04 - Logistic models

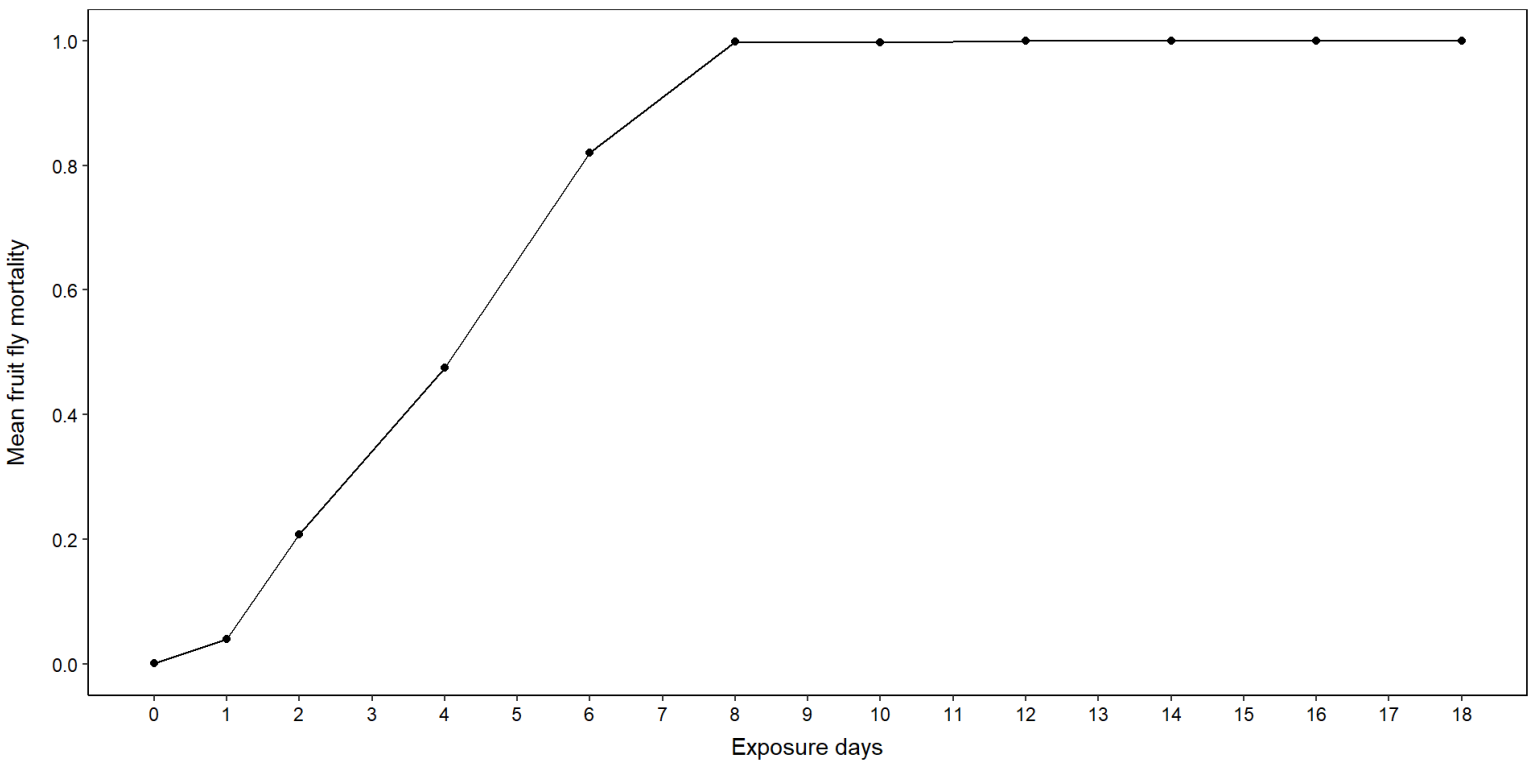
Guy F. Sutton
Centre for Biological Control
Rhodes University, South Africa
Email: g.sutton@ru.ac.za

