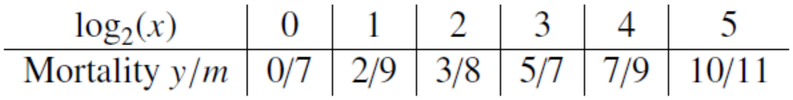
2132 HW3

Orly Olbum

Problem 1

In toxicology, the LD50 is the dose that causes a 50% mortality rate. Experiments are often carried out at a sequence of dose levels, x0, x1, x2,. . . , each dose being twice the preceding dose, where we will assume for simplicity that x0 = 1. The model most commonly used in toxicology is linear in log dose. Suppose that the following results have been obtained in an experiment at various multiples of the baseline dose.



$\log_2(x)$	0	1	2	3	4	5
Mortality y/m	0/7	2/9	3/8	5/7	7/9	10/11

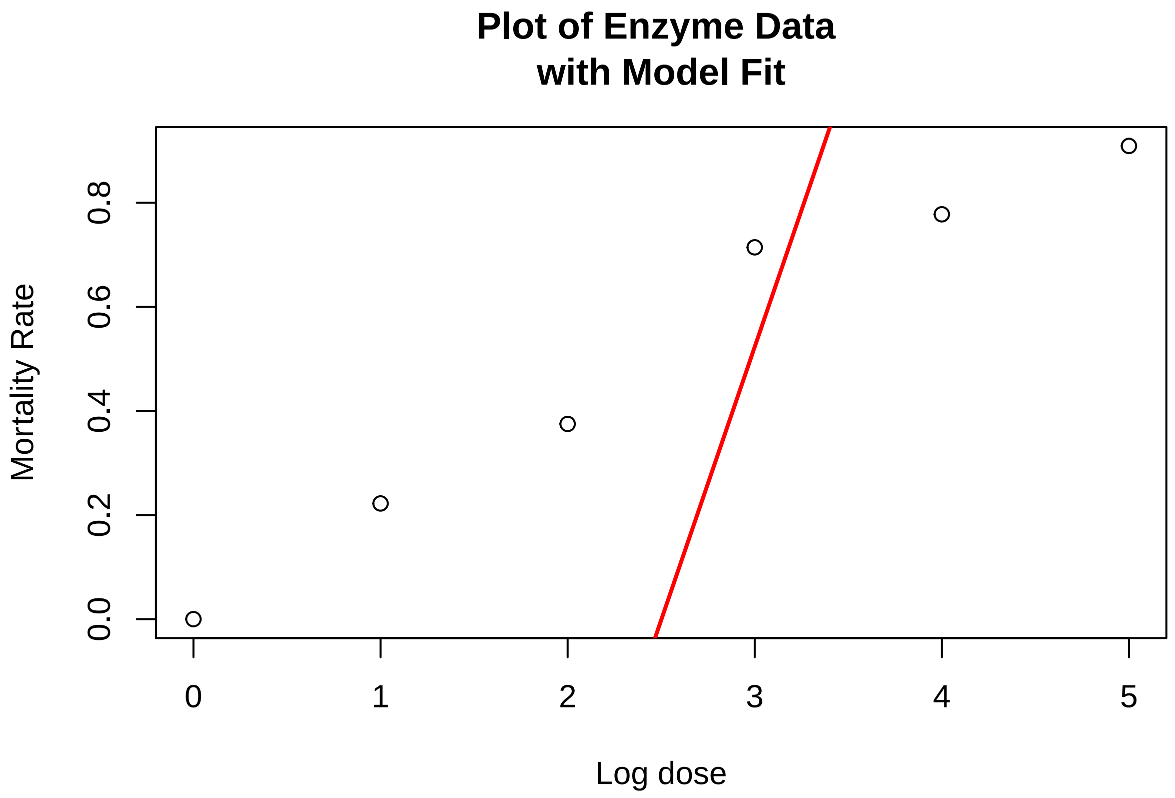
In the above table, x is the dose and y/m is the number of deaths (y) occurring in a sample of m individuals.

