# HW2_sj2921

*Shan*

*2/14/2019*

**Problem 1**

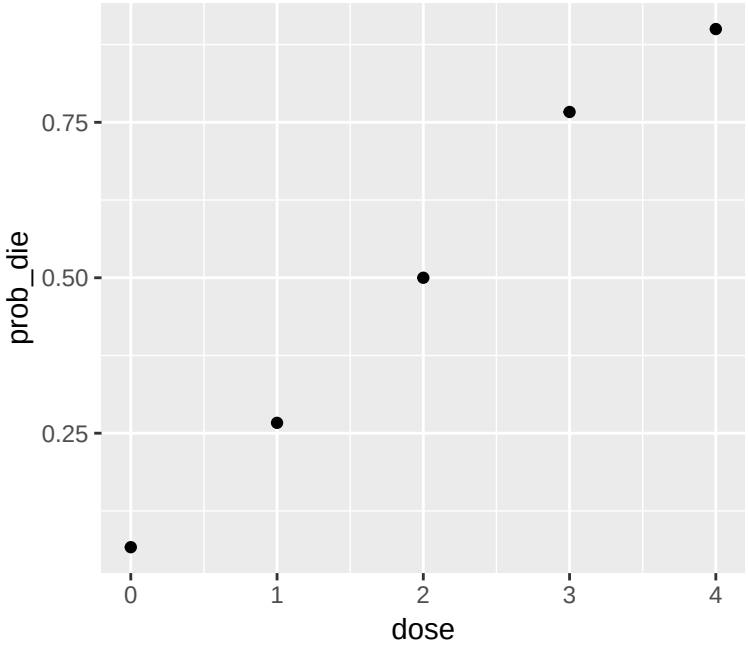
Import data

```
dose = c(0, 1, 2, 3, 4)
dying = c(2, 8, 15, 23, 27)
not_dying = 30 - dying
prob_die = dying/30
bio_df = data.frame(dose, dying, not_dying, prob_die)

## Look at the dataset
knitr::kable(bio_df,digits = 2, "latex", booktabs = T) %>%
kable_styling(latex_options = "striped", full_width = F)
```

| dose | dying | not_dying | prob_die |
|---|---|---|---|
| 0 | 2 | 28 | 0.07 |
| 1 | 8 | 22 | 0.27 |
| 2 | 15 | 15 | 0.50 |
| 3 | 23 | 7 | 0.77 |
| 4 | 27 | 3 | 0.90 |

```
ggplot(bio_df, aes(x = dose, y =  prob_die)) +
  geom_point()
```



### Fit the GLM model

$$g(P(dying)) = \alpha + \beta X$$

### Logit links

### Confidence interval

```
# CI for beta
vcov(bio.logit)
```

```
##             (Intercept)        dose
## (Intercept)  0.17463024 -0.06582336
## dose        -0.06582336  0.03291168
```

```
# variance-covariance matrix of beta MLE (fisher information inverse)

beta = bio.logit$coefficients[2]
se = sqrt(vcov(bio.logit)[2,2]) # (same as in above)
beta + c(qnorm(0.025), 0, -qnorm(0.025)) * se
```

```
## [1] 0.8063266 1.1618949 1.5174633
```

### deviance

```
logit.dev = sum(residuals(bio.logit, type = 'deviance')^2)
logit.dev  ## Residual deviance:  0.37875
```

```
## [1] 0.3787483
```

The estimated regression is

$$log(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}) = -2.3238 + 1.1619x$$

with $\beta_1$ being significant(p <0.01) and positive, implying that dose amount increases so does the probability of dying. As x increase 1 unit the odds of dying increase multiplicative by $e^{1.1619} = 3.196$. That is, for a dose increase, there is about a 219% increase in a chance of dying compared to a consuming 1 unit less dose(within the range of the data).

