Generalised Linear Models (GLMs)

http://stirlingcodingclub.github.io/link_functions/GLMs.pdf

Brad Duthie

25 November 2019

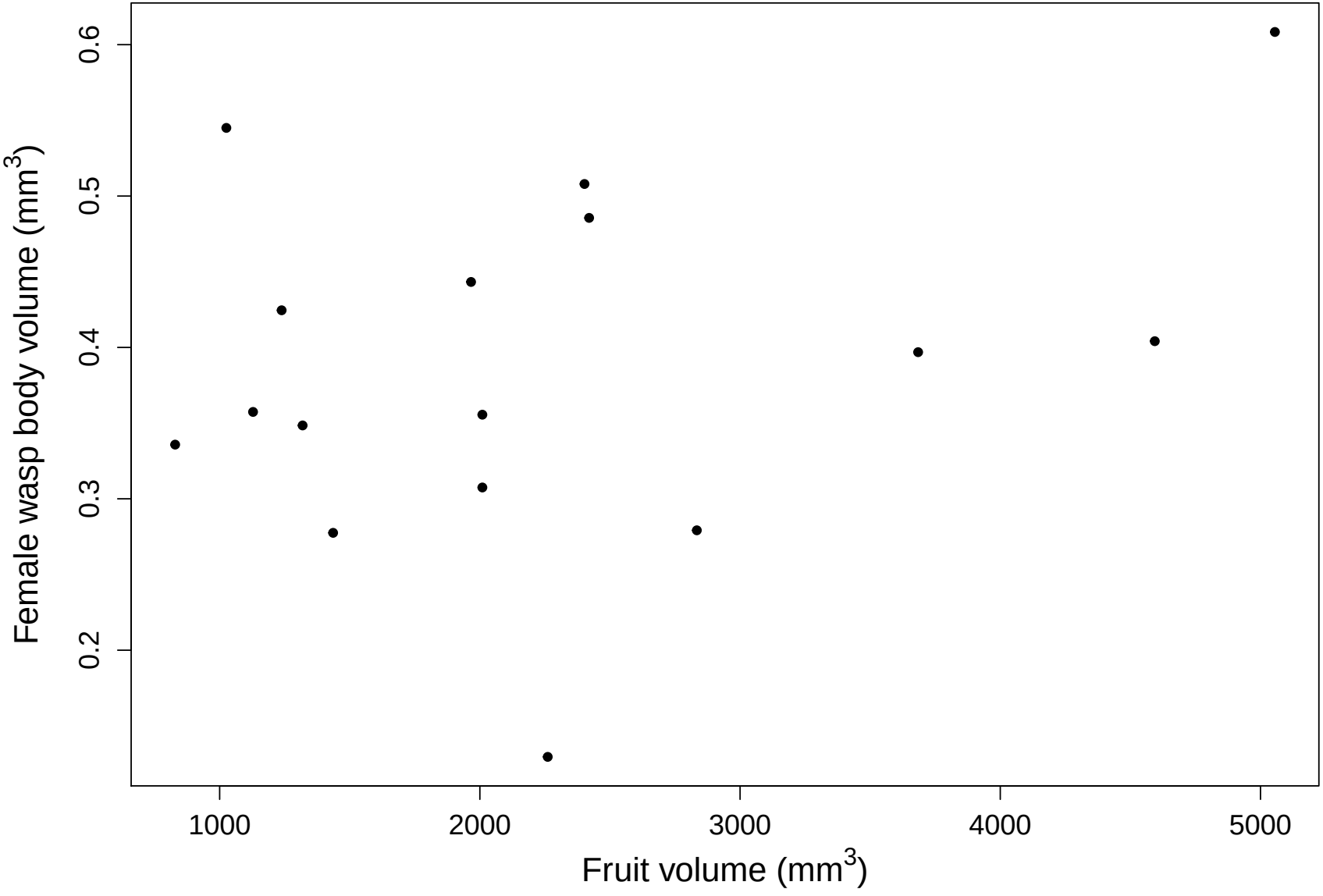
Quick review of simple linear regression

Quick review of simple linear regression

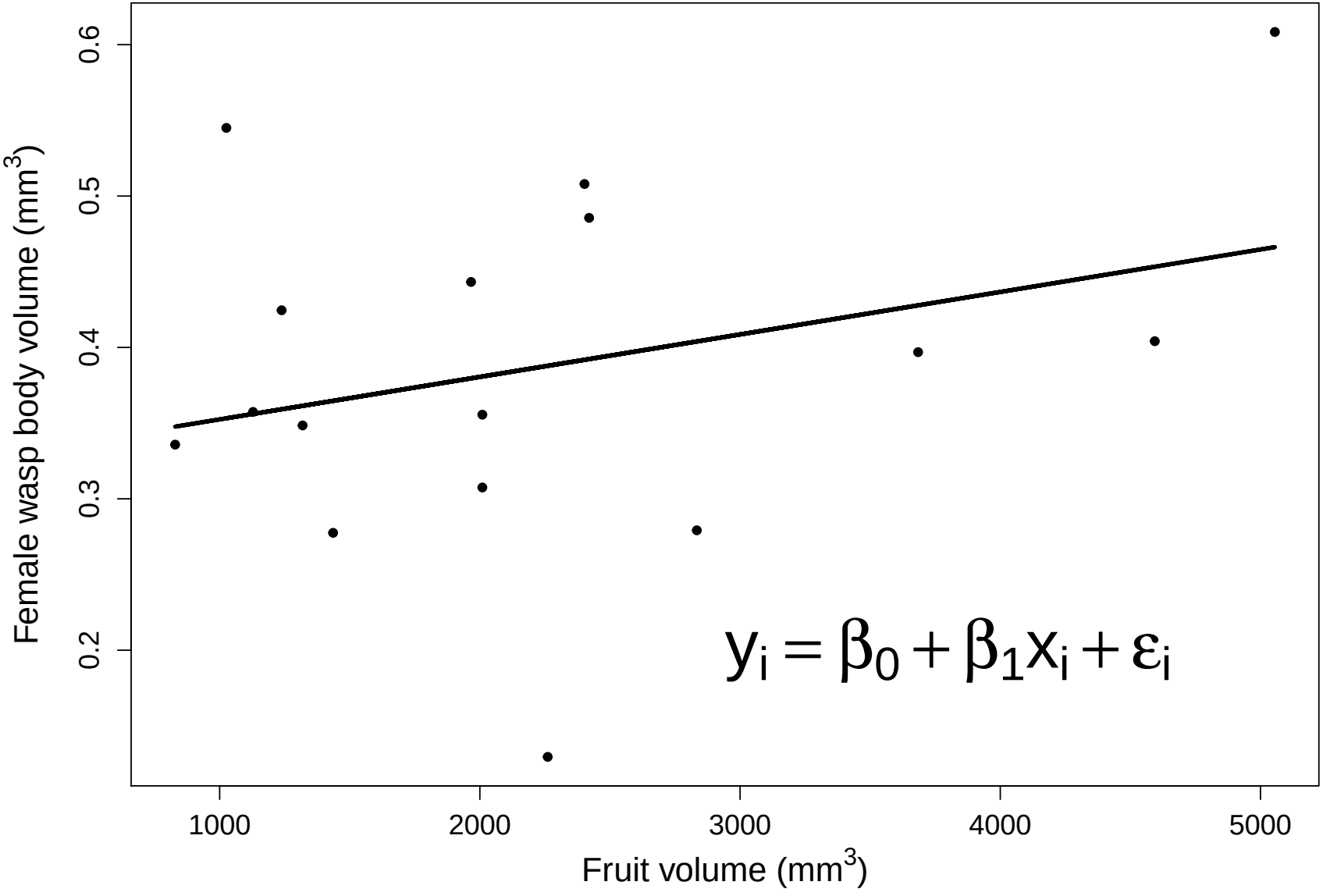
Fig trees in Baja, Mexico are visted by several species of wasps