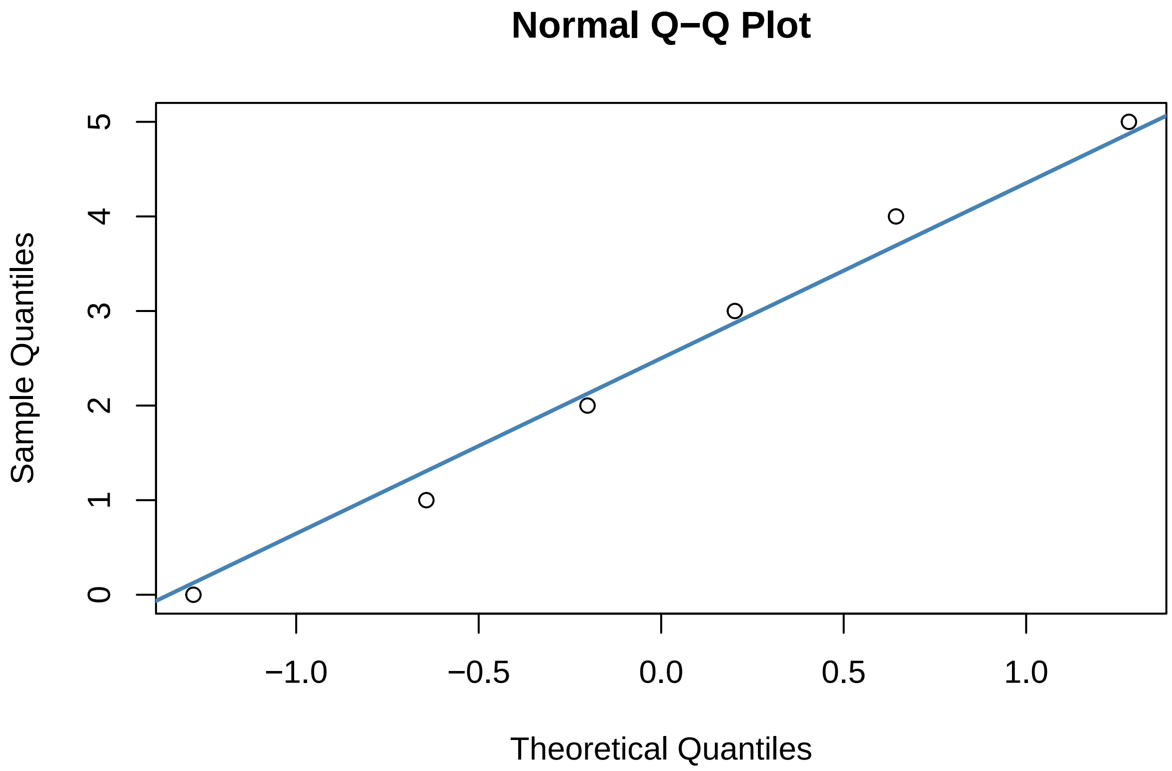
(a) Consider a model in which the logit of the mortality rate is linear in log dose. Report the probability model you are assuming, and define all coefficients in your model. Remember to state which observations you are assuming are independent.

Logit model: **logit(y/m) = B0 + B1Xi**, where Xi is log-base-2 of dose and y/m is probability of death. B0 is the log odds of mortality for Xi = 0, and B1 is the odds ration (OR) between Xi+1 and Xi. The assumptions for the logit model are as follows, with y being independent (Y given X):
- the link function must be strictly increasing or decreasing (satisfied by logit)
- the outcome must be binary - no transformation of Y, only of the mean of Y - there is a linear relationship between the logit and the independent variable
- there are on influential values
- there is no multicollinearity

(b) Fit the linear logistic model in which the logit of the mortality rate is linear in log dose, and plot the fitted mortality rate and raw mortality fractions y/m against log dose (remember the fitted mortality rate is a continuous function). Do you think the model you assumed in part (a) is reasonable?

Because the response variable mortality is presented as proportions between 0 and 1 and not binary as either 0 or 1, we use the quasibinomial family of distributions for the logit.


Plot of Enzyme Data
with Model Fit
Mortality Rate
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
Log dose



While we have very few data points and the model fit does not exactly appear to follow the data points, the normal qq-plot shown above does satisfy what we look for in a linear regression, which would indicate our model fits the data well. The model output shows that log dose is a significant predictor (at 95% confidence) for mortality rate with a p-value of **0.003**.

(c) Estimating gamma = LD50.

(i) Obtain an estimate for gamma.

