Regression Analysis - Quiz 5

Wei-Chen Chang r12227118

Due: 2023-11-26

Q1.

Check if logit, probit, and complimentary log-log link functions are satisfactory for the model.

Report statistical inference for the models, and the deviance and Pearson chi-square statistic for model goodness-of-fit.

Check the model assumptions and residuals, including possible transformations of the variable. Examine prediction quality by prediction error for the models. State your conclusion.

Ans

First, the data and the visualization of the data can be seen in Table 1 and Figure 1.

The regression results can be seen in Table 1.

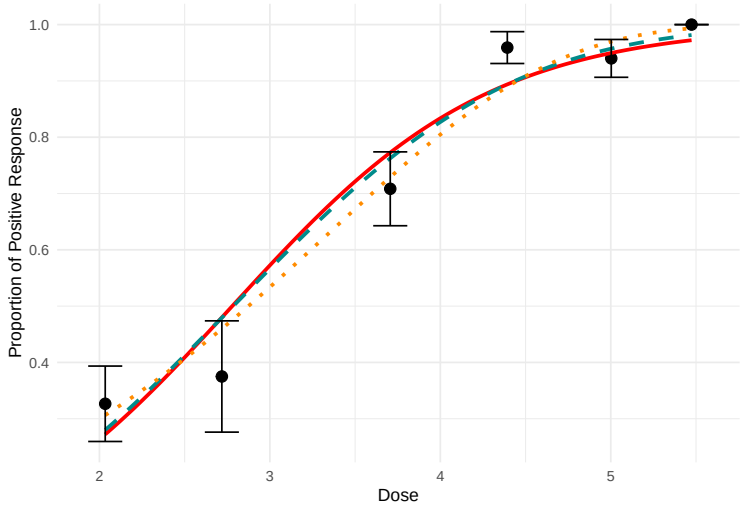

Figure 1: The Data and Fitting Curves. (error bar represent sample s.e., solid line for logit, dashed-line for probit, dot for Complementary log-log)

The residuals can be seen in Figure 2. It' noteworthy to point out that group 4 has a large positive residual in all the models.

For model choice, except Pearson's chi-square statistic, deviance, leave-one-out cross validation is used to calculate Sum of Prediction Squared Error (SPSE). All these results can be seen in 1 The calculation of SPSE is as below: First, extract data except the ith observation to fit the model, then calculate the squared

Table 1: Binominal Regression Models Results

	Dependent variable:					
	positive_prop					
	logistic	*probit*	*glm: binomial*	*glm: quasi-binomial*		
			link = cloglog	*link = logit*	*link = probit*	*link = cloglog*
	(1)	(2)	(3)	(4)	(5)	(6)
dose	1.320***	0.776***	0.761***	1.320**	0.776**	0.761**
	(0.176)	(0.096)	(0.100)	(0.206)	(0.109)	(0.110)
Constant	−3.669***	−2.160***	−2.554***	−3.669**	−2.160**	−2.554**
	(0.593)	(0.339)	(0.393)	(0.693)	(0.381)	(0.435)
Observations	6	6	6	6	6	6
Log Likelihood	−11.829	−11.605	−11.387			
Akaike Inf. Crit.	27.659	27.210	26.774			
Pearson's chi-square	5.454	5.072	4.900	5.454	5.072	4.900
Deviance	6.196	5.747	5.311	6.196	5.747	5.311
SPSE	0.071	0.060	0.030	0.071	0.060	0.030
Note:						*p<0.0; **p<0.01; ***p<0.001

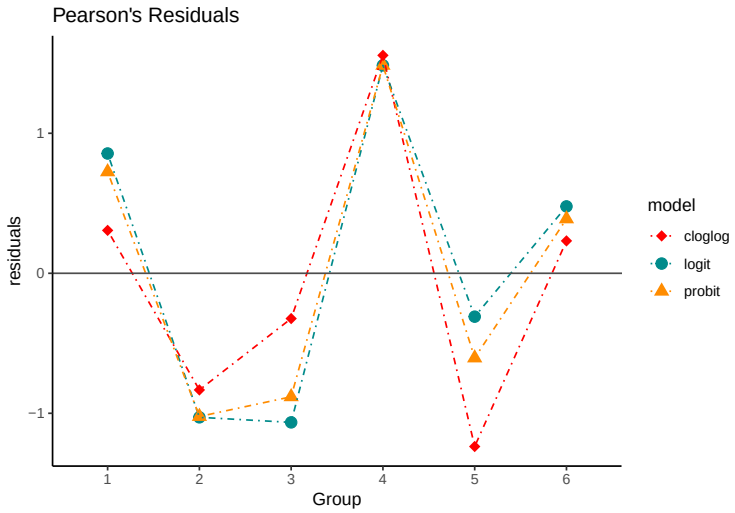

Figure 2: Peason's Residuals From Different Models

error between model prediction and the true value of ith observation, iterating i over the whole sample. Finally sum up the squared errors to get the SPSE. Lower SPSE indicates a better fit.

For all these indices, the complementary log-log yields the lowest value, it seems to have a best fit.