Binary/Proportion data

- Another set of common data types in ecology and agriculture are binomial and proportion data
  - Binomial data: There are only two categorical levels (e.g. dead/alive, present/absent, infected/healthy)
  - Proportion data: Any data in the range of [0,1] (e.g. Proportion of insects that survived or emerged)
- The Gaussian and Poisson/NB GLM analyses are completely in appropriate for analysing these data

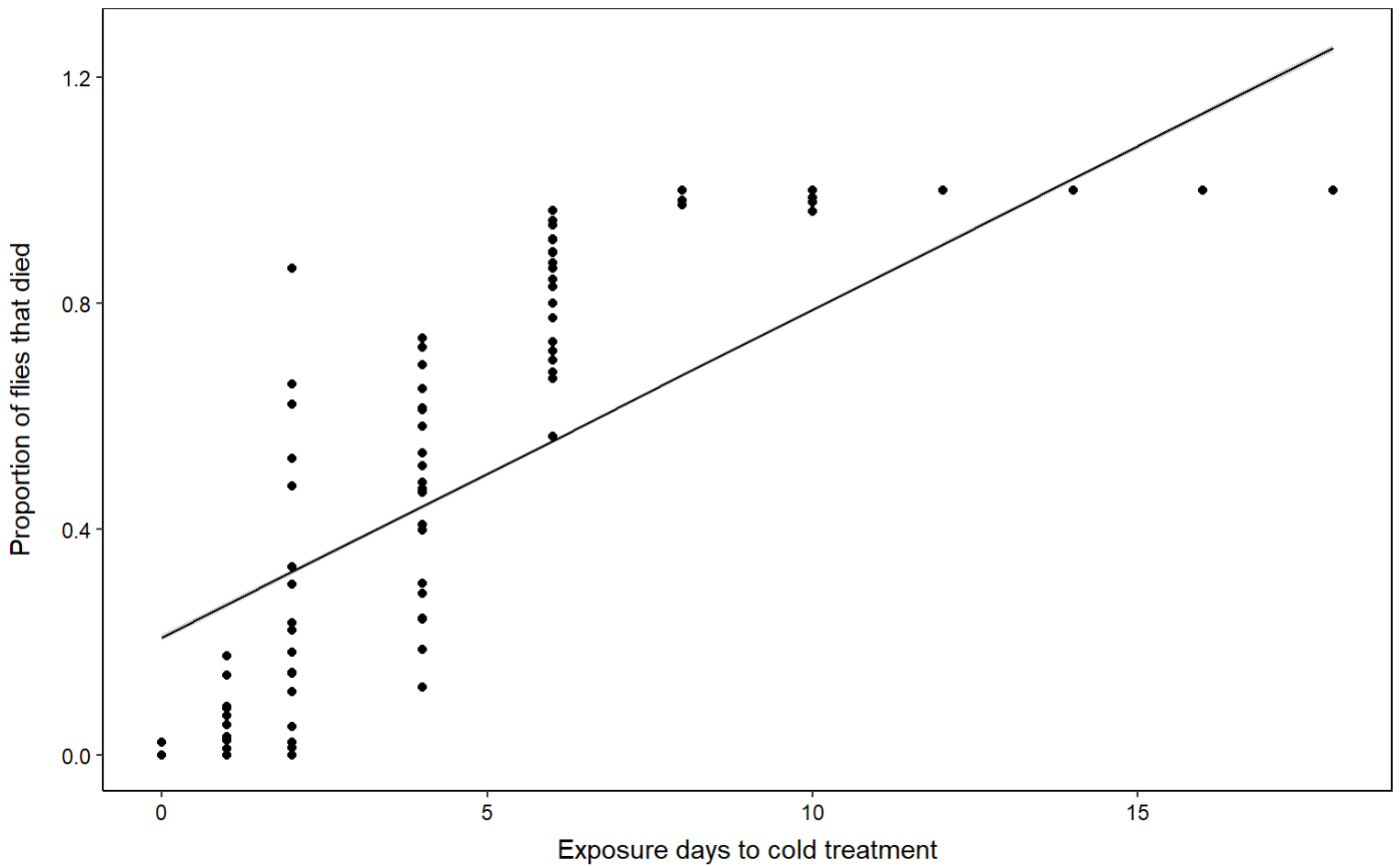
An example

- Let's consider a study looking at fruit fly mortality rates (`mortality`) when exposed to a cold treatment for 18 days (`exposure_days`.



