Regression analysis_Homework Assignment 6

心理所碩二 **R08227112** 林子堯

2020/12/07

A dose-response experiment yielded the following data:

```
library(tidyverse)
data ← data.frame(LogDose = c(0.71, 1.00, 1.31, 1.48, 1.61, 1.70),
                  GroupSize = c(49, 48, 48, 49, 50, 48),
                  Response = c(16, 18, 34, 47, 47, 48)) %>%
  mutate(NoResponse = GroupSize - Response,
         Propotion = Response / GroupSize)
knitr::kable(data, digits = 2)
```

LogDose	GroupSize	Response	NoResponse	Propotion
0.71	49	16	33	0.33
1.00	48	18	30	0.38
1.31	48	34	14	0.71
1.48	49	47	2	0.96
1.61	50	47	3	0.94
1.70	48	48	0	1.00

Fit a binomial regression model. Check if logit or probit models are satisfactory. Can you find a better fit? Please do residual analysis and check transformations of the dose scale and models with overdispersion.

Our binomial regression model is

$$Y_i \sim Binom(N_i, \pi_i)$$
$$\pi_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x_i}^\top \boldsymbol{\beta})$$

where $i = 1, \ldots, 6$ indicats each grouped observation, and y_i is the response number in the group size N_i with the response probability π_i, which is connected with linear predictor $\eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_i$ by link function $g(.)$.

Logit model

If we choose the logit link function,

```
binom_logit ← glm(formula = cbind(Response, NoResponse) ~ 1 + LogD
ose,
                  data = data,
                  family = binomial(link = "logit"))
## or can write as
# binom_logit ← glm(formula = Propotion ~ 1 + LogDose,
#                   data = data,
#                   weights = GroupSize,
#                   family = binomial(link = "logit"))
summary(binom_logit)
```

```
Call:
glm(formula = cbind(Response, NoResponse) ~ 1 + LogDose, family = b
inomial(link = "logit"),
    data = data)

Deviance Residuals:
       1         2         3         4         5         6
 1.77216  -1.77331  -1.52938   1.63856   0.03622   2.02531

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.4509     0.6104  -7.291 3.07e-13 ***
LogDose       4.4602     0.5155   8.652  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.963  on 5  degrees of freedom
Residual deviance:  15.412  on 4  degrees of freedom
AIC: 37.557

Number of Fisher Scoring iterations: 4
```

the fitted logit model is:

$$g(\hat{\pi_i}) = \ln \frac{\pi_i}{1-\pi_i} = -4.45 + 4.46(LogDose)_i$$

and each β is significance under the significance level $\alpha = 0.05$.

Probit model:

If we choose the probit link function,

```
binom_probit ← glm(formula = cbind(Response, NoResponse) ~ 1 + Log
Dose,
                   data = data,
                   family = binomial(link = "probit"))
summary(binom_probit)
```

```

Call:
glm(formula = cbind(Response, NoResponse) ~ 1 + LogDose, family = b
inomial(link = "probit"),
    data = data)

Deviance Residuals:
      1        2        3        4        5        6
 1.6390  -1.7851  -1.4346   1.5739  -0.2226   1.7598

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6317     0.3427  -7.680 1.59e-14 ***
LogDose       2.6397     0.2797   9.436  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.963  on 5  degrees of freedom
Residual deviance:  13.555  on 4  degrees of freedom
AIC: 35.699

Number of Fisher Scoring iterations: 5
```

the fitted probit model is:

$$g(\hat{\pi_i}) = \Phi^{-1}(\hat{\pi}_i) = -2.63 + 2.64(LogDose)_i$$

where $\Phi(.)$ is the cdf of the standard normal distribution. The result shows that each β is also significance.

