Practicing the concepts of Logit, Probit, c-log-log Models, Grouped and Sparse Data, Goodness of Fit

Hana Akbarnejad

2/18/2020

If needed, please find the Rmd file attached

problem 1

In this problem we analyze the data collected from a bioassay study, where X is the dosage and Y is the number of animals that have died after taking the drug.

Because we have a binary ooutcome (dying versus not dying), we should model Y-X relationship using logit, probit, cloglog link functions. In this case, the probability of the outcome (dying) is denoted as p and n is the number of animals tested which is 30.

```
# Problem 1-i

dose = c(0:4)
dying = c(2, 8, 15, 23, 27)
data1 = cbind(dose, dying)
data1 = as.data.frame(data1)
data1 %>% knitr::kable()

fit1 = glm(cbind(dying,30-dying)~dose, family=binomial(link='logit'))
fit2 = glm(cbind(dying,30-dying)~dose, family=binomial(link='probit'))
fit3 = glm(cbind(dying,30-dying)~dose, family=binomial(link='cloglog'))

### CI's
# CI for logit
vcov(fit1) # covariance of beta MLE (fisher information inverse)
beta=fit1$coefficients[2]
se=sqrt(vcov(fit1)[2,2]) # (same as in summary(glm_logit))
beta+c(qnorm(0.025),0,-qnorm(0.025))*se


# CI for probit
vcov(fit2) # covariance of beta MLE (fisher information inverse)
beta=fit2$coefficients[2]
se=sqrt(vcov(fit2)[2,2]) # (same as in summary(glm_logit))
beta+c(qnorm(0.025),0,-qnorm(0.025))*se

# CI for cloglog
vcov(fit3) # covariance of beta MLE (fisher information inverse)
beta=fit3$coefficients[2]
se=sqrt(vcov(fit3)[2,2]) # (same as in summary(glm_logit))
beta+c(qnorm(0.025),0,-qnorm(0.025))*se

### Deviance
summary(fit1)$deviance
```

```r
summary(fit2)$deviance
summary(fit3)$deviance

### P-value
pchisq(fit1$deviance, df=3, lower.tail=FALSE)
pchisq(fit2$deviance, df=3, lower.tail=FALSE)
pchisq(fit3$deviance, df=3, lower.tail=FALSE)

### Predictions
predicted1 = predict(fit1, data.frame(dose = 0.01), type = 'response')
predicted2 = predict(fit2, data.frame(dose = 0.01), type = 'response')
predicted3 = predict(fit3, data.frame(dose = 0.01), type = 'response')
```

dose	dying
0	2
1	8
2	15
3	23
4	27

Problem 1-i

In this part, we will model $g(P(dying)) = \alpha + \beta X$ with logit, probit and complementary log-log linkk and we are interested in calculating Estimate of β, 95% CI for β, Deviance, and $\hat{p(dying|x = 0.01)}$.

The table belows these estimates:

```r
my_table = matrix(c("1.162", "0.686", "0.747", "(0.806,1.517)", "(0.497,0.876)", "(0.532,0.961)",
                    "0.379", "0.314", "2.230", "0.09", "0.0853", "0.128"), ncol=4, byrow = FALSE)
colnames(my_table) = c("Estimate", "95% CI for beta", "Deviance", "p_hat(dying|x=0.01)")
rownames(my_table) = c("logit", "probit", "c-loglog")
my_table = as.table(my_table)
```

	Estimate	95% CI for beta	Deviance	p_hat(dying|x=0.01)
logit	1.162	(0.806,1.517)	0.379	0.09
probit	0.686	(0.497,0.876)	0.314	0.0853
c-loglog	0.747	(0.532,0.961)	2.230	0.128

- Comments of β Estimate, 95% CI, and $\hat{p(dying|x = 0.01)}$ for three models:

We can observe that logit model has the highest estimate for β and the probit model has the lowest. All estimates are positive which means that the risk of death increases with the increase in dose taken. The probability of death given the dose of 0.01 is similar in all three models, however the c-loglog model has the highest estimate.

for logit link function model, the estimate is 1.162 which is the log odds ratio of death in animals for each one unit increase in dose. The 95% CI for this model is (0.806,1.517) which means that we are 95% confident that the true value of beta falls somewhere between 0.806 and 1.517. Also, using this model we can say that the estimated probability of death is 0.09 when the animals take dose of 0.01.

For probit link function model, the estimate is 0.686 which is the difference in Z score associated with each one-unit increase in dose.The 95% CI for this model is (0.497,0.876) which means that we are 95% confident that the true value of beta falls somewhere between 0.497 and 0.876. Also, using this model we can say that the estimated probability of death is 0.0853 when the animals take dose of 0.01.