Gaussian GLM

Prediction | Code



Issues with Gaussian GLM

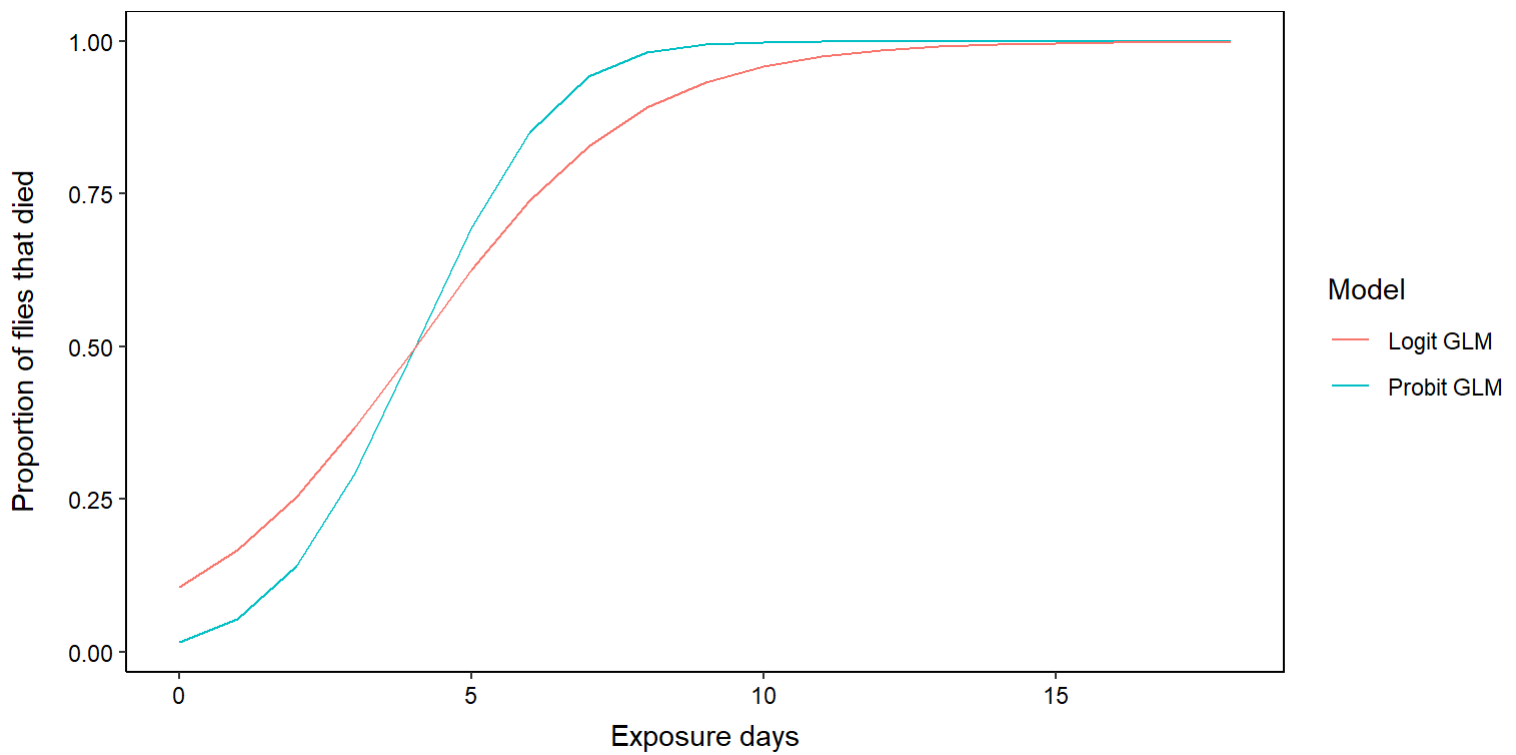
- What are some obvious issues with the Gaussian GLM?
  - It makes predictions outside the [0,1] interval
    - That doesn’t make sense, 1 represents all flies dying and 0 represents no flies dying!!!
  - There clearly isn’t a linear relationship (and there is eivdence for unequal variances!)