Complementary log-log model:

If we choose the complementary log-log link function,

```
binom_cloglog ← glm(formula = cbind(Response, NoResponse) ~ 1 + Lo
gDose,
                    data = data,
                    family = binomial(link = "cloglog"))
summary(binom_cloglog)
```

```

Call:
glm(formula = cbind(Response, NoResponse) ~ 1 + LogDose, family = b
inomial(link = "cloglog"),
    data = data)

Deviance Residuals:
      1        2        3        4        5        6
 1.3318  -1.1989  -0.9193   1.5105  -0.8325   1.1683

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2417     0.4128  -7.853 4.05e-15 ***
LogDose       2.7595     0.3036   9.090  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.9631  on 5  degrees of freedom
Residual deviance:   8.3959  on 4  degrees of freedom
AIC: 30.54

Number of Fisher Scoring iterations: 4
```

the fitted complementary log-log model is:

$$g(\hat{\pi}_i) = log(-log(1-\hat{\pi}_i)) = -3.24 + 2.76(LogDose)_i$$

The result shows that each β is also significance.

Model comparison

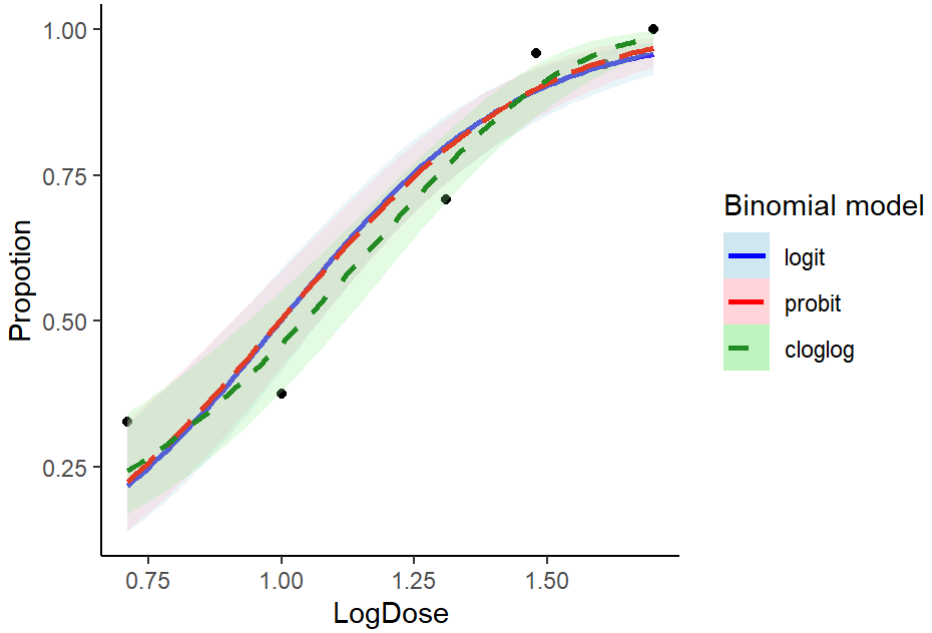
At first glance, the estimators $\hat{\boldsymbol{\beta}}$ form binomial models seem very different from each other. Since the logistic distribution function has variance $\pi^2/3$ and the extreme-value distribution function has variance $\pi^2/6$. We can rescale the estimated coefficients of the probit model and the complementary log-log model with factor $\pi/\sqrt{3}$ and $\sqrt{2}(=\frac{\pi/\sqrt{3}}{\pi/\sqrt{6}})$ respectively. The result is at following table, we can find adjust coefficients is closer to the logit model result.

```
beta ← data.frame(beta_logit = binom_logit$coefficients,
                  beta_probit = binom_probit$coefficients,
                  beta_probit_adj = binom_probit$coefficients * (p
i/sqrt(3)),
                  beta_cloglog = binom_cloglog$coefficients,
                  beta_cloglog_adj = binom_cloglog$coefficients *
 sqrt(2))
knitr::kable(beta, digits = 2)
```

	beta_logit	beta_probit	beta_probit_adj	beta_cloglog	beta_cloglog_adj
(Intercept)	-4.45	-2.63	-4.77	-3.24	-4.58
LogDose	4.46	2.64	4.79	2.76	3.90