For c-loglog link function model, the estimate is 0.747 which is the contribution to the complementary-log-log of mortality for a unit change in dose. Equivalently, we can say: the log-survival will drops by about (survival function: $1-\pi$) 2.110 (exp(0.747)) per one unit increase in dose. Or we can also say that 0.747 is the log hazard ratio of death per one unit increase in dose.

The 95% CI for this model is (0.532,0.961) which means that we are 95% confident that the true value of beta falls somewhere between 0.532 and 0.961.Also, using this model we can say that the estimated probability of death is 0.128 when the animals take dose of 0.01.

- Comments of Goodness of Fit of three models:

Since we have 30 animals tested in each group ($m_i \geq 30$), we have our data grouped, so we can use deviance as our goodness of fit. The deviance approximately follows χ^2_{n-p} where n = 5 (the number of groups) and p = 2 (the number of parameters). Which here is $D \sim \chi^2_3$.

For goodness of fit, the null hypothesis is that our model is correctly specified. Here we have a p-value of 0.944 for logit, a p-value of 0.957 for ptobit, and a p-value of 0.526 for c-lolog. Since all values are geater than 0.05, we fail to reject null hypothesis which suggests that all three models fit our data well.However, the amount of deviance is close for logit and probit models compared to c-loglog model. The amount of deviance is smaller using logit and probit link function models which suggests that the goodness of fit is better in these two models comapred to c-loglog.

Problem 1-ii

In this part we want to estimate LD50 (the dose at which 50% of animals die) with 90% confidence interval based on models built based on logit, probit, c-loglog link functions (considering that X is in natural logarithm scale).

```
## Problem 1-ii
# logit LD50 and CI
beta0=fit1$coefficients[1]
beta1=fit1$coefficients[2]
betacov=vcov(fit1) # inverse fisher information
x0fit=-beta0/beta1
exp(x0fit) # point estimate of LD50  ## this is our xLD50?
varx0=betacov[1,1]/(beta1^2)+betacov[2,2]*(beta0^2)/(beta1^4)-2*betacov[1,2]*beta0/(beta1^3)
c(x0fit,sqrt(varx0)) # point est and se
exp(x0fit+c(qnorm(0.05),-qnorm(0.05))*sqrt(varx0)) # 95% CI for LD50

# probit LD50 and CI
beta0_1=fit2$coefficients[1]
beta1_1=fit2$coefficients[2]
betacov_1=vcov(fit2) # inverse fisher information
x0fit_1=-beta0_1/beta1_1
exp(x0fit_1) # point estimate of LD50  ## this is our xLD50?
varx0_1=betacov_1[1,1]/(beta1_1^2)+betacov_1[2,2]*(beta0_1^2)/(beta1_1^4)-2*betacov_1[1,2]*beta0_1/(beta
c(x0fit_1,sqrt(varx0_1)) # point est and se
exp(x0fit_1+c(qnorm(0.05),-qnorm(0.05))*sqrt(varx0_1))
```

```r
# cloglog LD50 and CI
beta0_2=fit3$coefficients[1]
beta1_2=fit3$coefficients[2]
betacov_2=vcov(fit3) # inverse fisher information
x0fit_2= (log(log(2))- beta0_2)/beta1_2
exp(x0fit_2) # point estimate of LD50  ## this is our xLD50?
varx0_2=betacov_2[1,1]/(beta1_2^2)+betacov_2[2,2]*(beta0_2^2)/(beta1_2^4)-2*betacov_2[1,2]*beta0_2/(beta
c(x0fit_2,sqrt(varx0_2)) # point est and se
exp(x0fit_2+c(qnorm(0.05),-qnorm(0.05))*sqrt(varx0_2))
```

For logit, we know that the link function is:

$$g(\pi) = log(\pi/1-\pi) = \beta_0 + \beta_1 X$$

So, because $\pi = 0.5$ when calculating LD_{50}, we will have:

$$log(0.5/1-0.5) = log1 = 0$$

Which means that $\beta_0+\beta_1 X = 0$ and $x_0 = -\beta_0/\beta_1$. Substituting the coefficients we will have $exp(x_0) = 7.389$, and then we can calculate the CI.

For probit, we know that the link function is:

$$g(\pi) = \phi^{-1}(\pi) = \beta_0 + \beta_1 X$$

So, because $\pi = 0.5$ when calculating LD_{50}, we will have:

$$\phi^{-1}(0.5) = \beta_0 + \beta_1 X$$

Which means that $\beta_0 + \beta_1 X = 0$ and $x_0 = -\beta_0/\beta_1$. Substituting the coefficients we will have $exp(x_0) = 7.43583$, and then we can calculate the CI.

For logit, we know that the link function is:

$$g(\pi) = log(-log(1-\pi)) = \beta_0 + \beta_1 X$$

So, because $\pi = 0.5$ when calculating LD_{50}, we will have:

$$log(log(2)) = -0.366 = \beta_0 + \beta_1 X$$

Substituting the coefficients we will have $exp(x_0) = 8.84$, and then we can calculate the CI.

