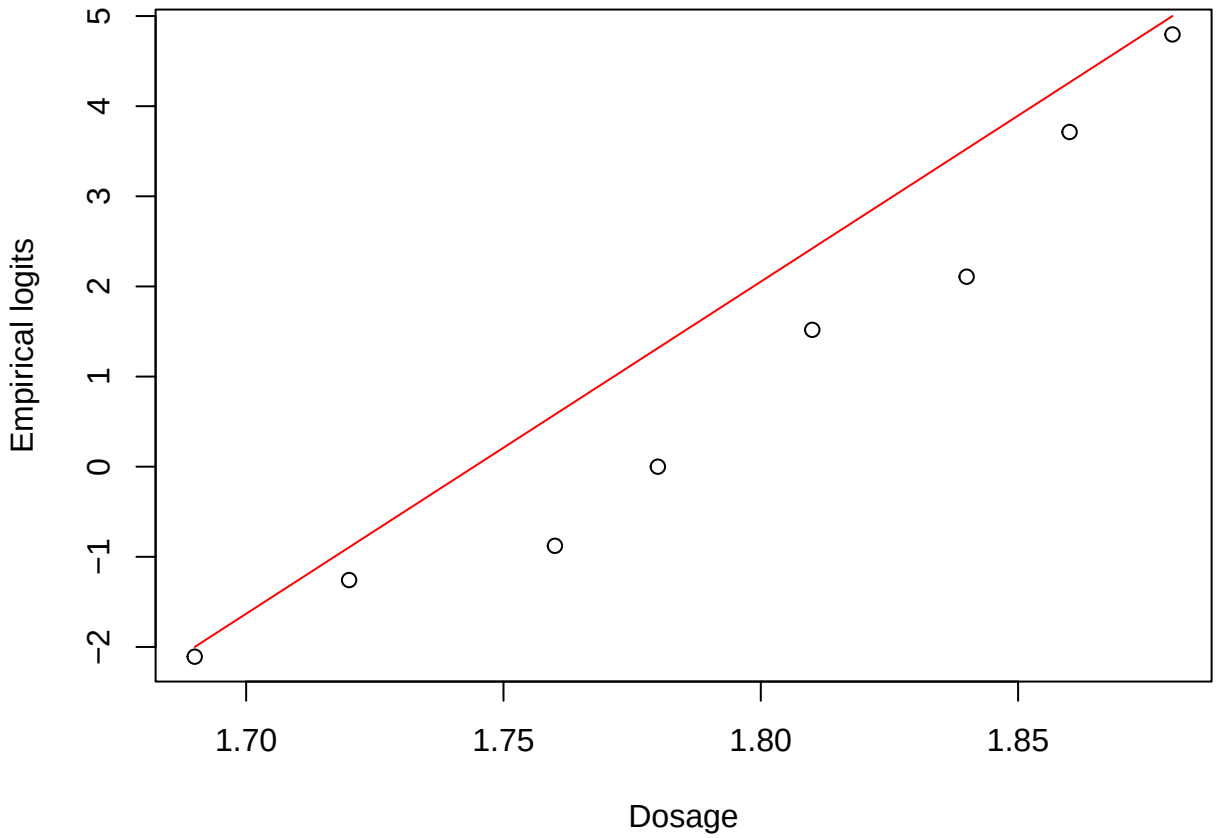
MAST90139 Assignment 2

Haonan Zhong 867492 (Wed 14:15 Qiuyi Li)

Question 1a

```
x <- c(1.69, 1.72, 1.76, 1.78, 1.81, 1.84, 1.86, 1.88)
n <- c(59, 60, 62, 56, 63, 59, 62, 60)
y <- c(6, 13, 18, 28, 52, 53, 61, 60)

## Compute the empirical logit
emp.logit <- log((y + 0.5)/(n - y + 0.5))
par(mar=c(4,4,1,1))
plot(x, emp.logit, xlab = 'Dosage', ylab = 'Empirical logits')
lines(c(1.69, 1.88), c(-2, 5), type = "l", col = "red")
```



We can see there's a somewhat linear trend presented in the plot.

