P8131 HW2

Shihui Zhu, sz3029

Problem 1

a) Fill out the table and give comments

Fit the model with logit, probit, and complementary log-log links

```
## # A tibble: 5 x 4
##   dose_x num_of_dying total_num  live
##    <int>        <dbl>     <dbl> <dbl>
## 1      0            2        30    28
## 2      1            8        30    22
## 3      2           15        30    15
## 4      3           23        30     7
## 5      4           27        30     3
```

Fit $g(P(dying)) = \alpha + \beta X$ using logit:

```
# Fit GLM
fit.logit = glm(cbind(num_of_dying, live) ~ dose_x, family = binomial(link = 'logit'),
          data = data1)
fit.logit$coefficients
```

```
## (Intercept)      dose_x
##   -2.323790    1.161895
```

So the fitted logit model is

$$\hat{\pi}(x) = \frac{e^{-2.323790+1.161895x}}{1 - e^{-2.323790+1.161895x}}$$

β is 1.161895.

Fit $g(P(dying)) = \alpha + \beta X$ using probit:

```
fit.probit = glm(cbind(num_of_dying, live) ~ dose_x, family = binomial(link = 'probit'),
          data = data1)
fit.probit$coefficients
```

```
## (Intercept)      dose_x
##  -1.3770923   0.6863805
```

So the fitted probit model is:

$$\hat{\pi}(x) = \phi(-1.3770923 + 0.6863805x)$$

β is 0.6863805.

Fit $g(P(dying)) = \alpha + \beta X$ using complementary log-log:

```
fit.cloglog = glm(cbind(num_of_dying, live) ~ dose_x, family = binomial(link = 'cloglog'),
          data = data1)
fit.cloglog$coefficients
```

```
## (Intercept)      dose_x
##  -1.9941520   0.7468193
```

So the fitted complementary log-log model is:

$$\hat{\pi}(x) = 1 - e^{-e^{-1.9941520+0.7468193x}}$$

β is 0.7468193.

Table:

```
# 95% CI for beta
invfisher.logit <- vcov(fit.logit) # inverse of fisher information matrix
invfisher.probit <- vcov(fit.probit)
invfisher.cloglog <- vcov(fit.cloglog)

# Compute CIs
CI.logit = fit.logit$coefficients + kronecker(t(c(0,qnorm(0.025),-qnorm(0.025))),
                                              t(t(sqrt(diag(invfisher.logit)))))
CI.probit = fit.probit$coefficients + kronecker(t(c(0,qnorm(0.025),-qnorm(0.025))),
                                              t(t(sqrt(diag(invfisher.probit)))))
CI.cloglog = fit.cloglog$coefficients + kronecker(t(c(0,qnorm(0.025),-qnorm(0.025))),
                                              t(t(sqrt(diag(invfisher.cloglog)))))
# Bind columns
out.logit = cbind(CI.logit[-1,,drop = FALSE],
                  sum(residuals(fit.logit,type='deviance')^2), # or fit.logit$deviance
                  predict(fit.logit, newdata = tibble(dose_x = 0.01), type = "response"))
out.probit = cbind(CI.probit[-1,,drop = FALSE],
                   sum(residuals(fit.probit,type = 'deviance')^2),
                   predict(fit.probit, newdata = tibble(dose_x = 0.01), type = "response"))
out.cloglog = cbind(CI.cloglog[-1,,drop = FALSE],
                    sum(residuals(fit.cloglog,type='deviance')^2),
                    predict(fit.cloglog, newdata = tibble(dose_x = 0.01), type = "response"))

# Bind rows
out <- rbind(out.logit, out.probit, out.cloglog)
colnames(out)=c('Estimate of Beta','95% CI lower','95% CI upper', "Deviance", "P(dying|x = 0.01)")
rownames(out)=c('logit', 'probit', 'complementary log-log')

out %>% knitr::kable(digits = 3)
```

	Estimate of Beta	95% CI lower	95% CI upper	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
complementary log-log	0.747	0.532	0.961	2.230	0.128

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

LD50 Estimate for logit model:

```
# Logit
x_0 <- -coef(summary(fit.logit))[1]/coef(summary(fit.logit))[2]
# Point Estimate
# MASS::dose.p(fit.logit, p = 0.5)
```

```r
x_0.se <- sqrt(
  t(c(-1/coef(summary(fit.logit))[2],
      coef(summary(fit.logit))[1]/coef(summary(fit.logit))[2]^2)) %*%
    invfisher.logit %*%
    c(-1/coef(summary(fit.logit))[2],
      coef(summary(fit.logit))[1]/coef(summary(fit.logit))[2]^2))
# 90% CI
ld50.logit <- c(exp(x_0), exp(x_0 - qnorm(0.95)*x_0.se), exp(x_0 + qnorm(0.95)*x_0.se))
```

LD50 Estimate for probit model:

```r
# Probit
x_0 <- -coef(summary(fit.probit))[1]/coef(summary(fit.probit))[2]
# Point Estimate
# MASS::dose.p(fit.probit, p = 0.5)
x_0.se <- sqrt(
  t(c(-1/coef(summary(fit.probit))[2],
      coef(summary(fit.probit))[1]/coef(summary(fit.probit))[2]^2)) %*%
    invfisher.probit %*%
    c(-1/coef(summary(fit.probit))[2],
      coef(summary(fit.probit))[1]/coef(summary(fit.probit))[2]^2))
# 90% CI
ld50.probit <- c(exp(x_0), exp(x_0 - qnorm(0.95)*x_0.se), exp(x_0 + qnorm(0.95)*x_0.se))
```

