Homework 3

Tim Vigers

26 February 2019

```
# Read in
copd <- read.csv("~/Documents/School/UC Denver/Biostatistics/Biostatistical Methods 2/Homework
s/Homework 3/hw3.txt", sep="")
# Format factor columns
copd[,c("copd","gender","smoker")] <- lapply(copd[,c("copd","gender","smoker")], as.factor)
```

1. COPD Models

```
# Model 1 with BMI squared
mod1 <- glm(copd ~ age + gender + smoker + BMI + BMIsquared, data = copd,family = "binomial")
```

a. Determine whether COPD is significantly associated with BMI squared using Wald statistic.

```
# Calculate Wald statistic
wald <- (coef(mod1)[6]/sqrt(diag(vcov(mod1)))[6])^2
wald
```

```
## BMIsquared
##   7.308759
```

$$\text{Wald statistic} = 7.309$$

Test against chi-squared distribution (1 DF):

```
1 - pchisq(7.309,1)
```

```
## [1] 0.00686101
```

Based on the Wald statistic, BMI squared is significantly associated with COPD (p = 0.007).

b. Determine whether COPD is significantly associated with BMI squared using LRT.

```
# Model without BMI squared
mod0 <- glm(copd ~ age + gender + smoker + BMI, data = copd,family = "binomial")
# LL for model 0 and model 1
LL0 <- logLik(mod0)
LL1 <- logLik(mod1)
LL0
```

```
## 'log Lik.' -514.709 (df=5)
```

```
LL1
```

```
## 'log Lik.' -510.7325 (df=6)
```

Likelihood ratio statistic = $2(LL1 - LL0) = 7.953$

Test against chi-squared distribution (1 DF):

```
1 - pchisq(7.953,1)
```

```
## [1] 0.004800773
```

Check LRT with R:

```
anova(mod0,mod1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: copd ~ age + gender + smoker + BMI
## Model 2: copd ~ age + gender + smoker + BMI + BMIsquared
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       829     1029.4
## 2       828     1021.5  1   7.9529 0.004801 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. AUC (c index)

Compare the predictive accuracy of each model. Model 0 (without BMI squared):

```
copd$prob_mod0 <- predict(mod0,type = "response")
roc_mod0 <- pROC::roc(copd ~ prob_mod0, data = copd)
roc_mod0
```

```
##
## Call:
## roc.formula(formula = copd ~ prob_mod0, data = copd)
##
## Data: prob_mod0 in 352 controls (copd 0) < 482 cases (copd 1).
## Area under the curve: 0.705
```

Model 1 (with BMI squared):

```
copd$prob_mod1 <- predict(mod1,type = "response")
roc_mod1 <- pROC::roc(copd ~ prob_mod1, data = copd)
roc_mod1
```

```
##
## Call:
## roc.formula(formula = copd ~ prob_mod1, data = copd)
##
## Data: prob_mod1 in 352 controls (copd 0) < 482 cases (copd 1).
## Area under the curve: 0.7121
```

$$\text{AUC without BMI}^2 = 0.705 \text{AUC with BMI}^2 = 0.712$$

The difference in predictive power between the two models is not particularly large, but the model with BMI squared is slightly better.

d. Is there evidence that COPD has a quadratic relationship with BMI?

Both the Wald statistic and likelihood ratio test suggest that BMI^2 contributes significantly to the model. Also, the model including BMI^2 has slightly higher predictive power than the model without BMI^2. Therefore, all of the evidence suggests that there is a quadratic relationship between BMI and presence of COPD.

e. Why do you think the BMI variable was centered?

I would guess that there was an issue with multicollinearity for BMI. Mean-centering a predictor variable can reduce multicollinearity, and does not require categorizing the variable (so you aren't throwing out as much information).

f. Calculate and interpret the estimated odds ratio for the effect of BMI on COPD for a patient with average BMI.

By inverting a Wald test:

$$95\%\ \text{CI} = (e^{\hat{\beta}_{\text{BMI}} - 1.96 * SE(\hat{\beta}_{\text{BMI}})}, e^{\hat{\beta}_{\text{BMI}} + 1.96 * SE(\hat{\beta}_{\text{BMI}})}) = (e^{-0.045 - 1.96 * 0.015}, e^{-0.045 + 1.96 * 0.015}) = (e^{-0.074}, e^{-0.016}) = (0.928, 0.985)$$

Check with R:

```
confint.default(mod1)
```

```
##                    2.5 %       97.5 %
## (Intercept) -5.642667879 -3.077391381
## age          0.058209737  0.097568453
## gender1     -0.515840010  0.079700013
## smoker1     -0.642877714  0.045112409
## BMI         -0.074759058 -0.015506197
## BMIsquared   0.001104102  0.006925159
```

Not exactly the same, but probably different due to rounding error.

By find a CI by inverting a LRT, solve the equation:

$$2 \log(\frac{L(\hat{\beta}_0, \hat{\beta}_1)}{L(\tilde{\beta}_0 | \beta_1 = \beta_1^*)}) = 3.841$$

This is tough to solve by hand, so use R's confint() function:

```
confint(mod1)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %       97.5 %
## (Intercept) -5.661287860 -3.093066657
## age          0.058487351  0.097888228
## gender1     -0.516896829  0.079026899
## smoker1     -0.642388194  0.046217286
## BMI         -0.075154570 -0.015821937
## BMIsquared   0.001194604  0.007039182
```

The interpretation for the Wald statistic CI is that for every 1 unit increase in BMI, the odds of developing COPD change 0.956 (95% CI: 0.928,0.985) fold on average (p = 0.003). For the LRT-based confidence interval, the interpretation is similar: For every 1 unit increase in BMI, the odds of developing COPD change 0.956 (95% CI: 0.928,0.984) fold on average (p = 0.003).

The confidence intervals are exactly the same when rounding to three digits. Because neither interval contains 0 (or 1 after exponentiation), they are both saying that the effect of BMI is significant. Also, the distributions of both the Wald and likelihood ratio test statistics are asymptotically equal (both chi square), and this large sample size allows us to assume that asymptotic theory holds.

2. Rickert et al.

a. Interpretation

In this study, group is the treatment variable, and refers to no education program (0) or the program being investigated (1). On average, participating in the education program increases the odds of an adolescent purchasing condoms 4.04 fold when adjusting for gender, SES, and total number of partners. The confidence interval means that we can be 95% certain that the odds change between 1.17 and 13.9 fold.

b. Parameter estimates

Because the estimates are provided on the OR scale, for each parameter except the intercept we just need to take the log of the OR:

$$\text{logitP}(Y_i = 1) = \hat{\beta}_0 + \hat{\beta}_{\text{group}} + \hat{\beta}_{\text{gender}} + \hat{\beta}_{\text{SES}} + \hat{\beta}_{\text{partners}} = \hat{\beta}_0 + \log(4.04) + \log(1.38) + \log(5.82) + \log(3.22)$$

So,

$$\hat{\beta}_{\text{group}} = 1.396\hat{\beta}_{\text{gender}} = 0.322\hat{\beta}_{\text{SES}} = 1.761\hat{\beta}_{\text{partners}} = 1.169$$

c. $\hat{\beta}_0$

In order to find $\hat{\beta}_0$ we would need to know how many female, low SES, adolescents who have had 0 sexual partners and didn't receive the education program bought condoms, and how many didn't buy them.

d. SE($\hat{\beta}_1$)

To find the standard error, we rearrange the following equation. It's only necessary to solve the lower or upper bound, since SE is constant:

$$95\%\ \text{CI} = (e^{\hat{\beta}_{\text{group}} - 1.96 * SE(\hat{\beta}_{\text{group}})}, e^{\hat{\beta}_{\text{group}} + 1.96 * SE(\hat{\beta}_{\text{group}})})$$

So:

$$1.17 = e^{\hat{\beta}_{\text{group}} - 1.96 * SE(\hat{\beta}_{\text{group}})} \log(1.17) = \hat{\beta}_{\text{group}} - 1.96 * SE(\hat{\beta}_{\text{group}}) SE(\hat{\beta}_{\text{group}}) = \frac{\log(1.17) - \hat{\beta}_{\text{group}}}{-1.96} = 0.632$$

Probably worth checking with the upper bound as well though:

$$13.9 = e^{\hat{\beta}_{\text{group}} + 1.96 * SE(\hat{\beta}_{\text{group}})} \log(13.9) = \hat{\beta}_{\text{group}} + 1.96 * SE(\hat{\beta}_{\text{group}}) SE(\hat{\beta}_{\text{group}}) = \frac{\log(13.9) - \hat{\beta}_{\text{group}}}{1.96} = 0.630$$

These are close enough that the discrepancy is most likely rounding error, so it's safe to say $SE(\hat{\beta}_{\text{group}}) \approx 0.63$.

e. Check OR for gender

We can check this interval and estimate using the same techniques as above. First calculate SE based on the lower bound:

$$1.23 = e^{\hat{\beta}_{\text{gender}} - 1.96 * SE(\hat{\beta}_{\text{gender}})} \log(1.23) = \hat{\beta}_{\text{gender}} - 1.96 * SE(\hat{\beta}_{\text{gender}}) SE(\hat{\beta}_{\text{gender}}) = \frac{\log(1.23) - \hat{\beta}_{\text{gender}}}{-1.96} = 0.059$$

And then the upper bound:

$$12.88 = e^{\hat{\beta}_{\text{gender}} + 1.96 * SE(\hat{\beta}_{\text{gender}})} \log(12.88) = \hat{\beta}_{\text{gender}} + 1.96 * SE(\hat{\beta}_{\text{gender}}) SE(\hat{\beta}_{\text{gender}}) = \frac{\log(12.88) - \hat{\beta}_{\text{gender}}}{1.96} = 1.136$$

In this case the estimates of SE are way off, which means something is wrong either the interval or the estimate. If we assume that the interval is correct, we can see if 1.38 is in fact the log odds ratio by plugging it in for $\hat{\beta}_{\text{gender}}$.

Lower:

$$1.23 = e^{1.38 - 1.96 * SE(\hat{\beta}_{\text{gender}})} \log(1.23) = 1.38 - 1.96 * SE(\hat{\beta}_{\text{gender}}) SE(\hat{\beta}_{\text{gender}}) = \frac{\log(1.23) - 1.38}{-1.96} = 0.598$$

Upper:

$$12.88 = e^{1.38 + 1.96 * SE(\hat{\beta}_{\text{gender}})} \log(12.88) = 1.38 + 1.96 * SE(\hat{\beta}_{\text{gender}}) SE(\hat{\beta}_{\text{gender}}) = \frac{\log(12.88) - 1.38}{1.96} = 0.597$$

These estimates for $SE(\hat{\beta}_{\text{gender}})$ are the same (within rounding error), which suggests that the estimate for gender was mistakenly reported on the log odds scale. So, assuming this is actually what happened, the true OR value is $e^{1.38} \approx 3.97$