P8131 Spring 2022 Homework #2 Solution

1. The table below gives the data collected from a bioassay study in which X variable (treated as continuous variable) is the concentration level. At each of five different dose levels (0-4), 30 animals are tested and the number of dying are recorded.

Dose (X)	0	1	2	3	4
Number of dying	2	8	15	23	27

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

(a) Fill out the table and give comments.

Model	Estimate of β	CI for β	Deviance	$\hat{p}(\text{dying}\|x = 0.01)$
logit	1.162	(0.806, 1.517)	0.379	0.0901
probit	0.686	(0.497, 0.876)	0.314	0.0853
c-log-log	0.747	(0.532, 0.961)	2.23	0.1282

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

Solution:

(a) From the output in the table, we see that the model with probit link seems to fit the data best since it has the smallest residual deviance. In fact, the GOF test for this model using deviance has a p-value $0.957 > 0.05$. Since CI for β in all three models are positive and do not include 0, we know that higher dose level is significantly associated with with higher probability of dying. At dose level 0.01, we can estimate the probability of dying to be 0.0853 using probit link model.

(b) Logit link:

$$-\frac{\hat{\beta}_0}{\hat{\beta}_1} = 2$$

Estimated LD50 in natural logarithm scale is 2, with standard error 0.178 and its 90% confidence interval is $(1.706, 2.294)$. Estimated LD50 in original scale is $\exp(2) = 7.389$, with 90% confidence interval $(5.510, 9.910)$.

Probit link:

$$\Phi^{-1}(0.5) = 0$$

$$-\frac{\hat{\beta}_0}{\hat{\beta}_1} = 2.006$$

Estimated LD50 in natural logarithm scale is 2.006, with standard error 0.174 and its 90% confidence interval is $(1.720, 2.293)$. Estimated LD50 in original scale is $\exp(2.006) = 7.436$, with 90% confidence interval $(5.583, 9.904)$.

C-log-log link:

$$\log(-\log(0.5)) = -0.367 = \hat{\beta}_0 + \hat{\beta}_1 \text{Amount}$$

$$-\frac{\hat{\beta}_0 + 0.367}{\hat{\beta}_1} = 2.179$$

Estimated LD50 in natural logarithm scale is 2.179, with standard error 0.185 and its 90% confidence interval is $(1.876, 2.483)$. Estimated LD50 in original scale is $\exp(2.179) = 8.841$, with 90% confidence interval $(6.526, 11.977)$.

2. The table below contains the enrollment data of some MPH program in a year

   - Amount: one-time two-year scholarship
   - Offer: the number of offers made with the corresponding scholarship
   - Enrolls: the number of offer accepted

Amount (in thousand dollars)	**Offers**	**Enrolls**
10	4	0
15	6	2
20	10	4
25	12	2
30	39	12
35	36	14
40	22	10
45	14	7
50	10	5
55	12	5
60	8	3
65	9	5
70	3	2
75	1	0
80	5	4
85	2	2
90	1	1

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

(a) How does the model fit the data?

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

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

Solution:

(a) We can fit a logistic regression model treating "Offer" as group sizes and "Enrolls" as number of outcomes in each group. Continuous variable "Amount" is the only predictor in the model. Since the group size is not very large (sparse data), we can use Hosmer-Lemeshow statistic to test model goodness of fit. We get a test statistic 1.611 with degrees of freedom $10-2=8$. The corresponding p-value from a chi-squared distribution is $0.991 > 0.05$. Therefore, we do not have enough evidence to reject our proposed model.

Note: Do not take points if they have deviance statistic 10.61, with df=15, p-value 0.78, or if they have generalized Pearson's χ^2 statistic 8.81. With same df, the p-value from GOF test is $0.89 > 0.05$. Make comments about using Hosmer-Lemeshow GOF.

(b) $\exp(0.031) = 1.031$. For every \$1,000 increase in scholarship, we expect to see 3.1% increase in odds of enrolling in the program. The 95% CI: $\exp(\hat{\beta} \pm 1.96\hat{\text{se}}(\hat{\beta})) = (1.012, 1.051)$.

(c) $\log \frac{0.4}{1-0.4} = -1.648 + 0.031\text{Amount} = \log \frac{2}{3}$. Hence we can solve the equation and get that we need provide \$40.13K scholarship for 40% yield rate. The 95% CI for this quantity is (30.56, 49.69).