Question 1b

```
logistic <- glm(y/n ~ x, family = binomial, weights = n)
summary(logistic)
```

```
##
## Call:
## glm(formula = y/n ~ x, family = binomial, weights = n)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8986  -0.5475   0.9842   1.3315   1.7179
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.103      5.164  -11.64   <2e-16 ***
## x             33.934      2.903   11.69   <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: 284.202  on 7  degrees of freedom
## Residual deviance:  13.633  on 6  degrees of freedom
## AIC: 43.831
##
## Number of Fisher Scoring iterations: 4
```

The estimate for the intercept is -60.10328, and the slope is 33.93416.

Question 1c

```
confint(logistic)[2,]
```

```
## Waiting for profiling to be done...
```

```
##    2.5 %   97.5 %
## 28.54467 39.96005
```

The 95% confidence interval for the slope is (28.54467, 39.96005).

Question 1d

Given that $logit(0.5) = \frac{0.5}{1-0.5} = 0$.

$$-60.1033 + 33.9341 \times dosage = 0$$

Solving the above equation yields dosage = 1.7712. Therefore, the estimate of the dosage that will kill 50% of the beetles is 1.7712.

Question 1e

```
# Compute the odds ratio
beta1 <- summary(logistic)$coef[2]
(odds <- exp(0.1 * beta1))
```

```
## [1] 29.76748
```

The estimated odds of being killed for a 0.1 increase in dosage is 29.76749.

```
exp(0.1*confint(logistic)[2,])
```

```
## Waiting for profiling to be done...

##    2.5 %   97.5 %
## 17.36518 54.38045
```

And the 95% confidence interval for the odds ratio is (17.36518, 54.38045).

Question 1f

```
etahat <- summary(logistic)$coef[1] + summary(logistic)$coef[2] * 1.8
(probability <- exp(etahat)/(1 + exp(etahat)))
```

```
## [1] 0.7267543
```

```
X.pred <- matrix(c(1, 1.8), nrow = 1, ncol = 2)
se <- sqrt(X.pred %*% summary(logistic)$cov.scaled %*% t(X.pred))
eta_l <- etahat - 1.96 * se
eta_r <- etahat + 1.96 * se
c(exp(eta_l)/(1 + exp(eta_l)), exp(eta_r)/(1 + exp(eta_r)))
```

```
## [1] 0.6671042 0.7792529
```

Therefore is estimated probability is 0.7268, and the 95% confidence interval is (0.6671042, 0.7792529).

Question 1g

```
# Testing using residual deviance
(p_value <- 1 - pchisq(deviance(logistic), df.residual(logistic)))
```

```
## [1] 0.03401062
```

Given the p-value is 0.03401062, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis, the model is not adequate.

