Logistic Regression

One Predictor

Demetris Athienitis

Section 1

Interpretation

We have already seen the simple logistic regression model

$$\text{logit}\left[\pi(x)\right] = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

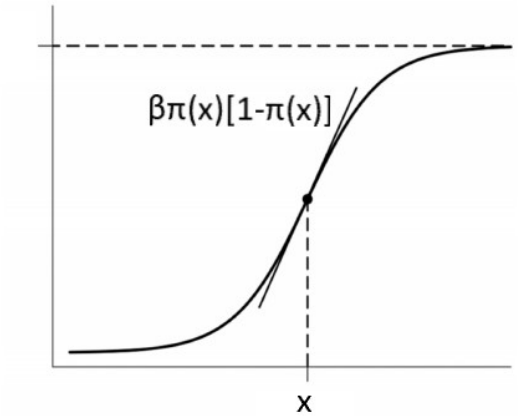
where β represents the strength of association between x and Y

- $\beta > 0$, then $\pi(x) \uparrow$ as $x \uparrow$
- $\beta < 0$, then $\pi(x) \downarrow$ as $x \uparrow$
- $\beta = 0$, then $\pi(x) = e^{\alpha}/(1 + e^{\alpha})$ a constant. $\pi(x) > 0.5$ when $\alpha > 0$

Parameters estimated via MLE are asymptotically normal.

Interpretation

The rate of change in $\pi(x)$ (by taking derivatives) is $\beta\pi(x)\,[1-\pi(x)]$



- Rate of change maximized when $\pi(x) = 0.5$
- This implies max rate of change is $\beta/4$ when $x = -\alpha/\beta$
- This value of $x = -\alpha/\beta$ is called the *median effective level* and it represents the level at which each outcome has a 50% chance

Interpretation

The term e^{β} is the odds ratio for a 1 unit increase in x. The odds of success

- at x

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x}$$

- at $x+1$

$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\alpha+\beta x} e^{\beta}$$

Hence, the odds ratio for $x+1$ versus x is

$$\text{OR} = \frac{\pi(x+1)/\left[1-\pi(x+1)\right]}{\pi(x)/\left[1-\pi(x)\right]} = e^{\beta}$$

Example (Horseshoe crab)

173 female crabs, and wish to model the presence or absence of male "satellites" dependent upon characteristics of the female horseshoe crabs.

$$Y_i = \begin{cases} 1 & \text{satellite present} \\ 0 & \text{otherwise} \end{cases}$$

Explanatory variables are:

- weight (in kg)
- width of shell
- color (medium light, medium, medium dark, dark)
- condition of spine (bad, good, excellent)

Example (continued)

```
> fit=glm(y ~ weight, family=binomial(link=logit))
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74
```

