HW 1

Steph Holm and Shelley Facente

February 7, 2019

Question 1:

Using the notation for generalized linear models presented in class, write out the equations for each of the 3 models, in terms of the variables in the dataset. Clearly define all parameters in each of the models. (15 points)

a. Logistic Regression Model:

Model:

$$logit(Pr(Y=1|X=x) = \beta_1 + (\beta_2 * BMI_{underweight}) + (\beta_3 * BMI_{overweight}) + (\beta_4 * BMI_{obese}) + (\beta_5 * currsmoke) + (\beta_6 * age) + (\beta_7 * female) + (\beta_8 * education_2) + (\beta_9 * education_3) + (\beta_{10} * education_4)$$

Equation:

$logit(Pr(Y=1|X=x) =$ `-1.1750028` + (`-0.3315533` * $BMI_{underweight}$) + (`0.4502186` * $BMI_{overweight}$) + (`1.2644009` * BMI_{obese}) + (`-0.1859176` * $currsmoke$) + (`0.0363935` * age) + (`0.1960934` * $female$) + (`-0.095462` * $education_2$) + (`-0.1699546` * $education_3$) + (`-0.5327318` * $education_4$)

b. Log-Binomial Model:

Model (Model would not converge despite providing starting values):

$$log(Pr(Y=1|X=x) = \beta_1 + (\beta_2 * BMI_{underweight}) + (\beta_3 * BMI_{overweight}) + (\beta_4 * BMI_{obese}) + (\beta_5 * currsmoke) + (\beta_6 * age) + (\beta_7 * female) + (\beta_8 * education_2) + (\beta_9 * education_3) + (\beta_{10} * education_4)$$

c. "Modified Poisson" Model:

Equation:

$log(Pr(Y=1|X=x) =$ `-0.9806268` + (`-0.1443865` * $BMI_{underweight}$) + (`0.1504552` * $BMI_{overweight}$) + (`0.3186672` * BMI_{obese}) + (`-0.0602246` * $currsmoke$) + (`0.0111295` * age) + (`0.0556455` * $female$) + (`-0.0247199` * $education_2$) + (`-0.0470482` * $education_3$) + (`-0.1919009` * $education_4$)

Question 2:

For the logistic regression model write the log-likelihood function in general terms - variable names and parameters - no data values. (you may reference your answer to question 1(a) to make the notation concise). (10 points)

$$\mathcal{L}(\beta|y,x) = \sum_i log\left[\frac{exp(logit(Pr(Y=1|X=x_i))}{1+exp(logit(Pr(Y=1|X=x_i))}\right]^{y_i} \text{ x } log\left[1-\frac{exp(logit(Pr(Y=1|X=x_i))}{1+exp(logit(Pr(Y=1|X=x_i))}\right]^{1-y_i}$$

Question 3:

Using the results from the models you estimated above, complete the following table. (15 points)

Table 2: Estimates of adjusted relative risk estimates of the association between baseline BMI status and incident hypertension. The Framingham Cohort Study. 1948-1972, Framingham, MA.

BMI	LogisticOR (95% CI)	Log-binomial RR (95% CI)	Poisson RR (95% CI)
< 18.5	`0.72` (`0.37`, `1.38`)	*Model*	`0.87` (`0.64`, `1.18`)
18.5 - 29.4	*ref*	*did*	*ref*
25.0 - 29.9	`1.57` (`1.29`, `1.91`)	*not*	`1.16` (`1.09`, `1.24`)
≥ 30	`3.54` (`2.29`, `5.47`)	*converge*	`1.38` (`1.27`, `1.49`)

Question 4:

Without considering any of the results, how well would you think the OR from the logistic model would approximate the RR from these data? When considering results from all of the models, which one do you think best characterizes the risk ratio? Why? (10 points)

We would expect the OR will be further from the null than the RR, and since this outcome is not rare (incidence = `0.658`*), we would expect this overestimate to potentially be substantial.*

We think the best model for characterizing the risk ratio is the "Modified Poisson" Model, since it allows for direct estimation of the RR using log as the canonical link, and while statistical efficiency is compromised when using this model compared to log-binomial regression, it is more likely to successfully converge.

Question 5:

Answer the following regarding the standardized measures of association: (15 points)

a.