The table below shows the calculated amounts:

```r
my_table2 = matrix(c("7.389", "(5.509,9.909)", "7.43583", "(5.582,9.904)",
                     "8.841249", "(6.636,11.779)"), ncol=2, byrow = TRUE)
colnames(my_table2) = c("LD50 Estimate Exponentiated", "90% CI for LD50")
rownames(my_table2) = c("logit", "probit", "c-loglog")
my_table2 = as.table(my_table2)
my_table2 %>%  knitr::kable()
```

	LD50 Estimate Exponentiated	90% CI for LD50
logit	7.389	(5.509,9.909)
probit	7.43583	(5.582,9.904)
c-loglog	8.841249	(6.636,11.779)

We can see that the logit and probit estimates and 90% confidence intervals are similar and that the estimate of the 50% lethal dose for c-loglog model is higher than logit and probit, and the confidence interval is wider than those two methods.

Problem 2

Problem 2-i

In this question, we have analyzed the relationship between the amount of one-time two-year scholarship, on the number of people who got enrolled in the program out of all who got the offer.

The data presented here is grouped data, but it is sparse data ($m_i < 30$). So, for assessing the goodness of fit we cannot use Deviance or generalized Pearson χ^2 because the asympthotic distribution doesn't hold anymore. Instead, we use Hosmer-Lemeshow statistic: X^2_{HL}.

```
# importing data
data = tibble(
  amount = seq(from = 10000, to = 90000, by = 5000),
  offered = c(4,6,10,12,39,36,22,14,10,12,8,9,3,1,5,2,1),
  enrolled = c(0,2,4,2,12,14,10,7,5,5,3,5,2,0,4,2,1)
)

# defining our x, y, m
x = data$amount
y = data$enrolled
m = data$offered

fit_glm = glm(cbind(y,m-y)~x, family=binomial(link='logit'))

library(ResourceSelection)
hl = hoslem.test(fit_glm$y, fitted(fit_glm), g=10) # fail to reject H0 >>> good fit!
```

Using this method, we have the X-squared = 1.6111 with degrees of freedom = 8 (χ^2_{g-2} where g = 10) and p-value = 0.9907. This result indicates that we fail to reject H_0 which means that the model fits the data well.

Problem 2-ii

looking at the relationship between the scholarship amount and the enrollment, we can see that the log odds ratio of enrollment changes by 3.095e-05 per 1 dollar increase in scholarship. Because we probably don't care about every dollar increase in scholarship, I change the scale and say that The log odds ratio of enrollment is 0.3095 per 10,000 dollars increase in scholarship amount. This intrepretation is equivalent to exponentiated interpretation, saying that: The odds ratio of enrollment is 1.363 per 10,000 dollars increase in scholarship amount.

```
# 95% CI
beta_glm = fit_glm$coefficients[2]
se_glm = sqrt(vcov(fit_glm)[2,2])
beta_glm + c(qnorm(0.025),0,-qnorm(0.025))*se_glm
```

The 95% CI for $\hat{\beta}_1$ is (1.197845e-05, 4.992240e-05) for unit 1 dollar and equivalently (0.12, 0.5) for the scale of $10,000 increase in the amount of scholarship.

Problem 2-iii

To determine the amount of scholarship we should provide to get 40% yield rate and the 95% CI, we should use $\pi = 0.4$. Using logistic regression, we will have:

$$g(\pi) = log(\pi/1 - \pi) = \beta_0 + \beta_1 X$$

So, we see that:

$$g(\pi) = log(0.4/1 - 0.4) = \beta_0 + \beta_1 X$$

So,

$$X = (log(0.4/0.6) - \beta 0)/\beta 1$$

```
beta_0 = fit_glm$coefficients[1]
beta_1 = fit_glm$coefficients[2]
my_log = log (.4/.6) # -0.4054651
x0_estimate = (my_log - beta_0)/beta_1    # 40134.29$ scholarship should be offered to have 40% of peop

betacov_3 = vcov(fit_glm)  # covariance of beta_0 and beta_1 to gain the variance of the estimate
var_x0_estimate = betacov_3[1,1]/(beta_1^2) + betacov_3[2,2]*((beta_0 - log(0.4/0.6))^2)/(beta_1^4)-2*b

x0_estimate+c(qnorm(0.025),-qnorm(0.025))*sqrt(var_x0_estimate)
```

We can see that $\hat{x_0} = 40134.29$ which means that we need 4.013429 ten-thousand dollars scholarship to get 40% yield rate and the 95% CI for for $\hat{x_0}$ is (30583.04, 49685.53) is dollars scale or equivalently (3.06, 4.97) in ten-thoundas dollars scale.