Modelling binomial data in R

- There are two basics options for modelling binary/proportion data:
  1. *Logistic GLM*: This is the default parameterisation in almost every field, except agriculture and toxicology, apparently
  2. *Probit GLM*: This seems to be much more popular in agriculture.

Logit vs probit GLM

- Different mathematical formula for calculating the curves
  - The results are usually qualitatively similar, very small differences in predictions



Model syntax

- Assuming our question is: *Does exposure time to cold treatment effect fly mortality?*

1. *Logistic GLM*:

- `mod_logistic <- glm(mortality ~ exposure_days, family = binomial(link = "logit"), data = data)`

2. *Probit GLM*:

- `mod_probit <- glm(mortality ~ exposure_days, family = binomial(link = "probit"), data = data)`

Fit the probit GLM

```r
mod_probit <-
  glm(mortality ~ 1 + exposure_days,
      family = binomial(link = "probit"),
      data = data)
summary(mod_probit)
```

```
Call:
glm(formula = mortality ~ 1 + exposure_days, family = binomial(link = "probit"),
    data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.137208   0.037948  -56.32   <2e-16 ***
exposure_days  0.529826   0.008781   60.34   <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: 15022.18  on 16018  degrees of freedom
Residual deviance:   963.87  on 16017  degrees of freedom
AIC: 3814.8

Number of Fisher Scoring iterations: 9
```