### P(dying| x = 0.01)

```
# CI for pi
predict(bio.logit,
        data.frame(dose = 0.01),
        se.fit = TRUE, type = 'response')$fit
```

```
##          1
## 0.09011997
```

### probit link

The estimated regression is

$$log(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}) = -1.378 + 0.686x$$

with $\beta_1$ being significant($p < 0.001$) and positive, implying that dose amount increases so does the probability of dying. As x increase 1 unit the odds of dying increase multiplicative by $e^{0.686} = 1.986$. That is, for a dose increase, there is about a 98.6% increase in a chance of dying compared to a consuming 1 unit less dose(within the range of the data).

### Confidence interval

```
# CI for beta
vcov(bio.probit)
```

```
##             (Intercept)         dose
## (Intercept)  0.05189588 -0.018760513
## dose        -0.01876051  0.009363749
```

```
# variance-covariance matrix of beta MLE (fisher information inverse)
beta = bio.probit$coefficients[2]
se = sqrt(vcov(bio.probit)[2,2]) # (same as in above)
beta + c(qnorm(0.025), -qnorm(0.025)) * se # 0 is for point estimate
```

```
## [1] 0.4967217 0.8760393
```

### deviance

```
probit.dev = sum(residuals(bio.probit, type = 'deviance')^2)
probit.dev  ## Residual deviance:  0.3136684
```

```
## [1] 0.3136684
```

### predict

```
# CI for pi
predict(bio.probit, data.frame(dose = 0.01),
        se.fit = TRUE, type = 'response')$fit
```

```
##         1
## 0.0853078
```

### c-log-log link((complementary log-log))

The estimated regression is

$$log(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}) = -1.9942 + 0.7468x$$

with $\beta_1$ being significant($p < 0.001$) and positive, implying that dose amount increases so does the probability of dying. As x increase 1 unit the odds of dying increase multiplicative by $e^{0.7468} = 2.1102$. That is, for a dose increase, there is about a 111% increase in a chance of dying compared to a consuming 1 unit less dose(within the range of the data).

**Confidence interval**

```
# CI for beta
vcov(bio.log)
```

```
##             (Intercept)        dose
## (Intercept)  0.09774304 -0.03111671
## dose        -0.03111671  0.01197721
```

```
# variance-covariance matrix of beta MLE (fisher information inverse)
beta = bio.log$coefficients[2]
se = sqrt(vcov(bio.log)[2,2]) # (same as in above)
beta + c(qnorm(0.025), 0, -qnorm(0.025)) * se # 0 is for point estimate
```

```
## [1] 0.5323200 0.7468193 0.9613187
```

**deviance**

```
c_log.dev = sum(residuals(bio.log, type = 'deviance')^2)
c_log.dev  ## Residual deviance:  2.23048
```

```
## [1] 2.230479
```

**predict:**

```
# CI for pi
predict(bio.log, data.frame(dose = 0.01), se.fit = TRUE, type = 'response')$fit
```

```
##         1
## 0.1281601
```

### 1.1 Table results

```
row1 = c("model", "Estimateof beta" , "CI for beta", "Deviance", "p(dying|x = 0.01)")
row2 = c("logit",  1.1619, "[0.8063, 1.5175]", 0.3788, 0.0901)
row3 = c("probit",  0.6864,  "[0.4967 0.8761]" , 0.3137, 0.0853 )
row4 = c("c-log-log", 0.7468, "[0.5323, 0.9613]" ,2.2305, 0.1282)
p1 = rbind(row1, row2, row3, row4)


kable(p1, "latex", booktabs = T ) %>%
kable_styling(full_width = F) %>%
column_spec(1, bold = T, color = "black") %>%
column_spec(2, width = "10em")
```

| | | | | | |
|---|---|---|---|---|---|
| **row1** | model | Estimateof beta | CI for beta | Deviance | p(dying\|x = 0.01) |
| **row2** | logit | 1.1619 | [0.8063, 1.5175] | 0.3788 | 0.0901 |
| **row3** | probit | 0.6864 | [0.4967 0.8761] | 0.3137 | 0.0853 |
| **row4** | c-log-log | 0.7468 | [0.5323, 0.9613] | 2.2305 | 0.1282 |

