HW7 (Due May 23)

In this HW, we will practice generalized linear model (GLM).

1. Using “bpd dataset”, fit logistic regression to model BPD as 1) a function of birth weight and 2) a function of birth weight, gestational age and toxemia. Summarize key results and comment.

BPD as function of birth weight/100:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 4.03429 0.69571 5.799 6.68e-09 \*\*\*

weight -0.42291 0.06408 -6.600 4.11e-11 \*\*\*

As birth weight increases the log-odds of BPD decrease by -0.42 for every 100 unit increase in weight. For a 100 gram increase in birth weight the odds of BPD decrease by 0.66 (exp^-0.42). This is significant at the 0.05 level.

BPD as function of birth weight/100, gestational age and toxemia:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 13.93608 2.98255 4.673 2.98e-06 \*\*\*

weight -0.26436 0.08123 -3.254 0.00114 \*\*

gestage -0.38854 0.11489 -3.382 0.00072 \*\*\*

toxemia -1.34379 0.60750 -2.212 0.02697 \*

Here all covariates are significantly (at the 0.05 level) associated with the probability of BPD. The (adjusted) effect of weight/100 in the log-odds is -0.26, which translates into that for a 100 gram increase in birth weight the odds of BPD decrease by 0.77 (exp^-0.26), when adjusting for gestational age and toxemia.

LRT comparing both models:

Model 1: bpd ~ weight

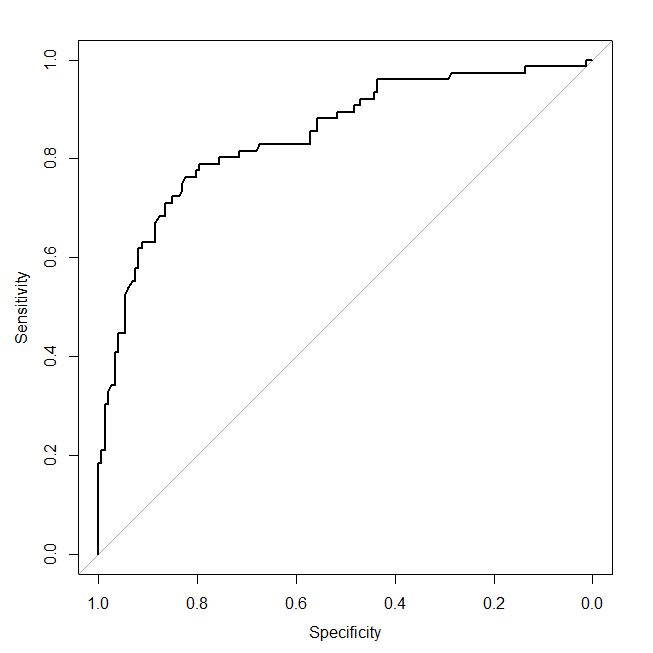
Model 2: bpd ~ weight + gestage + toxemia

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 221 223.72

2 219 203.71 2 20.013 4.51e-05 \*\*\*

Model that adjust for both gestational age and toxemia fits better to the data. This is the ROC curve for this model:



With an AUC of 0.85

1. Using “chd dataset”, fit Poisson regression to model CHD as 1) a function of smoking and 2) a function of smoking, behavior type and blood pressure. Summarize key results and comment.

CHD rate a function of smoking:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.799334 0.088520 -54.217 < 2e-16 \*\*\*

smoke 0.031754 0.005624 5.646 1.64e-08 \*\*\*

Smoking a package of cigarettes (20 cigarettes) increases the rate of CHD by 1.89 (exp^20\*0.03) times when compared to non-smokers. This is significant at the 0.05 level.

CHD rate a function of smoking, behavior type and blood pressure:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5.420153 0.130813 -41.434 < 2e-16 \*\*\*

smoke 0.027344 0.005614 4.871 1.11e-06 \*\*\*

behavior 0.752555 0.136202 5.525 3.29e-08 \*\*\*

bp 0.753376 0.129240 5.829 5.57e-09 \*\*\*

When adjusting by behavior type and blood pressure the effect of smoking is still associated with an increase in the rate of CHD. Now, smoking one pack of cigarettes (20 cigarettes) is associated with a 1.73 (exp^0.27) times higher risk of CHD when compared to non-smokers. This effect is significant at the 0.05 level.