Wasps use their ovipositors to drill into the side of the enclosed inflorescence (syconia, or colloquially "fruit")

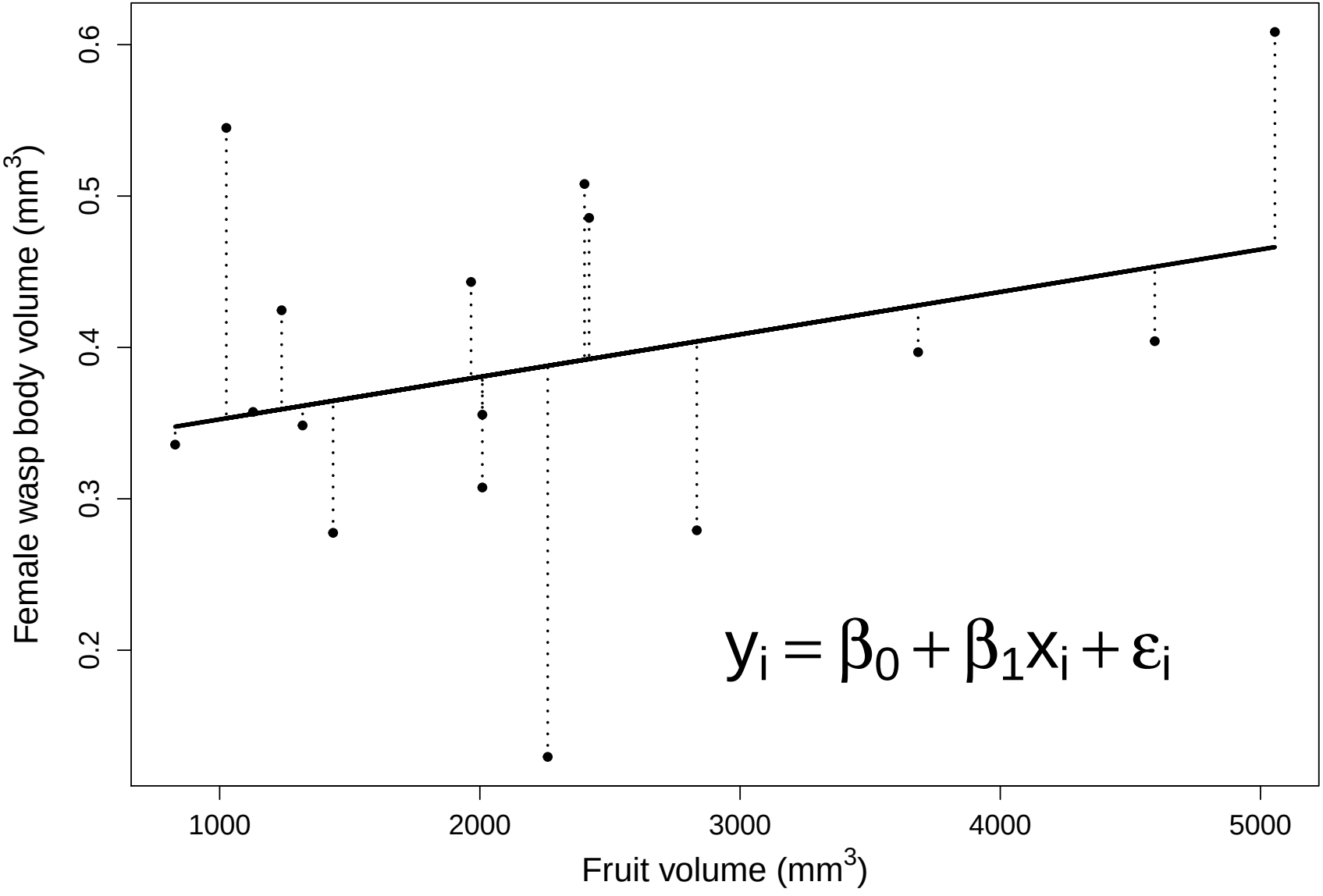
Quick review of simple linear regression



Quick review of simple linear regression



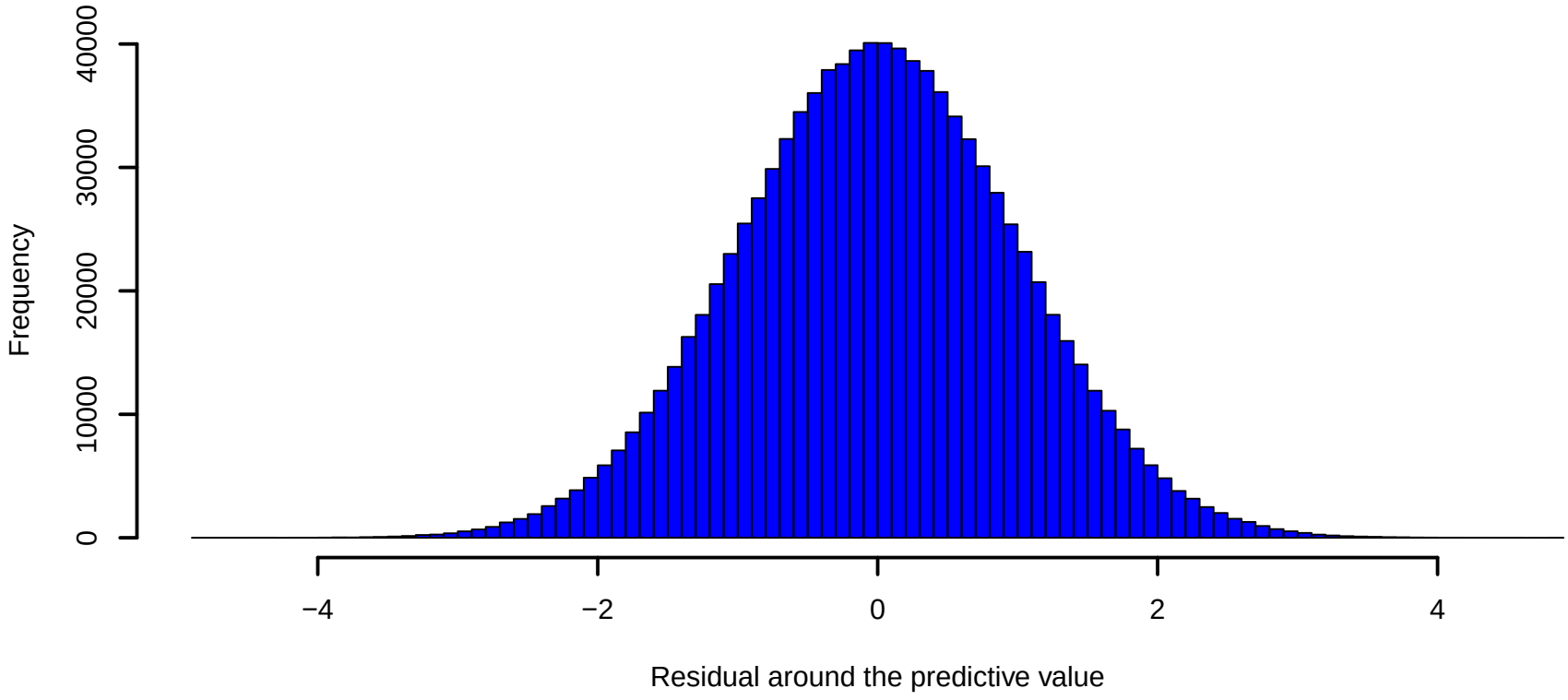
Quick review of simple linear regression