**(ii). LD50 with 90% CI**

**(1)logit**

```
# LD50 est and CI
beta0 = bio.logit$coefficients[1]
beta1 = bio.logit$coefficients[2]
betacov = vcov(bio.logit) # inverse fisher information
x0fit = -beta0/beta1
x0fit # Used for cross validation
```

```
## (Intercept)
##           2
```

```
exp(x0fit) # point estimate of LD50
```

```
## (Intercept)
##    7.389056
```

```
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
```

```
## (Intercept)        dose
##   2.0000000   0.1784367
```

```
exp((x0fit + c(qnorm(0.05),-qnorm(0.05))*sqrt(varx0)))
```

```
## [1] 5.509631 9.909583
```

```
# 90% CI for LD50
```

The 90% CI is (5.509631, 9.909583).

**(2) probit**

As the probit model has its unique link function, $g^{-1}(\eta) = \phi(eta)$, so we have $g^{-1}(0.5) = \phi(0.5) = 0$, then we can derive from this that the $x_0 = 0$. SO, the point estimate remains the same while the x est. value is still the same, g(0.5) = 0.

```
# LD50 est and CI
beta0 = bio.probit$coefficients[1]
beta1 = bio.probit$coefficients[2]
betacov = vcov(bio.probit) # inverse fisher information
x0fit = -beta0/beta1
x0fit # Used for cross validation
```

```
## (Intercept)
##     2.00631
```

```
exp(x0fit) # point estimate of LD50
```

```
## (Intercept)
##     7.43583
```

```
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
```

```
## (Intercept)        dose
##   2.0063102   0.1742755
```

```r
exp((x0fit + c(qnorm(0.05),-qnorm(0.05))*sqrt(varx0))) # 90% CI for LD50
```

```
## [1] 5.582588 9.904289
```

Probit: 90% Ci is [5.583, 9.902]

### (3)c-log-clog

The link function has changed into:

$$g_3(\pi) = log(-log(1-\pi))$$

so, we have $g_3(\pi) = log(-log(1-\pi)); g(0.5) = log(-log(1-0.5)) = log(-log(0.5))$

Then we can derive the derivatives from $\beta_0$ and $\beta_1$

```r
# LD50 est and CI
beta0 = bio.log$coefficients[1]
beta1 = bio.log$coefficients[2]
betacov = vcov(bio.log) # inverse fisher information
x0fit = (log(log(2))-beta0)/beta1

x0fit # Used for cross validation
```

```
## (Intercept)
##    2.179428
```

```r
exp(x0fit) # point estimate of LD50
```

```
## (Intercept)
##    8.841249
```

```r
varx0 = betacov[1,1]/(beta1^2) + betacov[2,2]*((beta0- log(log(2)))^2)/(beta1^4) - 2 * betacov[1,2] * (l
c(x0fit,sqrt(varx0)) # point est and se
```

```
## (Intercept)        dose
##   2.1794281   0.1845721
```

```r
exp((x0fit + c(qnorm(0.05),-qnorm(0.05))*sqrt(varx0))) # 90% CI for LD50
```

```
## [1]  6.526261 11.977407
```

The estimate value for three links are as following:

- The logit function:
- LD50 point est. 7.389056; 90% CI [5.5096, 9.9096]
- The probit function:
- LD50 est. 7.436 ; 90% CI [5.583, 9.904]
- C-log-log function:
- LD50 point est. 8.841 ; 90% CI [6.5263, 11.9774]

## Problem 2

### 1. Goodness of fit

```r
##(1) Matched the grouped model fit

beta0 = coef(mph.logit)[1]
beta1 = coef(mph.logit)[2]
beta_0 = amount * beta1
pihat = fitted(mph.logit)

### Pearson-chi-square residual
G.res = (y - offers  * pihat)/sqrt ( offers * pihat *(1 - pihat))
residuals(mph.logit, type = "pearson")
```

```
##          1          2          3          4          5          6
## -1.0243724  0.5717600  0.9810994 -0.9711570 -0.2647020  0.3289076
##          7          8          9         10         11         12
##  0.5320014  0.4781082  0.1583651 -0.6722082 -1.0076450 -0.2103582
##         13         14         15         16         17
##  0.1424593 -1.4004895  0.5054964  0.8650220  0.5661196
```

```r
sum(residuals(mph.logit, type = "pearson")^2)
```