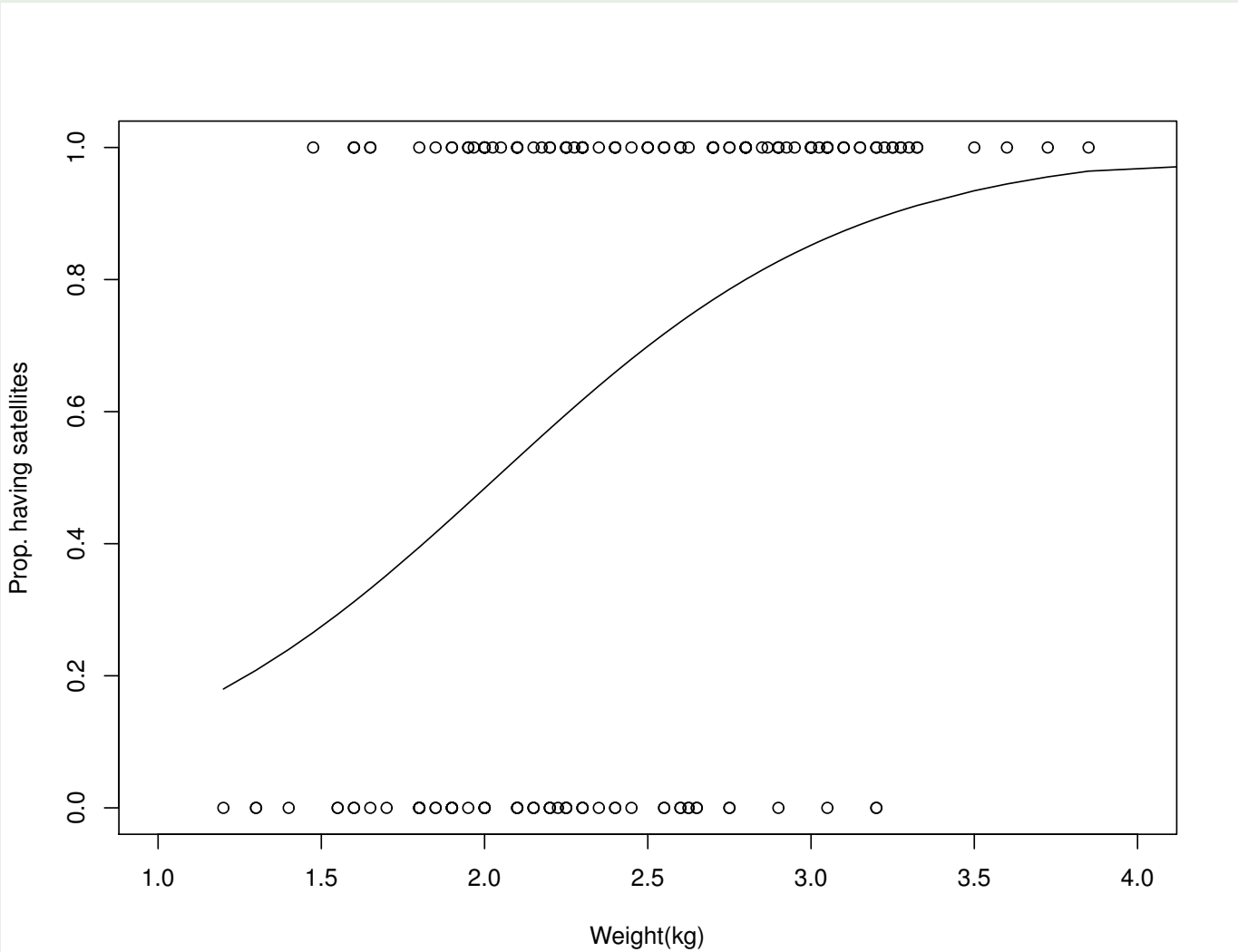
Fit is $\text{logit}[\hat{\pi}(x)] = -3.6947 + 1.8151x$, and β is positive.

$$\hat{\pi}(x) = \frac{\exp(-3.6947 + 1.8151x)}{1 + \exp(-3.6947 + 1.8151x)}$$

Example (continued)

- At the average weight of $x = \bar{x} = 2.44$, $\hat{\pi}(2.44) = 0.676$
- The rate of change at the average $x = 2.44$ is $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = 1.8151(0.676)(0.324) = 0.398$
- $\hat{\pi}(x) = 1/2$ when $x = \frac{-(-3.6947)}{1.8151} = 2.036$
- The estimated change in π per 1 kg increase is about 0.398 (in the neighborhood of the sample mean). However, the standard deviation of weight is $s_x = 0.58$ and hence a 1 unit increase (in kg), may be too much. So, the estimated change in π per 0.1 kg increase is about 0.0398
- For a 1 kg increase in weight, the estimated odds of the presence of a satellite are multiplied by $\exp(1.8151) = 6.14169$. For a 0.1 kg increase, the estimated odds of the presence of a satellite are multiplied by $\exp(0.1(1.8151)) = 1.2$, i.e. the odds increase by 20%.

Example (continued)



Section 2

Framework

Estimates from GLMs are asymptotically normal (due to MLE)

- $\hat{\alpha} \sim N\left(\alpha, \sigma_\alpha^2\right)$
- $\hat{\beta} \sim N\left(\beta, \sigma_\beta^2\right)$
- Jointly, the two estimates follow a multivariate normal distribution

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}\right)$$

The matrix on the right is called the variance-covariance matrix.