Using results from the logistic model, what is the standardized risk ratio comparing the population where everyone is obese to the one where everyone is ideal weight? Explain how this is conceptually different than the corresponding measures of association in question 3.

From the logistic model, the standardized risk ratio when comparing a counterfactual population where everyone is obese to one where everyone is ideal weight is `1.39`*. This is conceptually different from the corresponding measures of association in question 3 because in this model we are standardizing conditional on the distribution of covariates that exist in our underlying population. In question 3, the model was conditional holding other covariates constant.*

b.

Using results from the logistic model, what is the standardized risk difference comparing the population where everyone is obese to the one where everyone is ideal weight?

From the logistic model, the standardized risk ratio when comparing a counterfactual population where everyone is obese to one where everyone is ideal weight is `0.236`*.*

c.

What are the risk ratio and risk difference from the modified Poisson model? Between the logistic and Poisson model, which one would you choose to report and why?

From the "Modified Poisson" Model, the standardized risk ratio when comparing a counterfactual population where everyone is obese to one where everyone is ideal weight is `1.375`. *The standardized risk difference is* `0.226`. *We would choose to report Poisson because Poisson is calculating relative risks directly, whereas in a logistic model we are approximating relative risks using an odds ratio, and particularly in the case of a common outcome such as hypertension, this is inadvisable.*

Question 6:

Answer the following questions about penalized likelihood estimation for a logistic regression model as above: (20 points)

a.

Write out a penalized log-likelihood function using a general expression for a quadratic penalty for the logistic regression model you defined in question 2. Clearly define all parameters, including those on the penalty function. (You may use either vector or summation notation.)

$$\mathcal{L}_p(\beta|y,x) = \mathcal{L}(\beta|y,x) - (\beta - m)'R(\beta - m)/2$$

The penalized log-likelihood is equal to the usual log-likelihood (as defined in question 2), minus a quadratic penalty. The penalty is the inverse of the difference between the set of all beta coefficients and the vector m (means of the prior betas) multiplied by the tuning parameter (the prior precisions), again multiplied by the difference between beta and m, divided by two.

b.

In one sentence, explain in words what the penalty function does in penalized maximum likelihood estimation.

The penalty function adjusts your model to incorporate prior information, such that the likelihood values are increasingly decreased the further away they are from your prior betas.

c.

Which parameters are being penalized in the above? Why?

The likelihood of possible beta values, because you are calculating revised maximum likelihood that has been influenced by the prior information you have fed the model.

d.

If the prior precision corresponding to each beta in this model is equal to 0 (an infinitely wide prior confidence interval), show how this implies that the penalized likelihood in 6 (a) is equivalent to the one in question 2, regardless of the value of m.

$$\mathcal{L}_p(\beta|y,x) = \mathcal{L}(\beta|y,x) - (\beta - m)'0^*(\beta - m)/2 = \mathcal{L}(\beta|y,x)$$

As can be seen in the equation above, when the prior precision corresponding to each beta (R) is equal to zero, the penalty function drops to zero, leaving you with the usual log-likelihood.

Question 7.

Using the results from the models you estimated above, complete the following table. What happens to the OR and confidence intervals as the prior precision increases? (15 points)

Table 3: Penalized maximum likelihood estimates of the association between baseline BMI status and incident hypertension under different parameterizations of a quadratic penalty function. The Framingham Cohort Study. 1948-1972, Framingham, MA.

BMI	Unpenalized (Q2)	Scenario I: OR (95% CI)	Scenario II: OR (95% CI)	Scenario III: OR (95% CI)
< 18.5	0.72 (0.37, 1.38)	0.82 (0.45, 1.48)	0.89 (0.52, 1.55)	1.05 (0.67, 1.65)
18.5 - 29.4	*ref*	*ref*	*ref*	*ref*
25.0 - 29.9	1.57 (1.29, 1.91)	1.56 (1.28, 1.89)	1.54 (1.28, 1.87)	1.51 (1.26, 1.82)
≥ 30	3.54 (2.33, 5.58)	3.37 (2.26, 5.15)	3.23 (2.21, 4.83)	2.95 (2.1, 4.19)

As the prior precision increases, we expect the confidence intervals to decrease and the OR to be a better estimation of the true odds ratio.