Interpretation

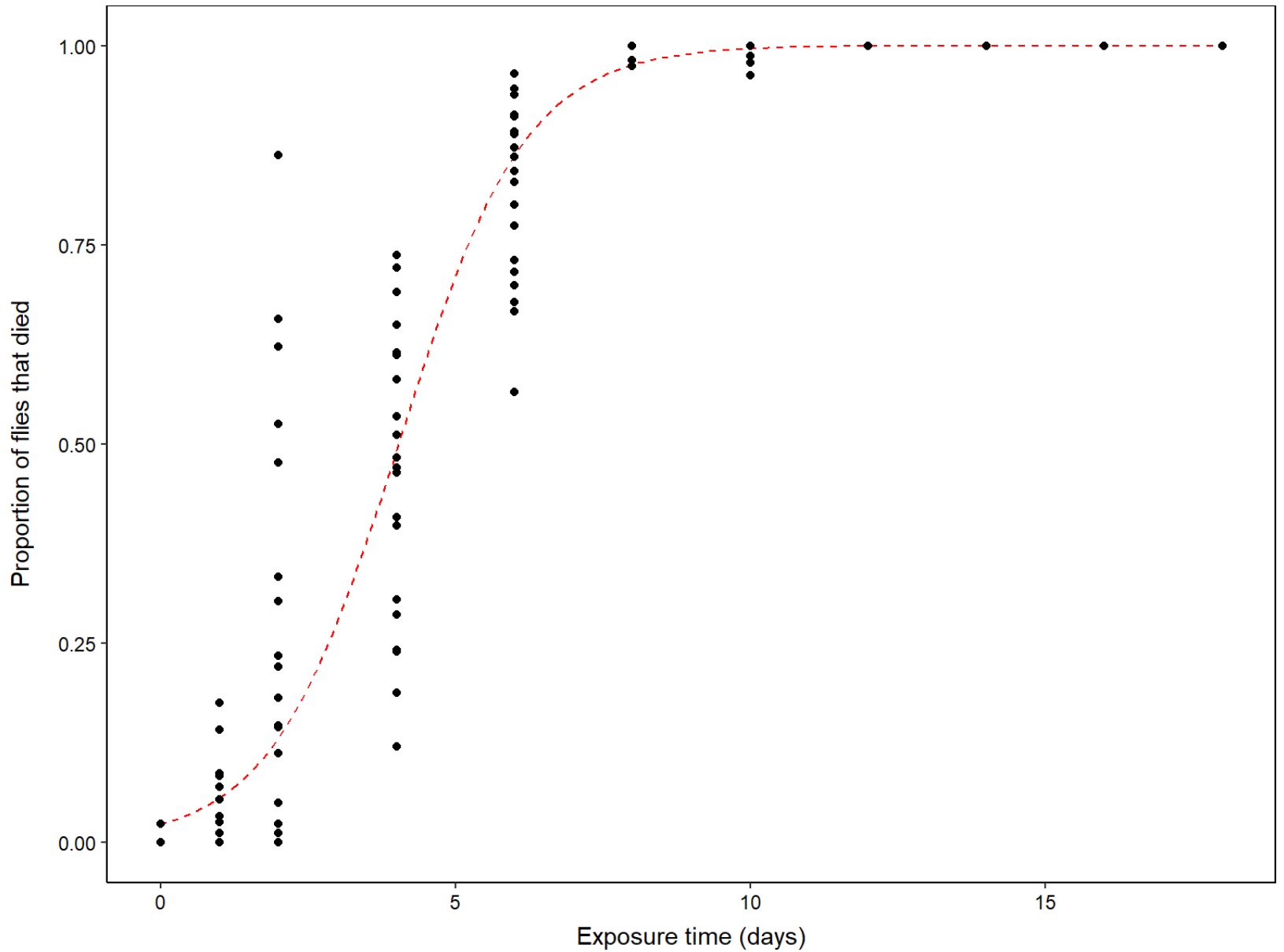
- The interpretation of probit GLM co-efficients is very close to being uninterpretable.
  - The coefficients represent differences in *z-scores* between treatment groups… What on earth is that?
  - Let’s not worry…
  - Instead, let’s fit the more commonly used logistic model

Fit the logit GLM

```r
m_logit <-
  glm(data = data,
      family = binomial(link = "logit"),
      mortality ~ 1 + exposure_days)
summary(m_logit)
```

```
Call:
glm(formula = mortality ~ 1 + exposure_days, family = binomial(link = "logit"),
    data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.75394    0.07517  -49.94   <2e-16 ***
exposure_days  0.93153    0.01740   53.55   <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: 15022  on 16018  degrees of freedom
Residual deviance:  1025  on 16017  degrees of freedom
AIC: 3824.7

Number of Fisher Scoring iterations: 8
```

Plot logit prediction



Likelihood Ratio Test

Test the hypothesis of a `exposure_time` effect on the fly `mortality`

```
# Fit null model
m_null <- glm(
  data = data,
  family = binomial(link = "logit"),
  mortality ~ 1
)

# Perform LRT
lmtest::lrtest(m_null, m_logit)
```

Likelihood Ratio Test

Test the hypothesis of a `exposure_time` effect on the fly `mortality`

```
Likelihood ratio test

Model 1: mortality ~ 1
Model 2: mortality ~ 1 + exposure_days
  #Df  LogLik Df Chisq Pr(>Chisq)
1   1 -9807.6
2   2 -1910.3  1 15794  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is support for a statistically significant relationship between `exposure_time` and fly `mortality` (proportion of flies that died) (X2 = 15794, df = 1, $P < 0.001$).

Interpret coefficients

```
Call:
glm(formula = mortality ~ 1 + exposure_days, family = binomial(link = "logit"),
    data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.75394    0.07517  -49.94   <2e-16 ***
exposure_days  0.93153    0.01740   53.55   <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: 15022  on 16018  degrees of freedom
Residual deviance:  1025  on 16017  degrees of freedom
AIC: 3824.7