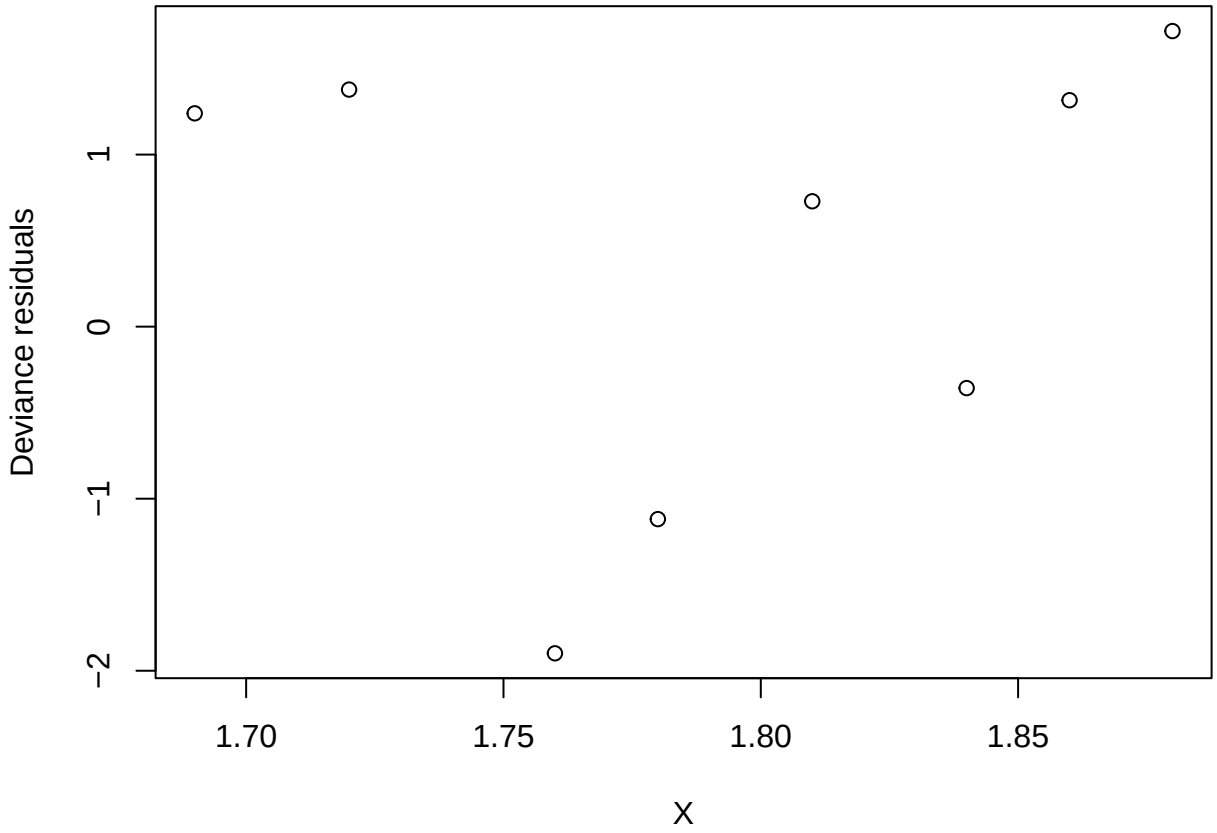
```r
# Testing using Pearson Chi-square test
(p_value <- 1 - pchisq(sum(resid(logistic, type = 'pearson')^2), df.residual(logistic)))
```

```
## [1] 0.05948877
```

In the case of using Pearson χ^2 test, the p-value is 0.05948877, which is slightly above the significance level of 0.05. Therefore, we claim the model is adequate.

Question 1h

```r
par(mar=c(4,4,1,1))
plot(x, resid(logistic, type = 'deviance'), xlab = 'X', ylab = 'Deviance residuals')
```



No evidence of a pattern can be seen from the plot, range of deviance residuals are bounded between -2 and 2.

Question 1i

```r
quad.logistic <- glm(y/n ~ x + I(x^2), family = binomial, weight = n)

# Performing likelihood ratio test between straight line model and quadratic logistic model
anova(logistic, quad.logistic, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y/n ~ x
## Model 2: y/n ~ x + I(x^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         6     13.633
## 2         5      5.107  1   8.5264   0.0035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given the p-value is 0.0035, which is less than the significance level of 0.05, the quadratic term is relevant. Therefore, we claim that the quadratic model provides a better fit.

Question 2a

```
educ <- rep(6:17, 2)
agree <- c(25, 27, 75, 29, 32, 36, 115, 31, 28, 9, 15, 3,
           17, 26, 91, 30, 55, 50, 190, 17, 18, 7, 13, 3)
disagree <- c(9, 15, 49, 29, 45, 59, 245, 70, 79, 23, 110, 29,
              5, 16, 36, 35, 67, 62, 403, 92, 81, 34, 115, 28)
total <- agree + disagree
gender <- c(rep(1, 12), rep(0, 12))
```