LD50 Estimate for cloglog model:

```r
# C log-log
x_0 <- (log(-log(0.5))-coef(summary(fit.cloglog))[1])/coef(summary(fit.cloglog))[2]
# Point Estimate
# MASS::dose.p(fit.cloglog, p = 0.5)
x_0.se <- sqrt(
  t(c(-1/coef(summary(fit.cloglog))[2],
      (coef(summary(fit.cloglog))[1] - log(-log(0.5)))/coef(summary(fit.cloglog))[2]^2)) %*%
    invfisher.cloglog %*%
    c(-1/coef(summary(fit.cloglog))[2],
      (coef(summary(fit.cloglog))[1] - log(-log(0.5)))/coef(summary(fit.cloglog))[2]^2))
# 90% CI
ld50.cloglog <- c(exp(x_0), exp(x_0 - qnorm(0.95)*x_0.se), exp(x_0 + qnorm(0.95)*x_0.se))
```

So the LD50 with 90% confidence interval based on the three models is:

	Estimate of LD50	90% CI lower	90% CI upper
logit	7.389	5.510	9.910
probit	7.436	5.583	9.904
complementary log-log	8.841	6.526	11.977

Problem 2

(a) How does the model fit the data?

Fit the model using logit

```
## # A tibble: 17 x 4
##    amount offers enrolls rejects
##     <dbl>  <dbl>   <dbl>   <dbl>
```

```
## 1      10     4      0      4
## 2      15     6      2      4
## 3      20    10      4      6
## 4      25    12      2     10
## 5      30    39     12     27
## 6      35    36     14     22
## 7      40    22     10     12
## 8      45    14      7      7
## 9      50    10      5      5
## 10     55    12      5      7
## 11     60     8      3      5
## 12     65     9      5      4
## 13     70     3      2      1
## 14     75     1      0      1
## 15     80     5      4      1
## 16     85     2      2      0
## 17     90     1      1      0
```

```
# Fit GLM
fit.logit = glm(cbind(enrolls, rejects) ~ amount, family = binomial(link = 'logit'),
          data = data2)
fit.logit$coefficients
```

```
## (Intercept)      amount
## -1.64763837  0.03095043
```

So the fitted logit model is

$$\hat{\pi}(x) = \frac{e^{-1.64763837+0.03095043x}}{1 - e^{-1.64763837+0.03095043x}}$$

This is grouped data. To evaluate how the model fits the data, we compute pearson chi-squared and deviance for the model. And we test on how the model is close to the full model:

H_0: The model is close to the full model

H_1: not close to full model

```
G.stat <- sum(residuals(fit.logit, type = 'pearson')^2);G.stat # pearson chisq
```

```
## [1] 8.814299
```

```
dev <- fit.logit$deviance;dev # deviance
```

```
## [1] 10.61271
```

```
# compare with chisq(17-2)
pval = 1 - pchisq(dev,15);pval # fit is not good(over dispersion; lack of covariate)
```

```
## [1] 0.7795345
```

The generalized pearson chi-squared statistics is 8.814299, and the deviance is 10.61271. The p-value is large so we do not reject the null hypothesis. The model fits the data fine.

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

```
# No scholarship
exp(coef(summary(fit.logit))[1])
```

```
## [1] 0.192504
```

```r
# OR_(n+1/n)
exp(coef(summary(fit.logit))[2])
```

```
## [1] 1.031434
```

The odds of enrolls when there is zero amount of scholarship is 0.192504. And for a one-unit increase in the amount of scholarship, we expect to see about 3.1% increase in the odds of enrolls among those students who were offered with scholarship.

The 95% CI is:

```r
CI1 = fit.logit$coefficients + kronecker(
  t(c(0,qnorm(0.025),-qnorm(0.025))),
  t(t(sqrt(diag(vcov(fit.logit))))))

out = cbind(exp(CI1)[-1,,drop=FALSE])

colnames(out)=c('Estimate for Beta','95% CI','95% CI')
rownames(out)=c('logit')
out %>% knitr::kable(digits = 3)
```

	Estimate for Beta	95% CI	95% CI
logit	1.031	1.012	1.051

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

```r
r_star <- log(0.4/0.6)
invfisher.logit <- vcov(fit.logit)
# Logit
x_0 <- (r_star-coef(summary(fit.logit))[1])/coef(summary(fit.logit))[2]
# Point Estimate
# MASS::dose.p(fit.logit, p = 0.4)
x_0.se <- sqrt(
  t(c(-1/coef(summary(fit.logit))[2],
      (coef(summary(fit.logit))[1] - r_star)/coef(summary(fit.logit))[2]^2)) %*%
    invfisher.logit %*%
    c(-1/coef(summary(fit.logit))[2],
      (coef(summary(fit.logit))[1] - r_star)/coef(summary(fit.logit))[2]^2))
# 95% CI
ld50.logit <- cbind(x_0, x_0 - qnorm(0.975)*x_0.se, x_0 + qnorm(0.975)*x_0.se)
colnames(ld50.logit) = c('Estimate for Scholarship Amount for 40% Enrollment Rate',
                         '95% CI Lower',
                         '95% CI Upper')
rownames(ld50.logit) = c('logit')
ld50.logit %>% knitr::kable(digits = 3)
```

	Estimate for Scholarship Amount for 40% Enrollment Rate	95% CI Lower	95% CI Upper
logit	40.134	30.583	49.686