P8131_hw_2

Renjie Wei rw2844

2/18/2022

Problem 1

Fit the model $g(\text{P(dying)}) = \alpha + \beta X$, with logit, probit, and complementary log-log links.

(a) Fill out the table and give comments.

Model	Estimate of β	95% CI of β	Deviance	p(dying|x=0.01)
Logit	1.162	0.806,1.517	0.379	0.090
Probit	0.686	0.497,0.876	0.314	0.085
c-log-log	0.747	0.532,0.961	2.230	0.128

Comments:

- All the estimated βs are greater than 0, which means the increase in dose may increase the probability of dying.
- 0 is not included in all the 3 95% CIs, that means we have 95% confidence to conclude that the dose level is significantly associated with the probability of dying.
- All the deviances follow the $\chi^2(3)$, the deviance of the probit link model is the smallest, which means it is the best fitted model among all the three models.
- About the probability of dying conditioned on 0.01 dose, the logit link model gives us a 0.0901,the probit link model gives a 0.0853 and the c-log-log link model gives a 0.1282.

(b) Suppose that the dose level is in natural logarithm scale, estimate LD50 with 90% confidence interval based on the three models.

Since we got the following link functions:

Logit link function:

$$g_1(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x$$

Probit link function:

$$g_2(\pi) = \Phi^{-1}(\pi) = \beta_0 + \beta_1 x$$

C-log-log link function:

$$g_3(\pi) = \log(-\log(1-\pi)) = \beta_0 + \beta_1 x$$

So we can get the point estimates of $LD50$ by solving the following equation:

$$g(0.5) = \beta_0 + \beta_1 x$$

Logit estimate:

$$\hat{x} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = f(\hat{\beta})$$

Probit estimate:

$$\hat{x} = \frac{\Phi^{-1}(0.5) - \hat{\beta}_0}{\hat{\beta}_1} = f(\hat{\beta})$$

C-log-log estimate:

$$\hat{x} = \frac{\log(-\log(0.5)) - \hat{\beta}_0}{\hat{\beta}_1} = f(\hat{\beta})$$

And we can get the asymptotic variance of $\hat{x}$:

$$\text{var}(\hat{x}) = \text{var}(f(\hat{\beta})) = (\frac{\partial f(\hat{\beta})}{\partial \beta_0})^2 \text{var}(\hat{\beta}_0) + (\frac{\partial f(\hat{\beta})}{\partial \beta_1})^2 \text{var}(\hat{\beta}_1) + 2(\frac{\partial f(\hat{\beta})}{\partial \beta_0})(\frac{\partial f(\hat{\beta})}{\partial \beta_1}) \text{cov}(\hat{\beta}_0, \hat{\beta}_1)$$

Then the asymptotic CI of $LD50$ is:

$$LD50 \in [e^{\hat{x} - z_{\alpha/2}\sqrt{\text{var}(\hat{x})}}, e^{\hat{x} + z_{\alpha/2}\sqrt{\text{var}(\hat{x})}}]$$

I write a function to calculate the $\text{var}(\hat{x})$, which is `se_of_h_beta`:

```
# a function to calculate se for any h(beta)
se_of_h_beta <- function(fit_object, h_expr){
  beta_0 = fit_object$coefficients[["(Intercept)"]]
  beta_1 = fit_object$coefficients[[2]]
  h_beta = eval(h_expr)

  I_beta = vcov(fit_object)
  #inv_I_beta = solve(I_beta) already inversed!
  partial_d_beta0 = eval(D(h_expr,"beta_0"))
  partial_d_beta1 = eval(D(h_expr,"beta_1"))

  partial_d_mtx = matrix(c(partial_d_beta0,partial_d_beta1),2,1)
  se = sqrt(t(partial_d_mtx) %*% I_beta %*% partial_d_mtx)
  return(c(h_beta,se))
}

LD50_logit = expression(-beta_0/beta_1)
e_logit = se_of_h_beta(logit.fit, LD50_logit)


LD50_probit = expression(-beta_0/beta_1)
e_probit = se_of_h_beta(probit.fit, LD50_probit)

g_pi_cloglog = log(-log(0.5))
LD50_cloglog = expression((g_pi_cloglog-beta_0)/beta_1)
e_cloglog = se_of_h_beta(cloglog.fit, LD50_cloglog)

summary_tab_1_b <-
  data.frame(
    Model = c("Logit", "Probit", "c-log-log"),
```

```
  LD50 = c(exp(e_logit[1]), exp(e_probit[1]), exp(e_cloglog[1])),
  CI_L = exp(c(e_logit[1], e_probit[1], e_cloglog[1]) - qnorm(.95)*c(e_logit[2], e_probit[2], e_clogl
  CI_U = exp(c(e_logit[1], e_probit[1], e_cloglog[1]) + qnorm(.95)*c(e_logit[2], e_probit[2], e_clogl
)
```