We also can present the response functions $\hat{\pi}_i = g^{-1}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ of the three binary regression models

```
ggplot(data, aes(x = LogDose, y = Propotion, weight = GroupSize)) +
  geom_point() +
  geom_smooth(aes(color = "blue", fill = "lightblue", linetype = "s
olid"),
              method = "glm", formula =  y ~ 1 + x,
              method.args = list(family = binomial(link = "logit"
)),
              alpha = 0.25) +
  geom_smooth(aes(color = "red", fill = "lightpink", linetype = "lo
ngdash"),
              method = "glm", formula =  y ~ 1 + x,
              method.args = list(family = binomial(link = "probit"
)),
              alpha = 0.25) +
  geom_smooth(aes(color = "forestgreen", fill = "lightgreen", linet
ype = "dashed"),
              method = "glm", formula =  y ~ 1 + x,
              method.args = list(family = binomial(link = "cloglog"
)),
              alpha = 0.25) +
  scale_color_identity("Binomial model",
                       breaks = c("blue", "red", "forestgreen"),
                       labels = c("logit", "probit", "cloglog"),
                       guide = "legend") +
  scale_fill_identity("Binomial model",
                      breaks = c("lightblue", "lightpink", "lightgr
een"),
                      labels = c("logit", "probit", "cloglog"),
                      guide = "legend") +
  scale_linetype_identity("Binomial model",
                          breaks = c("solid", "longdash", "dashed"
),
                          labels = c("logit", "probit", "cloglog"),
                          guide = "legend") +
  theme_classic()
```



It shows that the response function of logit and probit models are very similar and symmetric at the mean. In the contrast, the response function of the complementary log–log model is asymmetric and showing a faster approach towards 1 as η increasing. However, the complementary log–log model is seemingly closer to the true proportion ($\bar{y}_i = y_i/N_i$) than the other models.

To compare which model has the best goodness-of-fit, we can compared the Pearson statistic and the deviance of the each model. In addition, if we consider the goodness-of-fit and the model's complexity at the same time, the AIC or BIC is alternative criterion for the model choosing.

```
binom_models ← list(logit = binom_logit, probit = binom_probit, cl
oglog = binom_cloglog)
criterion ← data.frame(
  perason_statistic = sapply(binom_models, function(.binom){
    sum(residuals(.binom, type = "pearson")^2)
  }),
  deviance = sapply(binom_models, deviance),
  AIC = sapply(binom_models, AIC),
  BIC = sapply(binom_models, BIC)
)

knitr::kable(criterion, digits = 2)
```

	perason_statistic	deviance	AIC	BIC
logit	13.38	15.41	37.56	37.14
probit	11.93	13.55	35.70	35.28
cloglog	7.54	8.40	30.54	30.12

The result shows that the complementary log-log model has the smallest value in each criteria (including the Pearson statistic to BIC), indicating that this model is better than others. Otherwise, the Pearson statistic and the deviance are asymptotic chi-square distribution with degree of freedom 4 (= the number of groups - the number of estimated coefficients). Under the significance level $\alpha = 0.05$, the critical value is $\chi^2_{0.95,df=4} = 9.5$. The logit and probit model are probably lack of fit. In the conclusion, the complementary log-log model has the best fitted performance.