Generalising the linear model: error (i.e., residual) structure

General linear models assume normally distributed errors

Actual errors can violate the asumption of normality in several ways



Strong skew, kurtosis, Strict bounds (e.g., values between 0 and 1 as shown earlier, predicted values never below zero as with counts)

Common problems of general linear models

What if the response (y) variable residuals do not fit general linear model assumptions? This can happen under the following conditions:

- Residuals (ϵ) do not have a constant variance across x values (heteroscadisticity)
- Residuals (ϵ) are not normally distributed

[1] Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons.

Common problems of general linear models

What if the response (y) variable residuals do not fit general linear model assumptions? This can happen under the following conditions:

- Residuals (ϵ) do not have a constant variance across x values (heteroscadisticity)
- Residuals (ϵ) are not normally distributed

Four situations of interest[1]:

1. Count data
2. Proportion data
3. Binary responses
4. "Time to event" data

[1] Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons.

A concrete example: fig wasp survival

- Female *Heterandrium* wasps can produce two types of males
- Winged males disperse from their natal fruit to mate
- Wingless males engage in combat within fruit for access to females

A concrete example: fig wasp survival

- Female *Heterandrium* wasps can produce two types of males
- Winged males disperse from their natal fruit to mate
- Wingless males engage in combat within fruit for access to females
- **Do bigger wingless males have a higher probability of survival?**

Fig wasp survival data

head_width_mm	Survival
624.4652	0
711.9413	1
699.0758	0
779.3522	1
719.3534	1
790.9232	1
650.6125	0
723.1000	1
639.1395	0