The following table shows the results of estimated LD50.

Model	Estimate of LD50	90% CI lower	90% CI upper
Logit	7.389	5.510	9.910
Probit	7.436	5.583	9.904
c-log-log	8.841	6.526	11.977

Problem 2

Please analyze the data using a logistic regression and answer the following questions.

(a) How does the model fit the data?

Firstly, I made a dataframe for the data.

We assume the response of those who received offers Y_i in each group follow the same Bernoulli distribution $Y_i \sim Bin(1, \pi)$, then in each group j with size m_j, $Y = \sum_{i=1}^{m_j} Y_i$ has a Binomial distribution, that is $Y \sim Bin(m_j, \pi)$.

So we fit a logistic regression based on the above assumptions.

```
mph.fit <-
  glm(cbind(enrolls, offers-enrolls) ~ amount, family = binomial(link = "logit"), data = mph_enroll)
```

And I do Hosmer–Lemeshow test for the goodness-of-fit since the data is sparse.

```
library(ResourceSelection)
hoslem.test(mph.fit$y, fitted(mph.fit), g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  mph.fit$y, fitted(mph.fit)
## X-squared = 1.6111, df = 8, p-value = 0.9907
```

The result shows that the Hosmer-Lemeshow stastic χ^2_{HL} is `X-squared = 1.6111` with `df = 8` and `p-value = 0.9907`. Since `p-value` is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no evidence of that the model is lack of fit.

(b) How do you interpret the relationship between the scholarship amount and enrollment rate? What is 95% CI?

```
##
## Call:
## glm(formula = cbind(enrolls, offers - enrolls) ~ amount, family = binomial(link = "logit"),
##     data = mph_enroll)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4735  -0.6731   0.1583   0.5285   1.1275
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.64764    0.42144  -3.910 9.25e-05 ***
## amount       0.03095    0.00968   3.197  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21.617  on 16  degrees of freedom
## Residual deviance: 10.613  on 15  degrees of freedom
## AIC: 51.078
##
## Number of Fisher Scoring iterations: 4
```

The estimate coefficient β_1 is 0.031, which is equal to the $\log(OR_{\text{amout}})$. So the OR_{amout} is 1.031, which means the odds ratio of enrollment increases 3.14% per 1000 dollars increase in the scholarship. The 95% CI of OR is (`round(OR_CI_L,3)`,`round(OR_CI_U,3)`)

(c) How much scholarship should we provide to get 40% yield rate (the percentage of admitted students who enroll?) What is the 95% CI?

Since we use logit link function, we can get the estimate of scholarship (to get 40% yield rate) by solving the following equation:

$$\log(\frac{\pi}{1-\pi}) = \log(0.4/0.6) = \hat{\beta}_0 + \hat{\beta}_1 \times \widehat{\text{scholarship}}$$

Then we can get the estimate scholarship and the asymptotic variance and CI with the same method in problem 1.

We get the estimate of scholarship $\widehat{\text{scholarship}} = 40.134$, which means we should provide 40.134 thousand dollars of scholarship to get 40% yield rate. The 95% CI is (30.583,49.686)