

## MATH 490 Assignment 8

---

Omar Sandoval

May 8, 2014

**5.4** Refer to Problem 4.16 on the four scales of the Myers-Briggs (MBTI) personality test. Table 5.10 shows the result of fitting a model using the four scales as predictors of whether a subject drinks alcohol frequently.

**a.** Conduct a model goodness-of-fit test, and interpret.

Through our R code, and Table 5.10, we can see that deviance is 11.149 with degrees of freedom equal to 11. We can conclude that there is no lack of fit and that the model is good enough.

**b.** If you were to simplify the model by removing a predictor, which would you remove? Why?

We can see that it is possible for us to take out the JP term. It is the least significant when compared to other terms.

The Likelihood-ratio test gives us 0.7964 with P-value = 0.37.

**c.** When six interaction terms are added, the deviance decreases to 3.74. Show how to test the hypothesis that none of the interaction terms are needed, and interpret.

We can calculate the likelihood-ratio statistic as  $11.149 - 3.74 = 7.409$  with degrees of freedom = 6 and P-value = 0.28. We can conclude that the simpler model which has no interaction is good.

**5.15** According to the *Independent* newspaper (London, March 8, 1994), the Metropolitan Police in London reported 30,475 people as missing in the year ending March 1993. For those of age 13 or less, 33 of 3271 missing males and 38 of 2846 missing females were still missing a year later. For ages 14-18, the values were 63 of 7256 males and 108 of 8877 females; for ages 19 and above, the values were 157 of 5065 and 159 of 3250 females. Analyze these data, including checking model fit and interpreting parameter estimates.

We can see that  $X^2 = 0.1$  with degrees of freedom 2. We can see from the data that, when given age, the estimated odds that a female is still missing are  $e^{.38} = 1.46$  times higher when compared to males. Thus, the estimated odds are much higher for the 19 and over bracket.

**5.19** Problem 7.9 shows a  $2 \times 2 \times 6$  table for  $Y$  = whether admitted to graduate school at the University of California, Berkeley.

**a.** Set up indicator variables and specify the logit model that has department as a predictor (with no gender effect) for  $Y$  = whether admitted (1 = yes, 0 = no).

The equation would look like;  $\text{logit}(\pi) = \alpha + \beta_1 d_1 + \cdots + \beta_6 d_6$ . Here,  $d_i = 1$  for department  $i$  and  $d_i = 0$  for any time otherwise.

**b.** For the model in (a), the deviance equals 21.7 with  $df = 6$ . What does this suggest about the quality of the model fit?

It suggests that the model fit is not good.

**c.** For the model in (a), the standardized residuals for the number of females who were admitted are (4.15, 0.50, -0.87, 0.55, -1.00, 0.62) for Departments (1,2,3,4,5,6). Interpret.

We can see that department 1 does not have a good fit.

**d.** Refer to (c). What would the standardized residual equal for the number of males who were admitted into Department 1? Interpret.

The standardized residual would equal  $-4.15$ . The residuals for males and females have the same absolute value but differ in sign. If we were to compare the this fit to a model where gender holds an effect, we can say that fewer males were admitted in the latter.

**e.** When we add a gender effect, the estimated conditional odds ratio between admissions and gender (1 = male, 0 = female) is 0.90. The marginal table, collapsed over department, has odds ratio 1.84. Explain how these associations differ so much for these data.

We can say that male applicants apply in larger numbers to departments that have higher proportions of acceptances.

**5.22** For the example in Section 5.3.1,  $y = 0$  at  $x = 10, 20, 30, 40$ , and  $y = 1$  at  $x = 60, 70, 80, 90$ .

- a. Explain intuitively why  $\hat{\beta} = \infty$  for the model  $\text{logit}(\pi) = \alpha + \beta x$
- b. Report  $\hat{\beta}$  and its  $SE$  for the software you use.

Using R,  $\hat{\beta} = 3.84$  and  $\hat{\alpha} = -192.2$ .

- c. Add two observations at  $x = 50$ ,  $y = 1$  for one and  $y = 0$  for the other. Report  $\hat{\beta}$  and its  $SE$ . Do you think these are correct? Why?

Here,  $\hat{\beta} = 2.65$  and  $\hat{\alpha} = -132.3$

- d. Replace the two observations in (c) by  $y = 1$  at  $x = 49.9$  and  $y = 0$  at  $x = 50.1$ . Report  $\hat{\beta}$  and its  $SE$ . Do you think these are correct? Why?

Here,  $\hat{\beta} = 0.53$  and  $\hat{\alpha} = -26.39$ . I'm not entirely sure if these are correct. Although, there is no more separation of  $x$ -values from  $y = 0$  and  $y = 1$ . Also, maximum likelihood is not infinite. So yes, I think these values are correct.

**5.30** The following are true-false questions.

- a. A model for a binary response has a continuous predictor. If the model truly holds, the deviance statistic for the model has an asymptotic chi-squared distribution as the sample size increases. It can be used to test model goodness of fit.

False.

- b. For the horseshoe crab data, when width or weight is the sole predictor for the probability of a satellite, the likelihood-ratio test of the predictor effect has  $P$ -value  $< 0.0001$ . When both weight and width are in the model, it is possible that the likelihood-ratio tests for the partial effects of width and weight could both have  $P$ -values larger than 0.05

True, this is indeed possible.

- c. For the model,  $\text{logit}[\pi(x)] = \alpha + \beta x$ , suppose  $y = 1$  for all  $x \leq 50$  and  $y = 0$  for all  $x > 50$ . Then, the ML estimate  $\hat{\beta} = -\infty$

True. We see this up above in problem 5.22.