Fig wasp survival data

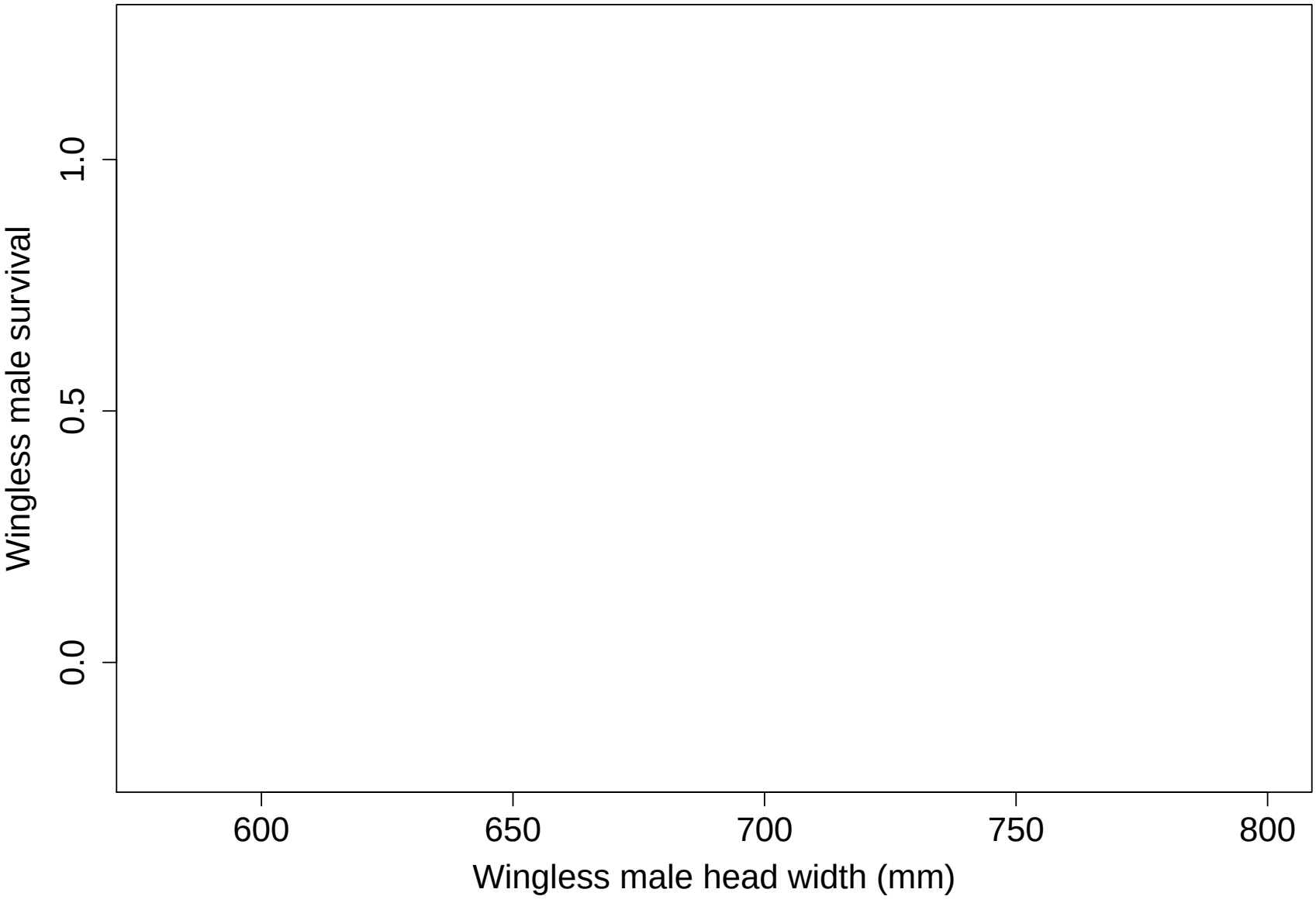


Fig wasp survival data

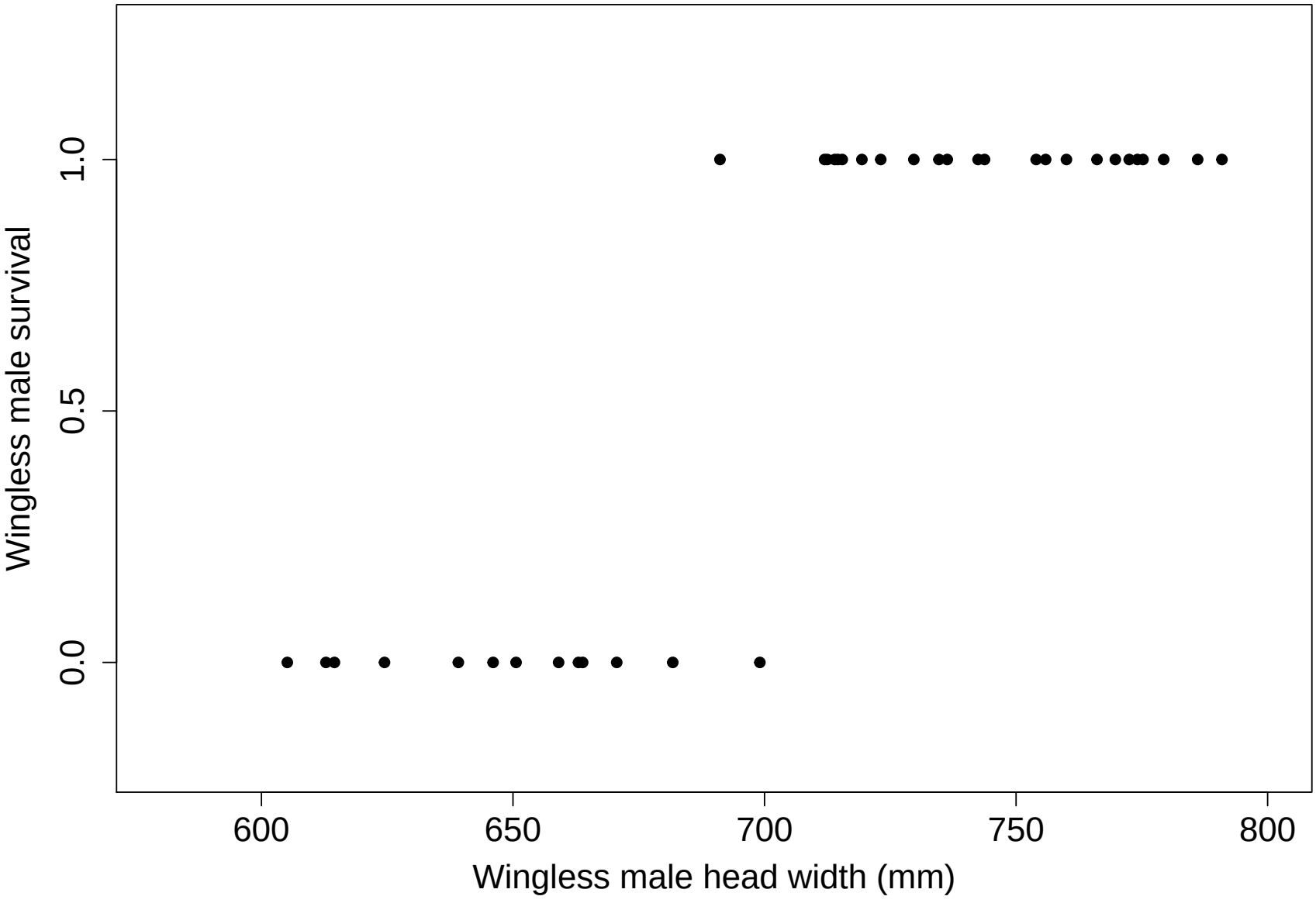