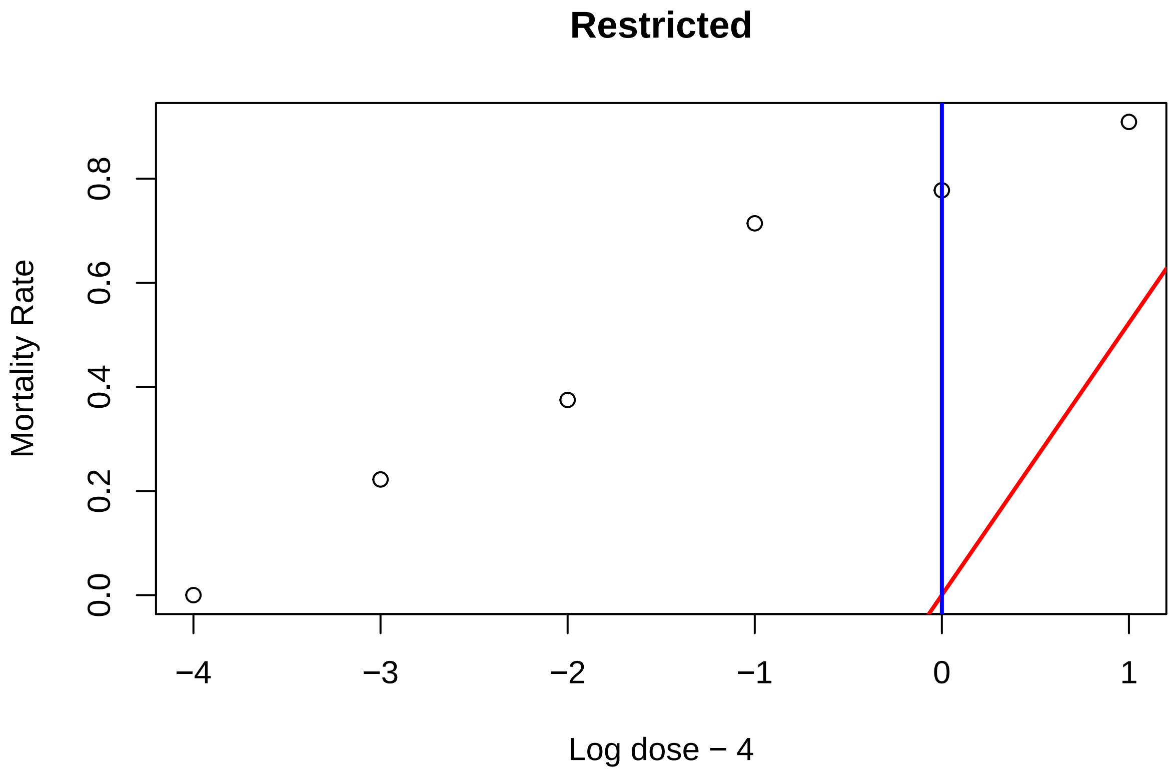
LD50 is the gamma for which the probability of mortality is 50%, which is equivalent to setting B0 + B1gamma = 0 and solving for gamma. We get gamma = -B0/B1 = **2.502**.

(ii) Consider the null hypothesis that log2(gamma) = 4 as a sub-model or restriction of the linear logistic model. Fit the sub-model and compute the log likelihood ratio statistic LR(4). If the null hypothesis is correct, what is the approximate distribution of LR(4)? Compute the p-value.

Log2(gamma) = 4 gives us -B0/B1 = 4, or B0 = -4B1, which reduces our model to one B instead of two. Once we shift the data our hypothesis essentially is that our model has no intercept (reduced model). We can then compare this to the full model from above, and determine if the intercept was adding anything or not.

The LR test statistic **-1.323** returns a p-value of **0.000**.

(iii) By plotting the restricted log likelihood against the hypothesized value of log2(gamma), construct a likelihood-based 95% confidence set for gamma.



From the calculated confidence interval for the one B left (B1), we can calculate an interval for gamma: **(0.083, 1.135)**.

STAT 2132
Orly Olbum

HW3 Problem ②

$\hat{\pi}(x)$ → estimated mortality
$y(x)$ → est. measured #deaths
$m(x)$ → est. # individuals at dose x

$$\hat{\sigma}^2 = \frac{1}{|I|-2} \sum_{x \in I} \frac{\{y(x) - m(x)\,\hat{\pi}(x)\}^2}{m(x)\,\hat{\pi}(x)\{1 - \hat{\pi}(x)\}}$$

where I → set of doses considered in this experiment

ⓘ Showthat $(|I|-2)\hat{\sigma}^2 = X^2$, where X^2 is Pearson's statistic

we have probability of death $\hat{\pi}(x)$, count of deaths $y(x)$, and individuals at x $m(x)$

If $$\hat{\sigma}^2(|I|-2) = \sum_{x \in I} \frac{\{y(x) - m(x)\,\hat{\pi}(x)\}^2}{m(x)\,\hat{\pi}(x)\{1 - \hat{\pi}(x)\}},$$

We know $$X^2 = \sum_{j=1}^{K} \frac{(n_j - np_j)^2}{np_j} \quad \left\{ \begin{array}{l} n_j = \text{\# successes} \\ p_j = \text{prob (success) \& } \sum p_j = 1 \end{array} \right\}$$

