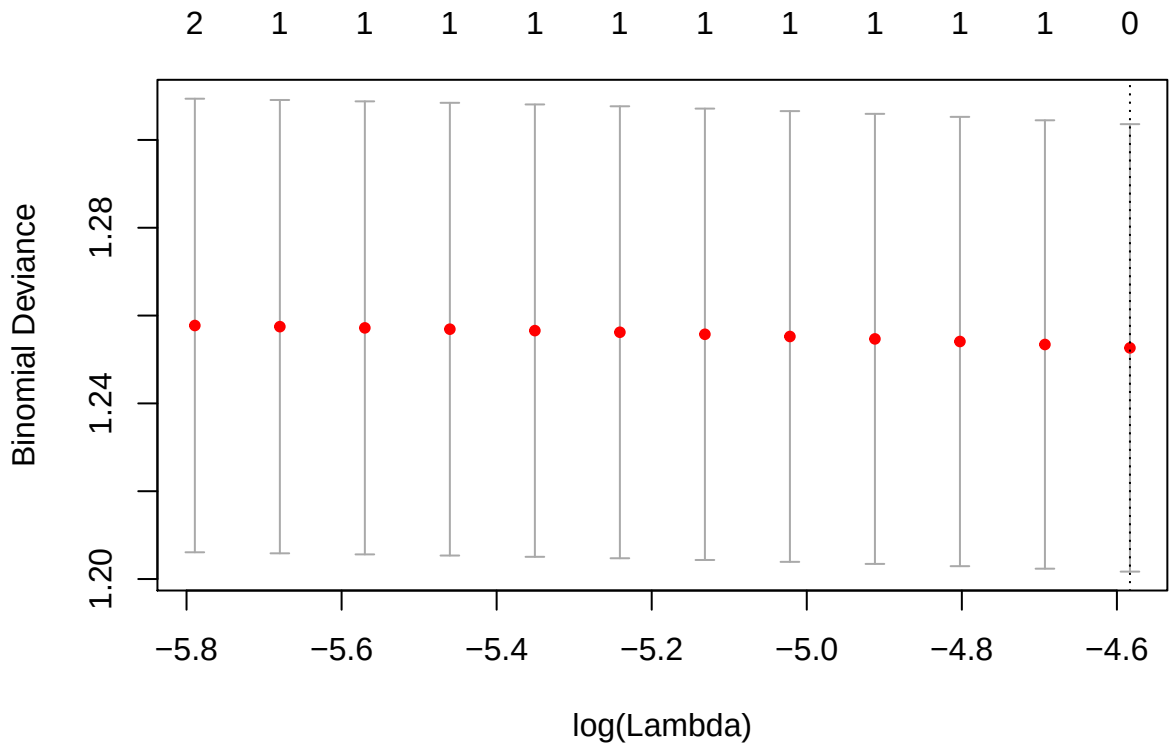
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  fit2$y, fitted(fit2)
## X-squared = 0.0409, df = 2, p-value = 0.9798
```

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).

```
## 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: $logit(p_{DepressedorSeverelyDrepressed|EducationLevel,Gender}) = -0.7979261$