Fig wasp survival data

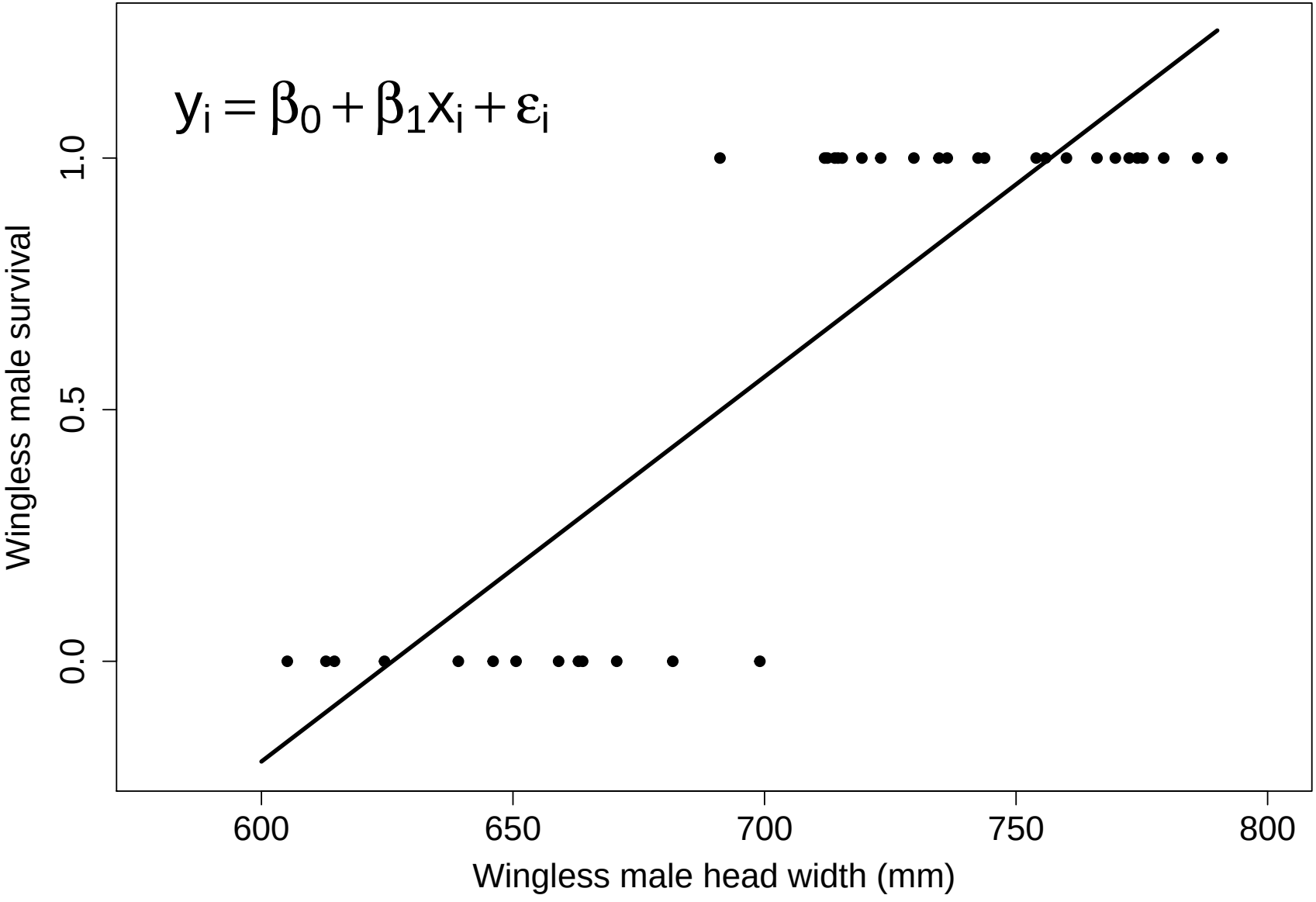
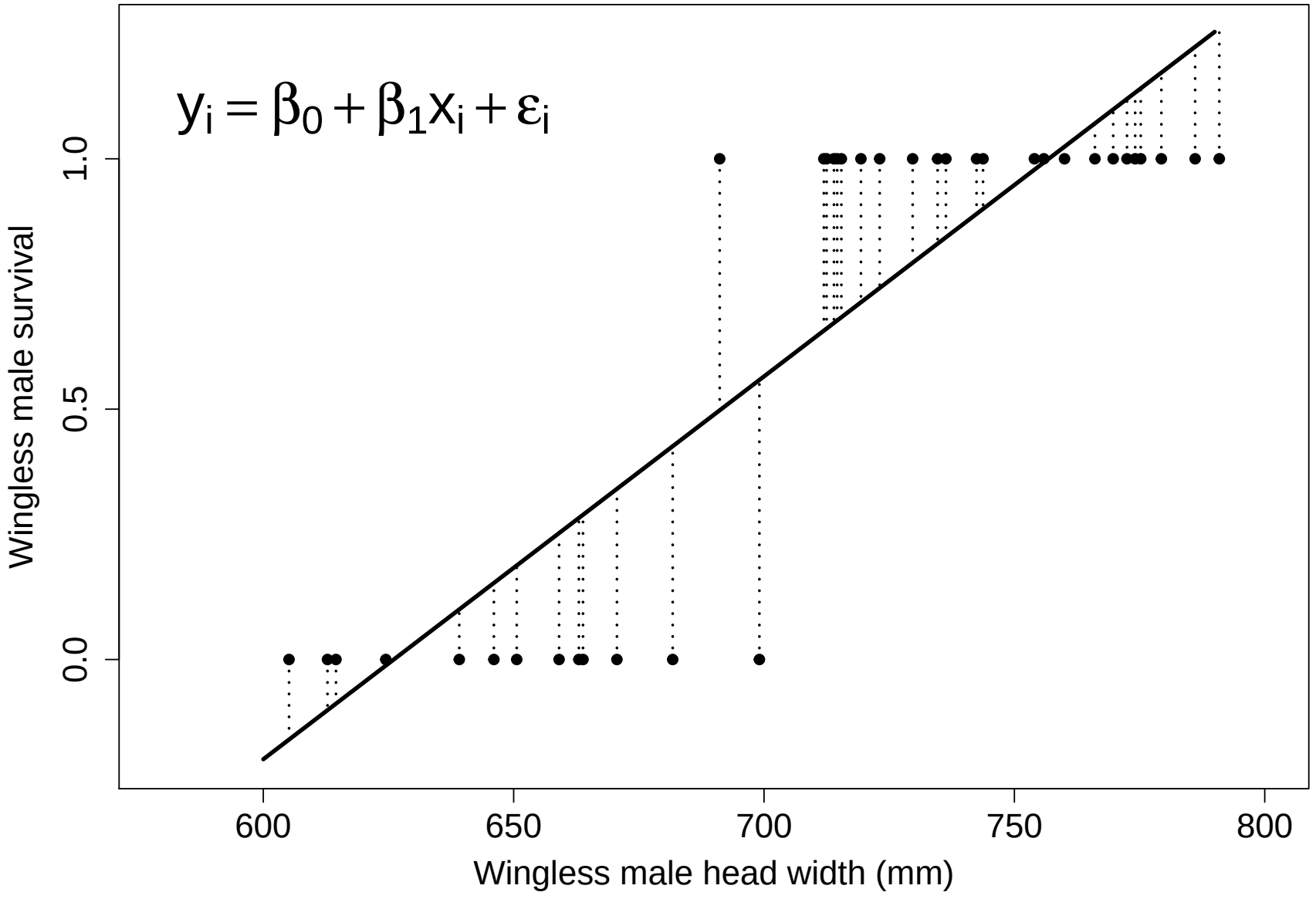
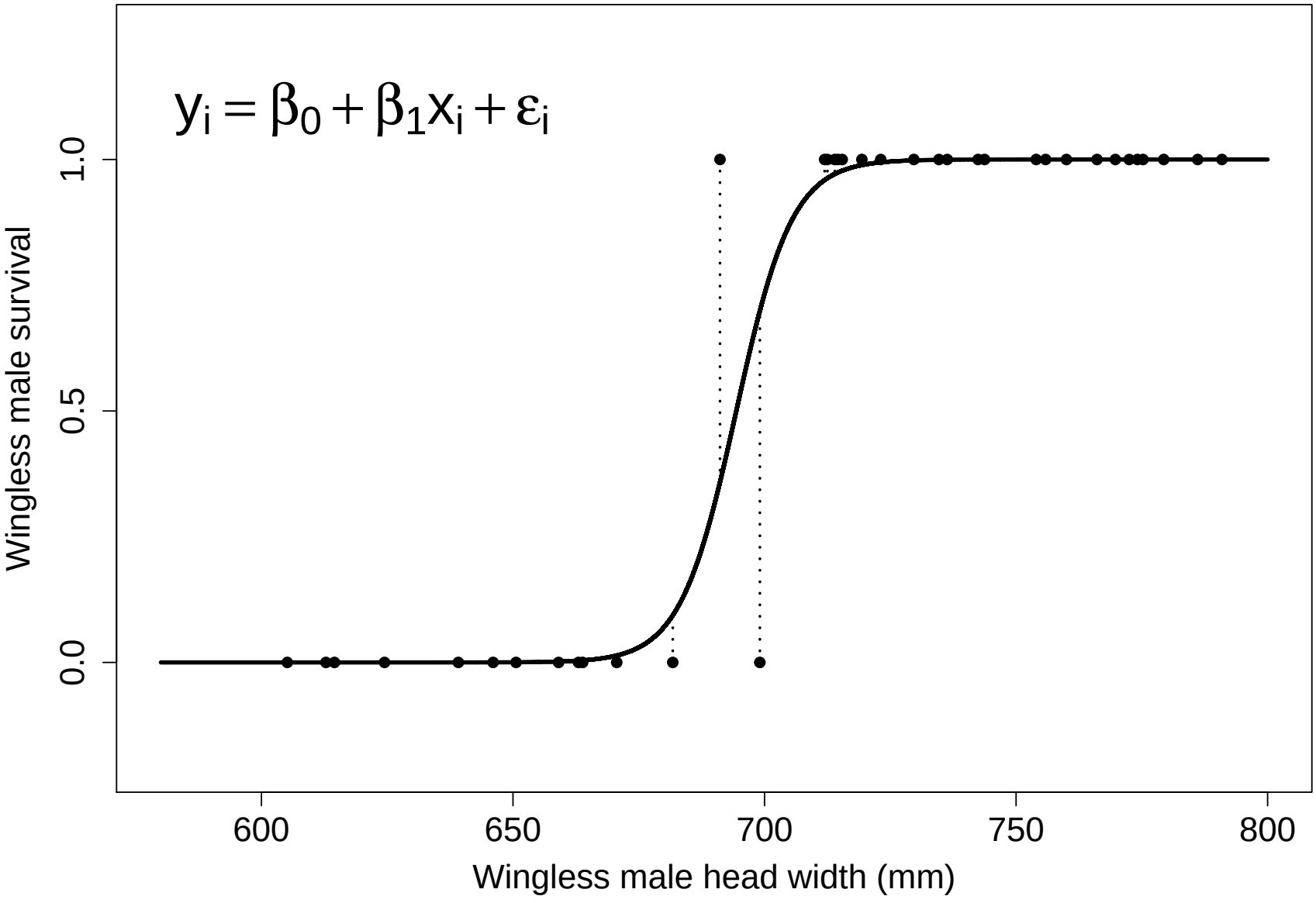