Number of Fisher Scoring iterations: 8
```

- (Intercept) = -3.75394
  - Always have to exponentiate (`exp(value)`) to get interpretable coefficients
  - `exp(-3.75394)` = 0.02
  - The expected `mortality` when `exposure_days` = 0 is 0.02.

Interpret coefficients

```
Call:
glm(formula = mortality ~ 1 + exposure_days, family = binomial(link = "logit"),
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.75394    0.07517  -49.94   <2e-16 ***
exposure_days  0.93153    0.01740   53.55   <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: 15022  on 16018  degrees of freedom
Residual deviance:  1025  on 16017  degrees of freedom
AIC: 3824.7

Number of Fisher Scoring iterations: 8
```

- Beta `exposure_days` = 1.998
  - Always have to exponentiate (`exp(value)`) to get interpretable coefficients
  - `exp(0.93153)` = 2.53
  - The proportion of fly `mortality` recorded increases by a factor of 2.53 for each additional day of exposure to cold treatment, on average

Calculating LC values

To find lethal concentration values (e.g. LC50, LC90, LC99 and LC999986), we can use the `MASS::dose_p` function.

```
lc_res <- MASS::dose.p(m_logit,
                       p = c(0.5, 0.9, 0.99, 0.999986))
lc_res
```

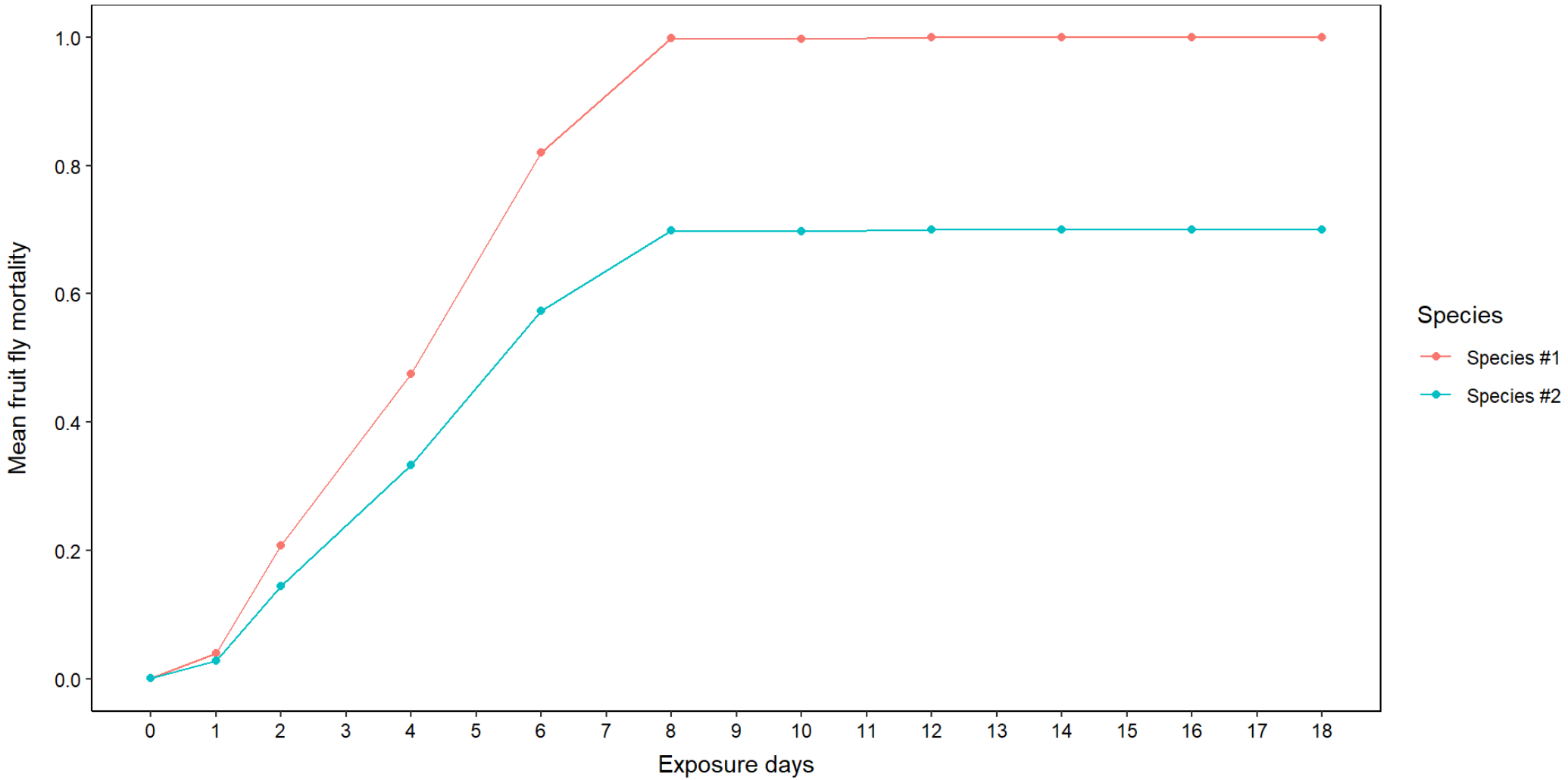
```
                    Dose          SE
p = 0.500000:   4.029861 0.03661887
p = 0.900000:   6.388587 0.05975282
p = 0.990000:   8.962734 0.10213822
p = 0.999986: 16.027797 0.23025544
```

Calculating confidence intervals around LC