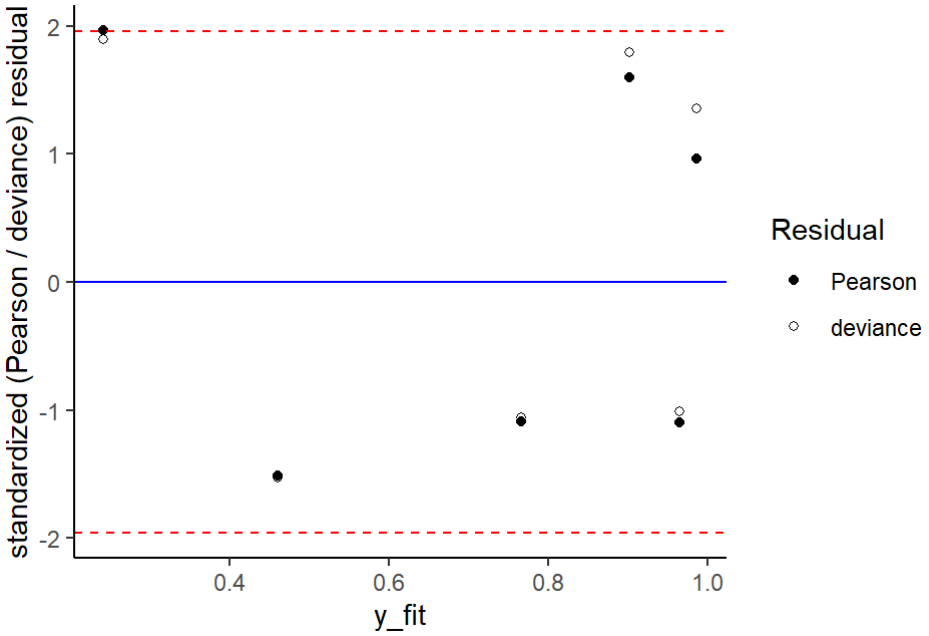
Residual analysis

The following table and plot are presented (standardized) Pearson residuals and (standardized) deviance residuals of the complementary log-log model. The standardized Pearson residuals and standardized deviance residuals are approximately $N(0,1)$. So we can detect whether is there outlier existing or any misspecified assumption in this model.

```
binom_cloglog_res = data.frame(
  LogDose = data$LogDose,
  y_bar = data$Propotion,
  y_fit = fitted(binom_cloglog),
  pearson_res = residuals(binom_cloglog, type = "pearson"),
  stand_pearson_res = rstandard(binom_cloglog, type = "pearson"),
  deviance_res = residuals(binom_cloglog, type = "deviance"),
  stand_deviance_res = rstandard(binom_cloglog, type = "deviance")
)
knitr::kable(binom_cloglog_res, digits = 2)
```

LogDose	y_bar	y_fit	pearson_res	stand_pearson_res	deviance_res	stand_deviance_res
0.71	0.33	0.24	1.38	1.96	1.33	1.90
1.00	0.38	0.46	-1.19	-1.51	-1.20	-1.53
1.31	0.71	0.77	-0.94	-1.09	-0.92	-1.06
1.48	0.96	0.90	1.35	1.60	1.51	1.79
1.61	0.94	0.96	-0.91	-1.10	-0.83	-1.01
1.70	1.00	0.99	0.83	0.96	1.17	1.36

```
ggplot(binom_cloglog_res, aes(x = y_fit)) +
  geom_point(aes(y = stand_pearson_res, shape = 19)) +
  geom_point(aes(y = stand_deviance_res, shape = 1)) +
  geom_hline(yintercept = qnorm(c(0.025, 0.5, 0.975), mean = 0, sd
 = 1),
             color = c("red", "blue", "red"),
             linetype = c("dashed", "solid", "dashed")) +
  scale_shape_identity("Residual", guide = "legend",
                       breaks = c(19, 1), labels = c("Pearson", "de
viance")) +
  labs(y = "standardized (Pearson / deviance) residual") +
  theme_classic()
```



Though the standard Pearson residual of the first observation is outer the criteria a little bit (but standard deviance residual is not), I thought there is not strong evidence to indicate it is an outlier. By the way, we can't find any systematic error between the residuals and y_fit (or LogDose) and the residuals are seemingly independent, so the complementary log-log model may have no problem.

Transformation

If we consider the higher order ($LogDose^2$) in to the model

$$Y_i \sim Binom(N_i, \pi_i)$$
$$\pi_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1(LogDose)_i + \beta_2(LogDose)_i^2)$$

where $g(.)$ still use the cloglog link.

```
binom_cloglog2 ← glm(formula = cbind(Response, NoResponse) ~ 1 + L
ogDose + I(LogDose^2),
                     data = data,
                     family = binomial(link = "cloglog"))
summary(binom_cloglog2)
```

```

Call:
glm(formula = cbind(Response, NoResponse) ~ 1 + LogDose + I(LogDose
^2),
    family = binomial(link = "cloglog"), data = data)

Deviance Residuals:
       1         2         3         4         5         6
 0.38617  -0.78569  -0.09554   1.69867  -1.37161   0.64422

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.9334     1.5074  -0.619    0.536
LogDose       -1.3348     2.6460  -0.504    0.614
I(LogDose^2)   1.6946     1.1061   1.532    0.125

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.9631  on 5  degrees of freedom
Residual deviance:   5.9574  on 3  degrees of freedom
AIC: 30.102

Number of Fisher Scoring iterations: 4
```