Fig wasp survival data



Using the logit link function



Using the logit link function

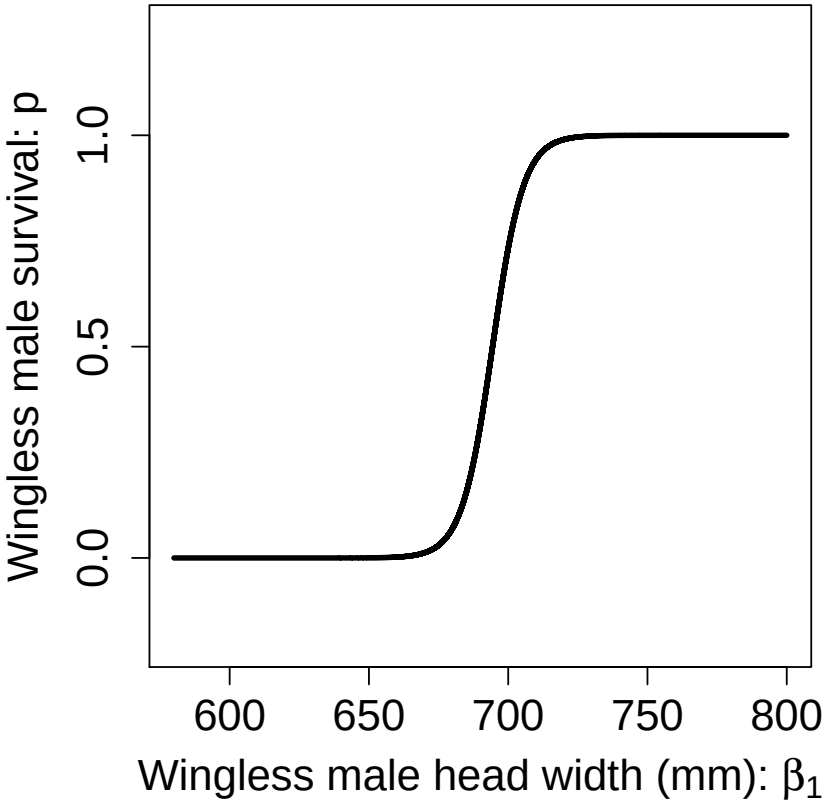
The logit link function linearises the binomial probability function

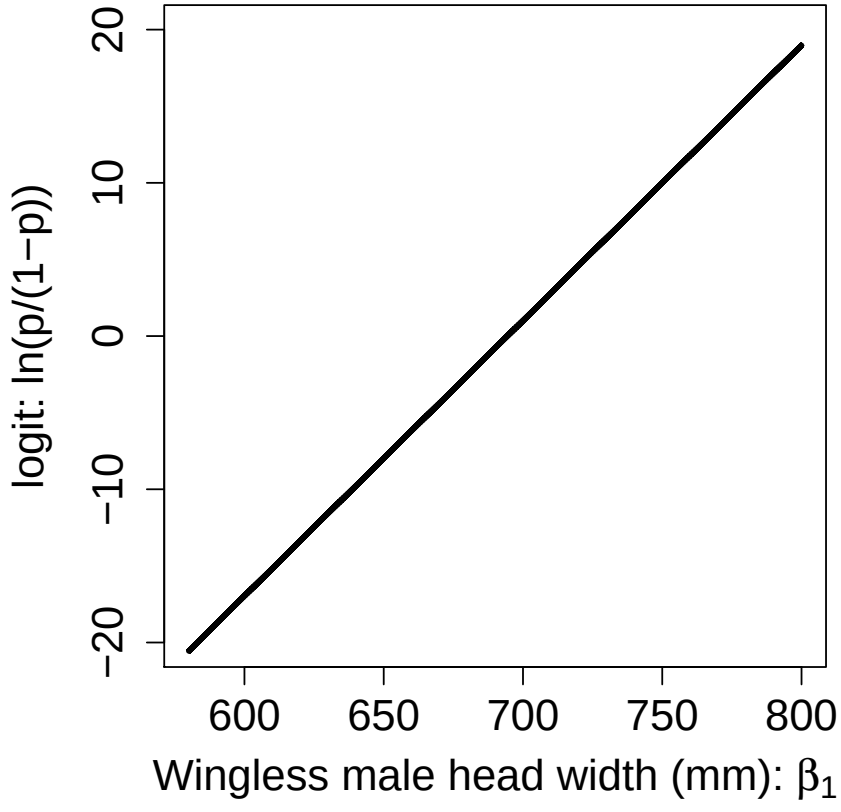
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Using the logit link function

The logit link function linearises the binomial probability function

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$





Using the logit link function

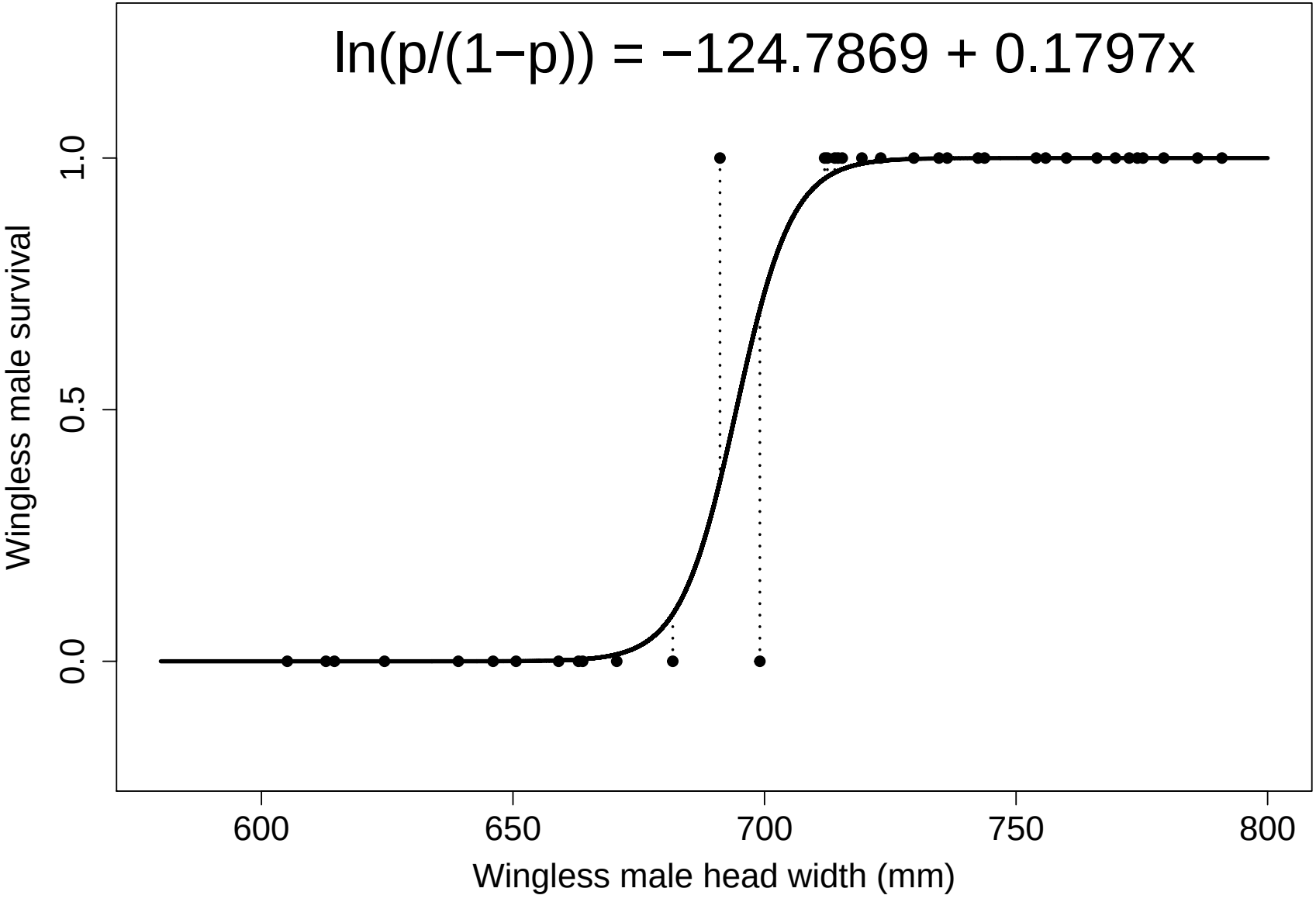
head_width_mm	Survival
624.4652	0
711.9413	1
699.0758	0
779.3522	1
719.3534	1
790.9232	1
650.6125	0
723.1000	1
639.1395	0