```
add.logistic <- glm(agree/total ~ factor(educ) + factor(gender),
                    family = binomial, weight = total)
pchisq(deviance(add.logistic), df = df.residual(add.logistic), lower.tail = F)
```

```
## [1] 0.1752889
```

As we can see from the result, given the p-value is higher than the significance level of 0.05, therefore our model is adequate and the effects of gender and years of education are additive on the logit scale.

Question 2b

```
logistic.0 <- glm(agree/total ~ educ * factor(gender), family = binomial, weight = total)
logistic.1 <- glm(agree/total ~ educ + factor(gender), family = binomial, weight = total)
logistic.2 <- glm(agree/total ~ factor(gender), family = binomial, weight = total)
logistic.3 <- glm(agree/total ~ educ,family = binomial, weights = total)
```

```
pchisq(deviance(logistic.0), df = df.residual(logistic.0), lower.tail = F)
```

```
## [1] 0.4427514
```

```
pchisq(deviance(logistic.1), df = df.residual(logistic.1), lower.tail = F)
```

```
## [1] 0.2434069
```

```
pchisq(deviance(logistic.2), df = df.residual(logistic.2), lower.tail = F)
```

```
## [1] 2.787083e-58
```

```
pchisq(deviance(logistic.3), df = df.residual(logistic.3), lower.tail = F)
```

```
## [1] 0.2859125
```

```
anova(logistic.1, logistic.0, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: agree/total ~ educ + factor(gender)
## Model 2: agree/total ~ educ * factor(gender)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     25.087
## 2        20     20.244  1    4.843  0.02776 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(logistic.3, logistic.0, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: agree/total ~ educ
## Model 2: agree/total ~ educ * factor(gender)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        22     25.236
## 2        20     20.244  2    4.992  0.08241 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the chi-square goodness of fit test, the only models that provide an adequate fit to the data are **logistic.0**, **logistic.1**, and **logistic.3**. Furthermore, we see that **logistic.0** is better than **logistic.1** based on LRT. However, **logistic.0** is not significantly better than **logistic.3**, given the p-value is larger than 0.05. Hence we conclude that **logistic.3** with education as the only predictor is a better model and it is simpler as well.

```
summary(logistic.3)$coef
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  2.8828628  0.2145396  13.43744 3.648330e-41
## educ        -0.3029687  0.0185458 -16.33624 5.450858e-60
```

Therefore, for every one level increased in years of education, the estimated odds of the respondent agreeing to the statement decreased by $1 - e^{-0.3029687} = 26.14\%$.

Question 2c

From part a, we discovered that the effects of gender and years of education are additive on the logit scale. From part b, we found out that the model with only education is the better model. Here, we will carry out further analysis to see which model is more appropriate.
Given that **logistic.3** is nested inside the additive model, we will use likelihood ratio test to compare them.

```
add.logistic <- glm(agree/total ~ factor(educ) + factor(gender),
                    family = binomial, weight = total)
logistic.3 <- glm(agree/total ~ educ,family = binomial, weights = total)
anova(logistic.3, add.logistic, test = "Chi")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: agree/total ~ educ
## Model 2: agree/total ~ factor(educ) + factor(gender)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        22     25.236
## 2        11     15.160 11   10.076   0.5235
```

Despite both model provides an adequate fit to the data, here we see that treating education as a factor and include gender in the model is actually not relevant compare to treating education as a variable. Therefore, we may conclude that **logistic.3** with education as the only predictor is a more appropriate model.