Appendix

```
# Problem 1
# i)
x = c(0:4)
death = c(2,8,15,23,27)
data1 = data.frame(death, x)

# logit link
glm1_logit <- glm(cbind(death, rep(30,5)-death)~x, data=data1,
family=binomial(link='logit'))
summary(glm1_logit)
round(glm1_logit$coefficients[2],3)
round(confint.default(glm1_logit),3) # CI
round(sum(residuals(glm1_logit,type='deviance')^2),3) # deviance
round(predict.glm(glm1_logit,newdata=data.frame(x=0.01),type='r'),4)
```

```
# probit link
qnorm(0.5, 0, 1)
glm1_probit <- glm(cbind(death, rep(30,5)-death)~x, data=data1,
family=binomial(link='probit'))
summary(glm1_probit)
round(glm1_probit$coefficients[2],3)
round(confint.default(glm1_probit),3) # CI
round(sum(residuals(glm1_probit,type='deviance')^2),3) # deviance
dev=sum(residuals(glm1_probit,type='deviance')^2)
round(predict.glm(glm1_probit,newdata=data.frame(x=0.01),type='r'),4)
1-pchisq(dev,5-2)

# c-log-log link
glm1_cll <- glm(cbind(death, rep(30,5)-death)~x, data=data1,
family=binomial(link='cloglog'))
summary(glm1_cll)
round(glm1_cll$coefficients[2],3)
round(confint.default(glm1_cll),3) # CI
round(sum(residuals(glm1_cll,type='deviance')^2),3) # deviance
round(predict.glm(glm1_cll,newdata=data.frame(x=0.01),type='r'),4)

# LD50
# logit
beta0=glm1_logit$coefficients[1]
beta1=glm1_logit$coefficients[2]
betacov=vcov(glm1_probit)
x0fit=-beta0/beta1
round(exp(x0fit),3)
varx0=betacov[1,1]/(beta1^2)+betacov[2,2]*(beta0^2)/(beta1^4)
-2*betacov[1,2]*beta0/(beta1^3)
c(x0fit,sqrt(varx0))
# CI
round((x0fit+c(qnorm(0.05),0,-qnorm(0.05))*sqrt(varx0)),3)
round(exp((x0fit+c(qnorm(0.05),0,-qnorm(0.05))*sqrt(varx0))),3)
# 90% CI for LD50

# probit
beta0=glm1_probit$coefficients[1]
beta1=glm1_probit$coefficients[2]
betacov=vcov(glm1_probit)
x0fit=-beta0/beta1
round(x0fit,3)
round(exp(x0fit),3)
```

```
varx0=betacov[1,1]/(beta1^2)+betacov[2,2]*(beta0^2)/(beta1^4)
-2*betacov[1,2]*beta0/(beta1^3)
c(x0fit,sqrt(varx0))
# CI
round((x0fit+c(qnorm(0.05),0,-qnorm(0.05))*sqrt(varx0)),3)
round(exp((x0fit+c(qnorm(0.05),0,-qnorm(0.05))*sqrt(varx0))),3)
# 90% CI for LD50

# cloglog
c = log(-log(0.5))
beta0=glm1_cll$coefficients[1]
beta1=glm1_cll$coefficients[2]
betacov=vcov(glm1_cll)
x0fit=(c-beta0)/beta1
round(x0fit,3)
round(exp(x0fit),3)
varx0=betacov[1,1]/(beta1^2)+betacov[2,2]*((c-beta0)^2)/(beta1^4)
+2*betacov[1,2]*(c-beta0)/(beta1^3)
c(x0fit,sqrt(varx0))
# CI
round((x0fit+c(qnorm(0.05),0,-qnorm(0.05))*sqrt(varx0)),3)
round(exp((x0fit+c(qnorm(0.05),0,-qnorm(0.05))*sqrt(varx0))),3)
# 90% CI for LD50

# Problem 2
amount=c(seq(10000,90000, by=5000))/1000
offer=c(4,6,10,12,39,36,22,14,10,12,8,9,3,1,5,2,1)
enroll=c(0,2,4,2,12,14,10,7,5,5,3,5,2,0,4,2,1)
data=data.frame(amount,offer,enroll)
data
plot(data$amount,data$enroll/data$offer)


glm_logit=glm(cbind(enroll,offer-enroll)~amount, data=data,
family=binomial(link='logit'))
summary(glm_logit) # wald test of coefficients

# Q1: model fitting
sum(residuals(glm_logit,type='pearson')^2) # pearson chisq
dev=sum(residuals(glm_logit,type='deviance')^2);dev # deviance
pval=1-pchisq(dev,17-2);pval # fit is ok, fails to reject
#
# check Hosmer-Lemeshow for sparse data
library(ResourceSelection)
```

```
hl <- hoslem.test(glm_logit$y, fitted(glm_logit), g=10)
# fitted: returns \hat{pi}
hl # again, fit is ok, fails to reject

# Q2: model interpretation
summary(glm_logit)
beta=glm_logit$coefficients[2]
exp(beta) # odds ratio per 1k increase in scholarship
se=sqrt(vcov(glm_logit)[2,2])
# CI for odds ratio
exp(beta+c(qnorm(0.025),0,-qnorm(0.025))*se)

exp(confint.default(glm_logit))

# Q3: 40% yield rate
predict(glm_logit, type='r')
predict(glm_logit)
(log(2/3)+1.648)/0.031

# CI
c = log(2/3)
beta0=glm_logit$coefficients[1]
beta1=glm_logit$coefficients[2]
betacov=vcov(glm_logit)
x0fit=(c-beta0)/beta1
round(x0fit,3)
varx0=betacov[1,1]/(beta1^2)+betacov[2,2]*((c-beta0)^2)/(beta1^4)
+2*betacov[1,2]*(c-beta0)/(beta1^3)
c(x0fit,sqrt(varx0))
# CI
round((x0fit+c(qnorm(0.025),0,-qnorm(0.025))*sqrt(varx0)),3)
```