CI on β

Hypothesis tests on a specific value, e.g. $H_0 : \beta = \beta_0$ are redundant when we can create an interval of plausible values.

Covered testing and inference on an individual coefficient in previous chapter. The same process will apply for all model coefficients.

We saw that the Wald $100(1-\alpha)\%$ CI on β is

$$\hat{\beta} \mp z_{1-\alpha/2}\left(s_{\hat{\beta}}\right)$$

with estimates and standard errors provided in `R` output. We also used `confint(.)` to obtain LRT CI.

CI on $\pi(x) = \alpha + \beta x$

1. First work with $\text{logit}\left[\hat{\pi}(x)\right] = \hat{\alpha} + \hat{\beta}x$, where

$$\underbrace{\hat{\alpha} + \hat{\beta}x}_{\text{logit}[\hat{\pi}(x)]} \sim N\left(\alpha + \beta x, \sigma_\alpha^2 + x^2\sigma_\beta^2 + 2x\sigma_{\alpha\beta}\right)$$

2. The $100(1-\alpha)\%$ CI for $\text{logit}\left[\pi(x)\right] = \alpha + \beta x$ is

$$\hat{\alpha} + \hat{\beta}x \mp z_{1-\alpha/2}\sqrt{s_\alpha^2 + x^2 s_\beta^2 + 2x s_{\alpha\beta}} \quad \rightarrow \quad (L, U)$$

   where s_α^2 and s_β^2 are the estimated variances, and $s_{\alpha\beta}$ is the estimated covariance. Recall,

$$V\left(\hat{\alpha} + \hat{\beta}x\right) = V(\hat{\alpha}) + x^2 V(\hat{\beta}) + 2x\text{Cov}(\hat{\alpha}, \hat{\beta})$$

3. The $100(1-\alpha)\%$ CI for $\pi(x)$, is then

$$\left(\frac{e^L}{1+e^L}, \frac{e^U}{1+e^U}\right)$$

CI on $\pi(x) = \alpha + \beta x$

In R

- The variance-covariance matrix for all parameters can be found for `glm` objects by using `vcov(model)`
- The estimate and the standard error for $\text{logit}\left[\hat{\pi}(x)\right] = \hat{\alpha} + \hat{\beta}x$ can obtained using

```
predict.glm(model,newdata,type="link",...)
```

Remark

We looked at CI for $\alpha + \beta x$, a linear combination of two parameters but this method can be extended to linear combinations of any length of parameters.

Example (Horseshoe crab continued)

Test $H_0 : \beta = 0$ via

- Wald test given in the summary output (and CI could be derived)
- Likelihood ratio test G^2 and preferably corresponding CI

```
> confint(fit,"weight")
    2.5 %    97.5 %
1.113790 2.597305
```

Hence we are confident that β is at least 1.124 and at most 2.597.

Example (Horseshoe crab continued)

There are 6 female crabs with a weight of 2.4 kg (or 2400 g), of whom only 4 have at least one satellite. Using the model we construct a 95% CI for $\hat{\pi}(2.4)$, by first constructing the CI for logit$[\hat{\pi}(2.4)]$

```
> eta=predict(fit,newdata=data.frame(weight=2.4),
+  type="link",se.fit=TRUE)
> eta
$fit
[1] 0.6616206
$se.fit
[1] 0.1780615
```

Note that the standard error is the same as

```
> sqrt(vcov(fit)[1,1]+2.4^2*vcov(fit)[2,2]+
  2*2.4*vcov(fit)[1,2])
[1] 0.1780615
```

Example (Horseshoe crab continued)

```
> eta.C.I.=eta$fit+c(-1,1)*qnorm(0.975)*eta$se.fit
> eta.C.I. # This is (l,u) interval
[1] 0.3126265 1.0106148

# This is (exp(l)/(1+exp(l)),exp(u)/(1+exp(u)))
> plogis(eta.C.I.)
[1] 0.5775262 0.7331404
```

We learned

- Interpretation of model and model parameters
- Inference on individual parameters (continued from previous chapter)
- Inference on linear combinations of parameters