Question 3a

Hypothesis 1

```
defendant_race <- factor(c(1, 1, 2, 2, 1, 1, 2, 2))
victim_race <- factor(c(1, 2, 1, 2, 1, 2, 1, 2))
death_penalty <- factor(c(1, 1, 1, 1, 2, 2, 2, 2))
freq <- c(19, 0, 11, 6, 132, 9, 52, 97)
data <- data.frame(cbind(defendant_race, victim_race, death_penalty, freq))

log_linear <- glm(freq ~ defendant_race + victim_race + death_penalty,
                  family = poisson, data = data)
pchisq(deviance(log_linear), df = df.residual(log_linear), lower.tail = F)
```

```
## [1] 7.83249e-29
```

As we can see, the p-value is smaller than 0.05. Therefore, the model is not adequate and the three factors are not mutually independent.

Hypothesis 2

```
log_linear <- glm(freq ~ death_penalty + defendant_race * victim_race,
                  family = poisson, data = data)
pchisq(deviance(log_linear), df = df.residual(log_linear), lower.tail = F)
```

```
## [1] 0.04336859
```

Given the p-value 0.0434 is smaller than the significance level of 0.05. Therefore, the model is not adequate. Thus, sentence is not independent of both the defendant's and the victim's race.

Hypothesis 3

```
log_linear <- glm(freq ~ defendant_race * death_penalty + defendant_race * victim_race,
                  family = poisson, data = data)
pchisq(deviance(log_linear), df = df.residual(log_linear), lower.tail = F)
```

```
## [1] 0.01915713
```

As the p-value is smaller than 0.05, we can conclude that the model is not adequate. And given defendant's race, sentence is not independent of the victim's race.

Hypothesis 4

```
log_linear <- glm(freq ~ victim_race * death_penalty + defendant_race * victim_race,
                  family = poisson, data = data)
pchisq(deviance(log_linear), df = df.residual(log_linear), lower.tail = F)
```

```
## [1] 0.3902578
```

As we can see, the p-value is greater than 0.05, and the model is a good fit. And given the victim's race, sentence is independent of the defendant's race.

Question 3b

```
defendant_race <- factor(c(1, 1, 2, 2))
victim_race <- factor(c(1, 2, 1, 2))
sentence <- c(19, 0, 11, 6)
total <- c(151, 9, 63, 103)
data <- data.frame(cbind(defendant_race, victim_race, sentence, total))
```

Hypothesis 1

```
logistic <- glm(sentence/total ~ 1, weights = total,
                data = data, family = binomial)
pchisq(deviance(logistic), df = df.residual(logistic), lower.tail = F)
```

```
## [1] 0.04336859
```

Given the p-value is less than 0.05, therefore the model is not adequate, and the three factors are not mutually independent.

Hypothesis 2

```
logistic <- glm(sentence/total ~ 1, weights = total,
                data = data, family = binomial)
pchisq(deviance(logistic), df = df.residual(logistic), lower.tail = F)
```

```
## [1] 0.04336859
```

The p-value is less than 0.05, hence not adequate. And sentence is not independent of both the defendant's and the victim's race.

Hypothesis 3

```
logistic <- glm(sentence/total ~ defendant_race, weights = total,
                data = data, family = binomial)
pchisq(deviance(logistic), df = df.residual(logistic), lower.tail = F)
```

```
## [1] 0.01915713
```

The log-linear model with interaction between defendant's race and sentence allows for an association between defendant's race and sentence, and is therefore equivalent to the logit model where the response depends only on defendant's race. As the p-value suggests, the model is not adequate. Therefore, given the defendant's race, sentence is not independent of the victim's race.

Hypothesis 4

```
logistic <- glm(sentence/total ~ victim_race, weights = total,
                data = data, family = binomial)
pchisq(deviance(logistic), df = df.residual(logistic), lower.tail = F)
```

```
## [1] 0.3902578
```

The log-linear model with interaction between victim's race and sentence allows for an association between victim's race and sentence, and is therefore equivalent to the logit model where the response depends only on victim's race. As the p-value suggest, the model is adequate. Thus, hypothesis 4 holds.