Binomial Prediction of Hodgkin's Survival

Greg Johnson

8/14/2017

```
HD<-read.csv("data/Hodgkins Data.csv")
```

Our data consist of the survival information of 342 subjects suffering from Hodgkin's. We know their white and red blood cell counts, whether they have HIV, their gender, the stage of Hodgkin's and whether they were alive after 5 years (binary outcome).

HIV as a Risk Factor

To assess the impact of having previously had mono/HIV as a risk factor for death prior to five years of a Hodgkin's diagnosis, we need a model that predicts a binary outcome variable. Thus we will fit a generalized linear model with a logit function as the link i.e. a logistic regression. To assess the impact of (previously having had) mono/HIV whilst controlling for the other predictors at our disposal (e.g. gender, age, stage of Hodgkin's, etc.), we can examine the parameter estimating the unique effect of mono/HIV on the probability of dying from Hodgkin's (within 5 years) whilst holding all other predictors constant.

As with any model, we need to check assumptions. Fortunately, logistic regression requires very few assumptions compared to ordinary regression. We will simply assume independence of observations and linearity of predictors and log odds but we can check for multicollinearity.

```
cor(HD[,c(3,4,7)])
```

```
##             age         rbc        wbc
## age  1.00000000 -0.03120439  0.2005624
## rbc -0.03120439  1.00000000 -0.1516761
## wbc  0.20056244 -0.15167614  1.0000000
```

We don't have a multicollinearity problem.

Lastly, since logistic regression is a generalized linear model, we don't have a clean analytic solution to the optimal parameter estimates - we need to use nonlinear estimation (i.e. MLE in this case) which requires a larger sample size. With an n= 342 we will be fine.

Now we can fit the logistic regression.

```
log.fit<-glm(alive5yr~stage+age+rbc+gender+HIV.mono+wbc,family=binomial(link="logit"),data=HD)
summary(log.fit)
```

```
##
## Call:
## glm(formula = alive5yr ~ stage + age + rbc + gender + HIV.mono +
##     wbc, family = binomial(link = "logit"), data = HD)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.83130   0.06567   0.20310   0.40163   2.15961
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.7864412  1.9636380   7.021  2.2e-12 ***
## stageII     -0.1620048  0.5517289  -0.294 0.769040
## stageIII    -0.7330153  0.5370526  -1.365 0.172289
## stageIV     -1.8005177  0.5418442  -3.323 0.000891 ***
## age         -0.0392651  0.0121003  -3.245 0.001175 **
## rbc          0.5170553  0.1435253   3.603 0.000315 ***
## genderM     -0.3078853  0.3875373  -0.794 0.426924
## HIV.monoY   -1.3861293  0.5872775  -2.360 0.018262 *
## wbc         -0.0011855  0.0001851  -6.405  1.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 352.02  on 341  degrees of freedom
## Residual deviance: 182.65  on 333  degrees of freedom
## AIC: 200.65
##
## Number of Fisher Scoring iterations: 6
```

The parameter estimate corresponding to the unique effect of having had HIV/mono is statistically significant, from which we can infer that the unique effect is likely not zero at the population level. We can get a nice interpretation from the parameter estimate in terms of the probability of dying from Hodgkin's within 5 years.

```r
1/exp(coef(log.fit)[["HIV.monoY"]])
```

```
## [1] 3.99934
```

We interpret the estimated HIV/mono parameter for the layperson as follows:

Having had HIV/mono previously results, on average, in a 299.9 % increase in the odds of dying from HD. This is true when we ignore gender, stage of Hodgkin's, and white and red blood cell count. Thus the data and our model identify HIV/mono as a huge risk factor for early death once someone has been diagnosed with Hodgkin's.

Stage IV vs. Stage I Hodgkin's

Say we want to know how much more at risk is an individual diagnosed with stage IV Hodgkins as opposed to someone diagnosed with stage I Hodgkins.

Similar to the previous problem, we can look at the estimated stage IV parameter. Because of how we coded the stage variable, the parameters or stage II, stage III, and stage IV are estimating the difference in probability of survival compare to stage I.

```r
1/exp(coef(log.fit)[["stageIV"]])
```

```
## [1] 6.05278
```

There is a 505% increase in the odds of dying from HD for someone with stage IV Hodgkins verus someone with stage I Hodgkin's. This aggrees with our intuition that the further the progression of the disease, the more likely we are to die from it (or some other complication).

Probability of 5-Year Survival

Say we have a 33-year old male with a red blood cell count of 4.3 cells/ul, a white blood cell count of 12,000 cells/mm3, no history of mono and no indication of HIV has been diagnosed with stage II Hodgkins. Based on the data available to us, we can predict the probability of being alive five years from now.

We can simply plug these values for the predictors into the inverse logit function to get a prediction of the probability. R has a great canned function for this:

```
  pr<-predict(log.fit,newdata=data.frame(stage="II",age=33,rbc=4.3,gender="M",HIV.mono="N",wbc=12000),t

  pr[[1]]
```

```
##         1 
## 0.5045477
```

Our intuitive expanation is: on average, 33-year old males with rbc 4.3, wbc of 12000, no history of mono/HIV, and stage II Hodgkins have a probability of 50.45% of being alive with Hodgkin's within the next 5 years. Out of all people that match this description, 95% of them have a probability of being alive with Hodgkin's between:

```
pr[[1]]+c(-1.96*pr[[2]],1.96*pr[[2]])
```

```
##         1         1 
## 0.2379455 0.7711499
```