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

```r
## Deviance
dev = sum(residuals(mph.logit, type = "deviance")^2)
dev
```

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

```r
## compare with chi-square
pval = 1 - pchisq(dev, 15) ## 2 parameters: 17-2 =15
## pval is 0.77 > 0.05, fail to reject the null hypothesis, our model fits well enough.
```

The residual deviance and test results shows that the pvalue is 0.77, which means we cannot reject the null hypothesis, so our model can be a suitable fit.

### 2. Relationship between the scholarship amount and enrollment rate

```r
beta0
```

```
## (Intercept)
##   -1.647638
```

```r
beta1
```

```
##     amount
## 0.03095043
```

$$log(\frac{\pi}{1-\pi}) = amount \cdot \beta_1 + \beta_0$$

$$log(\frac{\pi}{1-\pi}) = amount \cdot 0.031 - 1.648$$

*Interpretation*:

(1). In this study, we find that when there is no scholarship provided for mph students, log odds of enrollment would be -1.648;

(2). For one thousand dollar increase of scholarship provided for mph students, the log odds of enrollment would increase by 0.031 keeping other factors the costant.

What is 95% CI?

```
# CI for beta
vcov(mph.logit)
```

```
##               (Intercept)        amount
## (Intercept)  0.177611410 -3.809551e-03
## amount      -0.003809551  9.369767e-05
```

```
# variance-covariance matrix of beta MLE (fisher information inverse)
beta = mph.logit$coefficients[2]
se = sqrt(vcov(mph.logit)[2,2]) # (same as in above)
beta + c(qnorm(0.025), 0, -qnorm(0.025)) * se # 0 is for point estimate
```

```
## [1] 0.01197845 0.03095043 0.04992240
```

```
# CI for odds ratio: exp(beta); tranfer back.
exp(beta + c(qnorm(0.025),0,-qnorm(0.025)) * se)
```

```
## [1] 1.012050 1.031434 1.051190
```

The 95% CI for $\beta_1$ is [0.01198, 0.04992], we are 95% confident that the coefficient of amount falls between 0.01198 and 0.04992.

The 95% CI for odds ratio is [1.01205, 1.05119], we are 95% confident that the odds ratio of amount falls between 1.012050 and 1.051190, which is greater than 1, implying a positive correalation of scholarship and enrollment.

**3. Get 40% yield rate (the percentage of admitted students who enroll?) What is the 95% CI?**

```
# 40 yield rate est and CI
vcov(mph.logit) # inverse fisher information
```

```
##               (Intercept)        amount
## (Intercept)  0.177611410 -3.809551e-03
## amount      -0.003809551  9.369767e-05
```

```
x_fit = (log(0.4/(1 - 0.4)) - beta0)/beta1  ## point est.  40.13429
beta1_sq = beta1^2
# point estimate of 40% yield rate
varx0 = betacov[1,1] *  (1/beta1)^2 + (log(2/3) - beta0)^2*betacov[2,2] / (beta1^4) + 2 * (betacov[1,2]

c(x_fit,sqrt(varx0)) # point est and se
```

```
## (Intercept)      amount
##    40.13429   132.79449
```

```
x_fit + c(qnorm(0.025),-qnorm(0.025))*sqrt(varx0) # 95% CI for yield rate 40%
```

```
## [1] -220.1381  300.4067
```

- We should provide **$40,134 dollars** scholarship to get 40% yield rate;
- Through using the Yield rate of 40%, we get the point est. of scholarship amount, the 95% Confidence interval is [30.58304, 49.68553], implying that we are 95% confident that the est. of scholarship amount falls between $30,583 and $49,685.