Overdispersion

Nuisance parameter (ϕ) can be estimated by $\frac{\text{Deviance}}{n-p}$ or $\frac{\text{Pearson's } \chi^2}{n-p}$

All models have deviance around 6, Pearson's Chi-Square around 5 and $n-p = 6-2 = 4$. In all the models $\hat{\phi} > 1$, the data is likely suffers from overdispersion. To deal with the problem, setting the argument `family ="quasibinomial"` in the `glm()`. to perform a quasi-likelihood using $\hat{\phi} = \frac{\text{Pearson's } \chi^2}{n-p}$. The results can be also seen in Table 1. Coefficients have no changes but larger s.e. (in parentheses) for the estimates.

Lastly, To compared the coefficients between models, we might want to refooting the variance.Thus for probit model we times the coefficients by $\frac{\pi}{\sqrt{3}}$and for complementary log-log model, times by $\frac{\pi}{\sqrt{6}}$. The result can be seen in Table 2.

Table 2: Refooting Coefficients

	Intercept	dose
Logit	-3.669	1.320
Probit	-3.918	1.407
C_log-log	-3.275	0.976

The effect of dose on positive response is lower in complementary log-log link function. And from the evaluation of model choice, we found the using complementary log-log has the best fitting, evaluation of the effect of dose using complementary log-log link function is more appropriate.

Q2

Furthermore, consider the link family:

$$g(\mu;\alpha) = \{\log \frac{(1-\mu)^{-\alpha}-1}{\alpha}; \alpha > 0\},$$

where α is a parameter to be determined. Develop a method of determining the optimal value of α. How would you choose the link function in practice for the data set, using the above link family?

Ans:

Observation:

Figure 3 shows different values of α in $g(\cdot)$ and some common link functions. We can see that $g(\mu;\alpha=1) = logit(\mu)$ and as $\alpha \to 0, g(\cdot)$ is close to complementary log-log function.

For the selection of α, one may perform a grid search. First choosing an interval over $[L_0, U_0]$ (L_0 should be ≥ 0), and compared the fit (using deviance, for example) over models with $\alpha = L_0, \alpha = \frac{U_0-L_0}{2} = M_0, \alpha = U_0$, (i.e., the boundary and the midpoint). For the model with the best fit, we can narrow down the searching interval to $[L_i, U_i]$ and searching for α with the best fit again. (the rule can be found in Appendix) The searching would be stopped if the fit of $\alpha = L_i, \alpha = U_i$ is similar (e.g., the difference of deviance $< \epsilon$) or the number of iteration over a maximum N. The searching should be successful if the fitting indices should be convex over α and the α with the best fit is included in $[L_0, U_0]$.

Here apply the algorithm to the data, we found that the $\alpha = 0.2469661$ has the best fit.

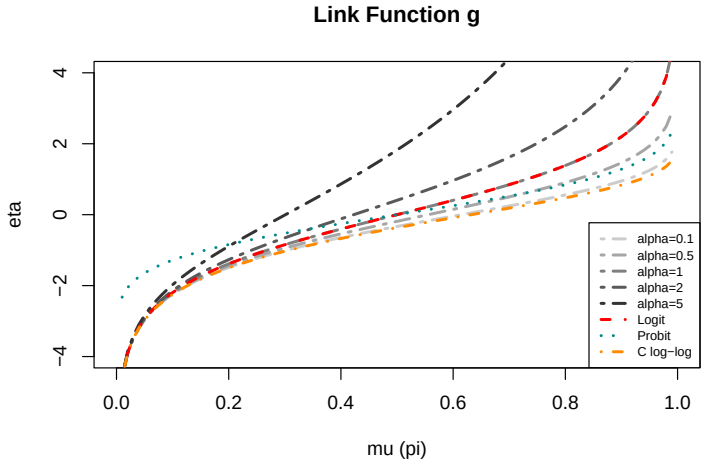

Figure 3: Plot of Link Functions

Comparison and Choice between link functions.

For Logit and Probit link, first, they are both symmetric to $\eta = 0$. If the grouped data have similar proportions with extreme values (close to 0 or 1), these two link function would be appropriate. The choice between 2 models can based on the interpretation of the data: is the odds-ratio interpretation for β in Logit or the latent response model in Probit be more appropriate to the data? Or just compare the goodness of fit between 2 models.

On the other hand, complementary log-log and the $g(\alpha \neq 1)$ function, are both asymmetric. For $\alpha < 1$, both function were similar, its inverse has approaching to 1 much faster as η increase. Thus, when the data has many groups with high proportion value, these two link function were more appropriate. Lastly, when $\alpha > 1$, function $g(\cdot)$ approaches to 1 in a slower speed, when the data has many groups with low proportion value, the link function were more appropriate.

Or one can naively use $g(\cdot)$ to fit the data if a algorithm for choosing α is available.

Appendix:

All the codes used in the report can be found in *Here*

The Algorithm of Updating Searching Intervals

In each iteration, we compared the fit over upper/lower-bound and the midpoint$(U_i, L_i, M_i(= \frac{U_i - L_i}{2}))$. On updating rules, consider 3 cases:

1. $\alpha = L_i$ has the minimum deviance: update $L_{i+1} = L_i - \Delta_i$, $U_{i+1} = M_i$
2. $\alpha = U_i$ has the minimum deviance: update $L_{i+1} = M_i$, $U_{i+1} = U_i + \Delta_i$
3. $\alpha = M_i$ has the minimum deviance: update $L_{i+1} = M_i + \frac{U_i - L_i}{4}$, $U_{i+1} = M_i - \frac{U_i - L_i}{4}$

Where $\Delta_i = \frac{M_i}{i \times 10}$ is quite an arbitrary form.