Generalised linear model in R

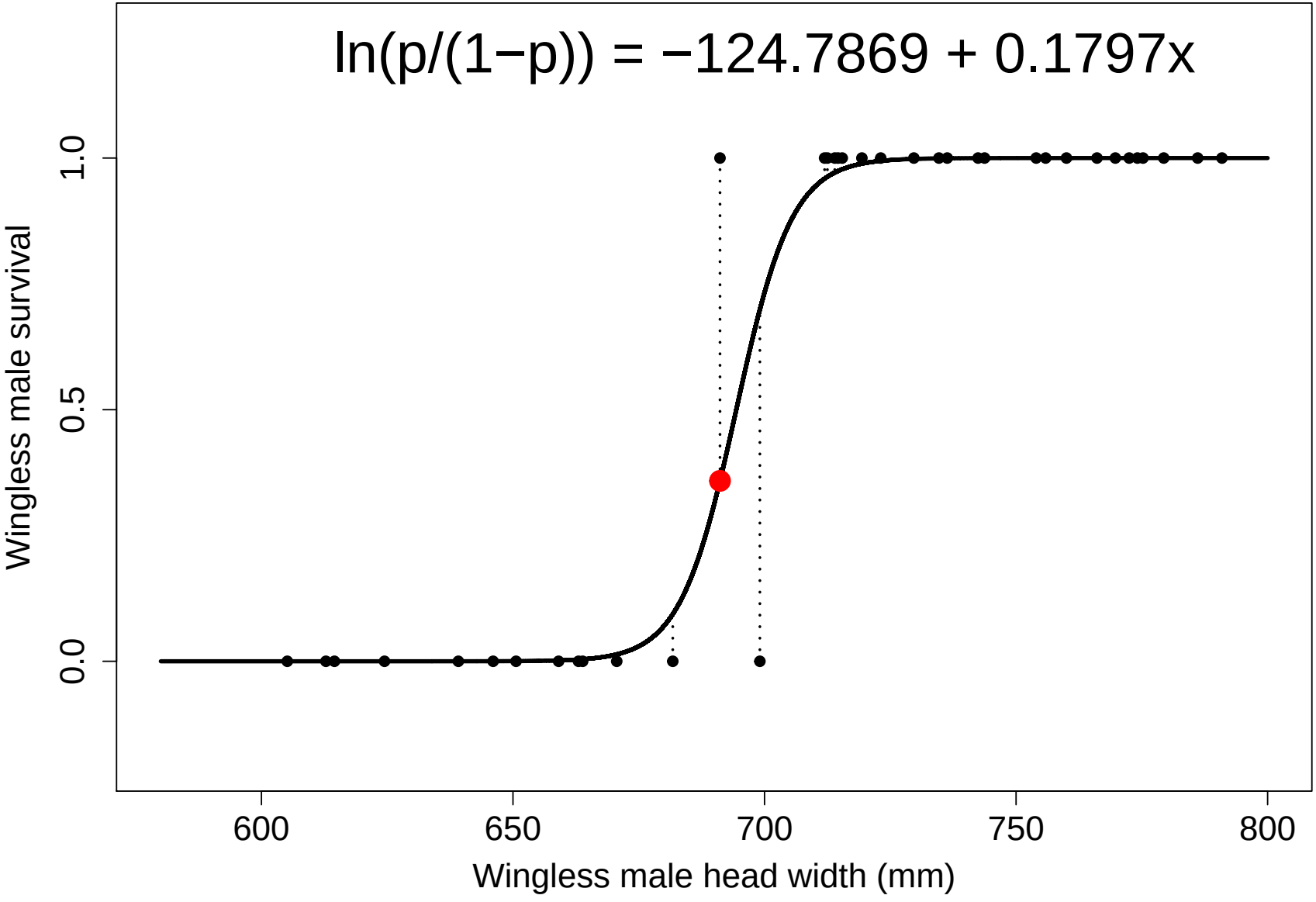
```
dat  <- read.csv("het_heads.csv");
ghus <- glm(formula = Survival ~ head_width_mm,
            family  = binomial(link = "logit"), data = huse);
print(ghus);
```

```
##
## Call:  glm(formula = Survival ~ head_width_mm, family = binomial(lin
##     data = huse)
##
## Coefficients:
##   (Intercept)  head_width_mm
##     -124.7869         0.1797
##
## Degrees of Freedom: 36 Total (i.e. Null);  35 Residual
## Null Deviance:       47.97
## Residual Deviance: 5.053     AIC: 9.053
```