The fitted result show that coefficients β_0, β_1 & β_2 are not significance any more. Even though, it has the smaller AIC (or BIC) value than the original model.

```
anova(binom_cloglog, binom_cloglog2, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: cbind(Response, NoResponse) ~ 1 + LogDose
Model 2: cbind(Response, NoResponse) ~ 1 + LogDose + I(LogDose^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     8.3959
2         3     5.9574  1   2.4385   0.1184
```

But from the likelihood ratio test, there is not significance difference between the original model and the higher order model. By the principle of simplicity, I think the original complementary log-log model is the better choice. From the residual plot, we also can't find any other relationship

between residuals and LogDose We don't need higher order on the LogDose.

On the other side, if we don't take the "log" on the Dose covariate, our new model is

$$Y_i \sim Binom(N_i, \pi_i)$$
$$\pi_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1(Dose)_i)$$

where $g(.)$ still use the cloglog link.

```
binom_cloglog_dose ← glm(formula = cbind(Response, NoResponse) ~ 1
+ I(exp(LogDose)),
                       data = data,
                       family = binomial(link = "cloglog"))
summary(binom_cloglog_dose)
```

```

Call:
glm(formula = cbind(Response, NoResponse) ~ 1 + I(exp(LogDose)),
    family = binomial(link = "cloglog"), data = data)

Deviance Residuals:
      1        2        3        4        5        6
 0.6201  -0.9363  -0.2353   1.7049  -1.2738   0.6931

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.71768    0.35534  -7.648 2.04e-14 ***
I(exp(LogDose))  0.80114    0.09022   8.880  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.9631  on 5  degrees of freedom
Residual deviance:   6.3261  on 4  degrees of freedom
AIC: 28.471

Number of Fisher Scoring iterations: 4
```

The fitted new complementary log-log model is:

$$g(\hat{\pi}_i) = log(-log(1 - \hat{\pi}_i)) = -2.72 + 0.80(Dose)_i$$

The result shows that each β is also significance. Furthermore, it also shows that the new model have the better fit than the original model! The Pearson statistic (= 6), deviance (= 6.3), AIC(= 28) and BIC(= 28) are all smaller. Maybe it is the better candidate model.

Model with overdispersion

However, we may observe overdispersion in the original complementary log-log model. We assume

$$Var(y_i) = \phi \frac{\pi_i(1-\pi_i)}{N_i}.$$

The overdispersion paramter ϕ can be estimated as the average Pearson statistic χ^2 or the average deviance D:

$$\hat{\phi}_p = \frac{\chi^2}{G-p} \quad \text{or} \quad \hat{\phi}_D = \frac{D}{G-p}$$

where G is the number of the grouped observation.

```
(phi_p ← criterion$perason_statistic[3] / binom_cloglog$df.residua
l)
```

```
[1] 1.884424
```

```
(phi_D ← criterion$deviance[3] / binom_cloglog$df.residual)
```

```
[1] 2.098972
```

We find $\hat{\phi}_p$ and $\hat{\phi}_D$ are larger than 1, so it indicates that there is overdispersion.

An appropriate approach to this situation is using a quasi-likelihood model.

```
quasibinom_cloglog ← update(binom_cloglog,
                            family = quasibinomial(link = "cloglo
g"))
summary(quasibinom_cloglog)
```

```
Call:
glm(formula = cbind(Response, NoResponse) ~ 1 + LogDose, family = q
uasibinomial(link = "cloglog"),
    data = data)

Deviance Residuals:
      1        2        3        4        5        6
 1.3318  -1.1989  -0.9193   1.5105  -0.8325   1.1683

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.2417     0.5667  -5.721  0.00462 **
LogDose       2.7595     0.4167   6.621  0.00270 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.88447
1)

    Null deviance: 123.9631  on 5  degrees of freedom
Residual deviance:   8.3959  on 4  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

The estimated coefficient $\hat{\beta}_0 = -3.24$ & $\hat{\beta}_1 = 2.76$ is as same as the original complementary log-log model, but with $\hat{\phi} = 1.88$. Other difference is the standard errors of the $\hat{\boldsymbol{\beta}}$ become larger, but luckily, $\hat{\boldsymbol{\beta}}$ are still significance.