And we have n_j = measured deaths = $y(x)$
n = individuals = $m(x)$
p_j = mortality = $\hat{\pi}(x)$

where $y(x)$ are observed values
$m(x)\,\hat{\pi}(x)$ are expected
$m(x)\,\hat{\pi}(x)[1 - \hat{\pi}(x)]$ is variance
which has the form and qualifications of Pearson's statistic,

so $(|I|-2)\hat{\sigma}^2 = X^2$

(ii) Asymptotic dist of $\hat{\sigma}^2$ as $m(x) \to \infty$ for $x \in I$, and what is the mean? what in the data suggests sample size may be too small to assume asymptotic dist is approx. correct?

we know that a χ^2 variable has gamma distribution, if χ^2_m (df=m) $\sim$ gamma$(\frac{m}{2}, \frac{1}{2})$

Since we saw from (i) that

$\hat{\sigma}^2(|I|-2)$ is pearson's statistic X^2;

$\hat{\sigma}^2$ has chisquare distribution and as $m(x) \to \infty$, asymptotically has gamma distribution is $(5-1)(2-1) = 4$

$\sim$gamma$(2, \frac{1}{2})$

with mean $= \alpha/\beta = {}^{m/2}/_{\frac{1}{2}} = m = 4$

Having an entry of 0 might indicate that our sample size is too small to assume asymptotic distribution as presented above.

(iii) say $\pi(x)$ from ⓐ is correct, & $y(x) \perp y(x'), x \neq x'$. If $\hat{\sigma}^2$ is large, what does this say about var$\{y(x)\}$? what's causing the difference?

The variance of $y(x)$ will be higher because we are looking at whole numbers and not proportions

Additionally, $\hat{\sigma}^2$ is based on our data (low $\overset{n}{=}$ sample, high variance) where theoretically we might expect lower variance from a larger sample.

→

(iv) Is CI from ⓒ too narrow/wide if $\hat{\sigma}^2$ were large? small?

For large σ^2, the CI will be wide because our sample size is small. If $\hat{\sigma}^2$ were small, our CI would be too narrow.

(v) what is $\hat{\sigma}^2$ in these data? Is it what we expected based on ⓐ being correct?

$$\hat{\sigma}^2 = \frac{1}{(|I|-2)} \cdot \sum_{x \in I} \frac{\{y(x) - m(x)\hat{\pi}(x)\}^2}{m(x)\hat{\pi}(x)[1-\hat{\pi}(x)]}$$

where $\hat{\pi}(x) \Rightarrow \text{logit}(\pi_i) = \beta_0 + \beta_1 X_i$

$$= -2.6263 + 1.0498(\text{logdose})$$

so $\hat{\pi}(x) = \dfrac{\exp(-2.6 + 1.0 \cdot \text{logdose})}{1 + \exp(-2.6 + 1.0 \times \text{logdose})}$

and $\hat{\sigma}^2 = \frac{1}{(|I|-2)} \cdot 1.1498$

This is much lower than var$\{y(x)\}$, as expected from (iii)

~~~
setwd("C:/Users/orlyo/OneDrive/Desktop/Grad School/Spring 2021/STAT 2132 - Applied Stat. Methods
2/Homeworks/HW3")
data = read.csv("data.csv")

## Problem 1

```{r}
model1 = glm(mortality ~ log.2.x, family = quasibinomial(link = "logit"), data = data)
# summary(model1)
plot(data$log.2.x, data$mortality, main = "Plot of Enzyme Data \nwith Model Fit",
    xlab = "Log dose", ylab = "Mortality Rate")
abline(model1, col = "red", lwd = 2)

qqnorm(data$log.2.x, pch = 1, frame = TRUE)
qqline(data$log.2.x, col = "steelblue", lwd = 2)
```

```{r}
gamma.hat = model1$coefficients[1]/model1$coefficients[2]*-1
# gamma.hat
```

```{r}
data$new.x = data$log.2.x-4
model2 = glm(mortality ~ new.x + 0, family = quasibinomial(link = "logit"), data = data)

lr.stat = model1$deviance - model2$deviance
# lr.stat
pvalue.H0.lr = pchisq(q = lr.stat, df = 4, lower.tail = T)
# pvalue.H0.lr
```

```{r}
plot(data$new.x, data$mortality, main = "Restricted", xlab = "Log dose - 4", ylab = "Mortality Rate")
abline(model2, col = "red", lwd = 2)
abline(v = 0, col = "blue", lwd = 2)

ci = confint(model2)
```
~~~