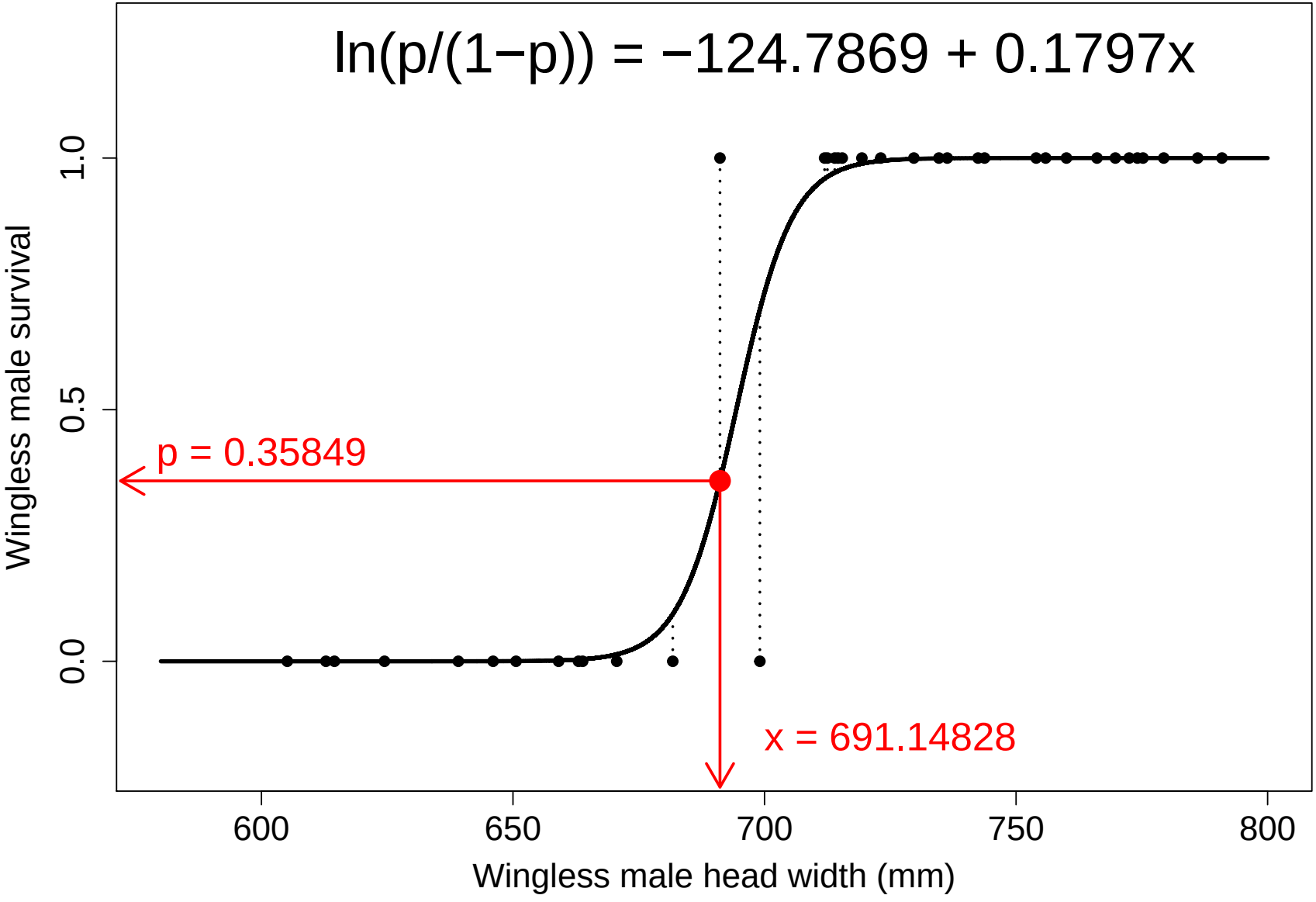
Generalised linear model in R



Generalised linear model in R



What the link function is doing



What the link function is doing

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$ln\left(\frac{p}{1-p}\right) = -124.7869377 + 0.1797082x$$

What the link function is doing

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$ln\left(\frac{p}{1-p}\right) = -124.7869377 + 0.1797082x$$

$$ln\left(\frac{0.3584852}{1-0.3584852}\right) = -124.7869377 + 0.1797082(691.14828)$$

What the link function is doing

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$ln\left(\frac{p}{1-p}\right) = -124.7869377 + 0.1797082x$$

$$ln\left(\frac{0.3584852}{1-0.3584852}\right) = -124.7869377 + 0.1797082(691.14828)$$

$$ln\left(\frac{0.3584852}{1-0.3584852}\right) = -0.5819447$$

What the link function is doing

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$ln\left(\frac{p}{1-p}\right) = -124.7869377 + 0.1797082x$$

$$ln\left(\frac{0.3584852}{1-0.3584852}\right) = -124.7869377 + 0.1797082(691.14828)$$

$$ln\left(\frac{0.3584852}{1-0.3584852}\right) = -0.5819447$$

$$ln\left(0.5588106\right) = -0.5819447$$

What the link function is doing

$$y = -124.7869377 + 0.1797082(700)$$

$$y = 1.0087817$$

$$ln\left(\frac{p}{1-p}\right) = 1.0087817$$

$$\frac{p}{1-p} = e^{1.0087817}$$

$$p = \frac{e^{1.0087817}}{1+e^{1.0087817}} = 0.7327817$$

Generalising the linear model

GLMs: Generalised linear models

- Not to be confused with **general** linear models (also sometimes called GLMs)

- have three properties
  1. Error structure (e.g., binomial)
  2. Linear predictor (e.g., $-124.79 + 0.1797x$)
  3. Link function (e.g., $ln\left(\frac{p}{1-p}\right)$)

Generalising the linear model: error (i.e., residual) structure

Generalised linear models are characterised by independent random variables (i.e., y_1, y_2, ..., y_n) with an expected value $E(y_i) = \mu_i$, and a density function (error) **from the exponential family**.

A density function $f(y_i; \theta_i)$ is in the exponential family if it can be expressed as follows,

$$f(y_i; \theta_i) = e^{y_i\theta_i + b(\theta_i) + c(y_i)}.$$

In the above, θ_i is a parameter of the family.

Statistical distributions in the exponential family include the Poisson, binomial, exponential, and gamma (also the normal).

---

[1] Rencher, AC, & Schaalje, GB. 2008. *Linear models in statistics*. John Wiley & Sons, 446-448.

[2] Nelder, JA, & Wedderburn, RW. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370-384.

Generalising the linear model: error (i.e., residual) structure

There are four common error structures:

1. Poisson errors (for count data)
2. Binomial errors (for proportion data)
3. Exponential errors (for time to event)
4. Gamma errors (for data with constant coefficient of variation)

Generalising the linear model: linear predictor

The linear predictor (η) is the sum of linear effects of 1 or more explanatory variables (β),

$$\eta = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}.$$

GLMs compare a *transformed* value from η to observations:

- The transformation is specified by the link function (see next)
- The fitted value is the predicted value multiplied by the reciprocal of the link function

Generalising the linear model: link function

The link function describes how the expected value of the response variable (μ_i) relates to η,

$$g(\mu_i) = \eta.$$

- Note that this relates the **mean** of a response variable (i.e., $E(y_i) = \mu_i$) to the linear predictor; it is not transforming individual values of y_i.

Generalising the linear model: link function

The link function describes how the expected value of the response variable (μ_i) relates to η,

$$g(\mu_i) = \eta.$$

- Note that this relates the **mean** of a response variable (i.e., $E(y_i) = \mu_i$) to the linear predictor; it is not transforming individual values of y_i.
- The model prediction is not $E(y_i)$, except in the special case of the *identity link* (i.e., $g(\mu_i) = \mu_i = \eta$); i.e., a general linear model.

Linear predictors and link functions

Common GLMs and associated canonical link-distribution pairs.[1]

Model	**Response variable**	**Predictor variable(s)**	**Residual dist.**	**Link**
Linear regression	Continuous	Continuous/ categorical	Gaussian (normal)	Identity $g(\mu) = \mu$
Logistic regression	Binary	Continuous/ categorical	Binomial	Logit $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$
Log-linear models	Counts	Categorical	Poisson	Log $g(\mu) = \ln(\mu)$

[1] Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons.

Linear predictors and link functions

Common GLMs and associated canonical link-distribution pairs.[1]

Model	**R glm argument**
Linear regression	`family = gaussian(link = "identity")`
Logistic regression	`family = binomial(link = "logit")`
Log-linear models	`family = poisson(link = "log")`

```
ghus <- glm(formula = Survival~head_width_mm,
            family = binomial(link = "logit"),
            data = huse);
```

[1] Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons.

Further reading suggestions

- Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons. **(Chapter 17)**
- Crawley, MJ. 2012. *The R book*. John Wiley & Sons. **(Chapters 13, 14, 16)**
- Generalised Linear Mixed Models: http://glmm.wikidot.com
- Rencher, AC, & Schaalje, GB. 2008. *Linear models in statistics*. John Wiley & Sons, 446-448.
- Nelder, JA, & Wedderburn, RW. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370-384. [PDF]
- Generalized Linear Models undertanding the link function. 15 OCT 2018. Bluecology blog.