Comparing both models using LRT:

Model 1: chd ~ smoke + offset(logpyrs)

Model 2: chd ~ smoke + behavior + bp + offset(logpyrs)

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 14 89.382

2 12 21.240 2 68.142 1.596e-15 \*\*\*

Adjusted model seems to have better fit to the data at hand.

1. Fit #2 with negative binomial. How would you compare the two models, and which do you prefer and why?

Beforehand I would say I prefer the negative binomial one (or any model that adjusts for over-dispersion such us estimating a (phi) dispersion parameter, or estimating a normal distribution of the errors, or sandwich estimator of SE’s), as it takes into account any possible over-dispersion of the count data, which in turn yields more realistic standard errors of parameter estimates and hence inferences are not as liberal as the ones of the Poison model without over-dispersion adjustment.

Now, when fitting the negative binomial for the model with all predictors, it did not converge (still gave estimates but I suppose these are not very reliable, especially standard errors). Here they are anyway:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5.419487 0.131613 -41.177 < 2e-16 \*\*\*

smoke 0.027354 0.005653 4.839 1.30e-06 \*\*\*

behavior 0.751869 0.136982 5.489 4.05e-08 \*\*\*

bp 0.753046 0.130073 5.789 7.06e-09 \*\*\*

Now, for the model with only smoking, here are the results:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.52655 0.24669 -18.349 <2e-16 \*\*\*

smoke 0.02626 0.01365 1.925 0.0543 .

We see that smoking is now borderline non-significant, mostly due to an increase in the SE of the regression coefficient for smoking. Nevertheless, most of this over-dispersion could be due the other two variables not included in this model (as shown in the book). Curiously enough, in the book they did not fit NB with all of the predictors…perhaps because it wouldn’t converge. Instead they estimated the extra dispersion parameter, phi.

1. Using “arthritis dataset”, fit proportional odds regression model for global impression scale at month 6 with two covariates, age and treatment. Summarize key results and comment. [We did not learn this in class but worth trying!]

Here are the results of the proportional odds model of global impression on age and treatment:

Coefficients:

Value Std. Error t value

age 0.02048 0.009833 2.083

trt -0.60791 0.214223 -2.838

Intercepts:

Value Std. Error t value

1|2 -1.2119 0.5317 -2.2793

2|3 0.5251 0.5249 1.0003

3|4 2.1494 0.5381 3.9944

4|5 4.1365 0.6122 6.7571

R does a different (compared to the one presented in the book) parameterization of the model, which is

Placing a negative sign in front of the makes the regression coefficients for proportional odds model to be associated with the probability of being in the higher numbered categories.

So, based on this model, we see that as baseline age increases the log-odds of being in the higher impression scale categories (meaning poorer arthritis statues) increase, for every year increase in age the odds of less favorable response increase by 1.02 (exp^0.02) for either treatment level. Based on the t-value, this effect is significant at 0.05 level.

Likewise we see that, when adjusted for baseline age, treatment decreases the odds of being in higher impression scale category, with an estimated odds ratio of 0.54 (exp^-0.61). Based on the t-value, this effect is significant at 0.05 level.

1. In #4,
2. Comment on including age in the model.

By including age we’re controlling by its potential confounding effect on the association between the exposure of interest and the outcome. This is especially true considering we have different baseline ages.

1. Some suggest using age vs. age10=age/10. Comment on this issue.

Arguably this would make the results more interpretable, as it is easier to visualize the effect of 10 years than the effect of just 1 (considering the small effect size for age in years, 0.02). In terms of model fit, it shouldn’t make any difference fitting a model with age/10 or fitting a model with age and then multiplying the effect of age for 10 (SE of age effect can accordingly be adjusted by estimating the variance of 10\*). Here are the results of such a model:

Coefficients:

Value Std. Error t value

I(age/10) 0.2048 0.09832 2.083

trt -0.6079 0.21422 -2.838