```r
# Plug in the LC values you want using `p` argument in `dose_p`
# p = 0.50 = LC50, for example
xp <- MASS::dose.p(m_logit, p = c(0.50, 0.90, 0.99, 0.999986))
xp.ci <- xp + attr(xp, "SE") %*% matrix(qnorm(1-0.05/2)*c(-1,1), nrow = 1)
zp.est <- cbind(xp, attr(xp, "SE"), xp.ci[,1], xp.ci[,2])
dimnames(zp.est)[[2]] <- c("LD", "SE", "LCL", "UCL")
zp.est
```

```
                      LD         SE       LCL       UCL
p = 0.500000:  4.029861 0.03661887  3.958089  4.101632
p = 0.900000:  6.388587 0.05975282  6.271473  6.505700
p = 0.990000:  8.962734 0.10213822  8.762546  9.162921
p = 0.999986: 16.027797 0.23025544 15.576505 16.479089
```

Comparing LC between two treatments



Fit seperate logit GLM per treatment group

Fit a logit GLM to the data for `species #1`

```
m_sp1 <-
  glm(data = data,
      family = binomial(link = "logit"),
      subset = c(species == "Species #1"),
      mortality ~ 1 + exposure_days)
```

Fit seperate logit GLM per treatment group

Fit a logit GLM to the data for `species #2`

```
m_sp2 <-
  glm(data = data,
      family = binomial(link = "logit"),
      subset = c(species == "Species #2"),
      mortality ~ 1 + exposure_days)
```

Lethal ratio test

The ratio test is taken from Wheeler et al. (2006). The test compares LC values from two different probit models. Below, we will compare LC50 between species #1 and species #2

```
ratios <- ecotox::ratio_test(
  model_1 = m_sp1,
  model_2 = m_sp2,
  type = "logit",
  percentage = 50,
  log_x = FALSE)
print(ratios)
```

```
# A tibble: 1 × 7
  compare           percentage dose_1 dose_2    se test_stat p_value
  <chr>                  <dbl>  <dbl>  <dbl> <dbl>     <dbl>   <dbl>
1 Model 1 - Model 2         50   4.02   8.75 0.540      1.44   0.151
```

There is no evidence for a statistically significantly difference in LC50 between species #1 and species #2 ($z = 1.44$, $P = 0.15$).