Logistic Regression

Nov 21, 2019

Example: Birthweight and bronchopulmonary dysplasia

Can we estimate probability of bronchopulmonary dysplasia (BPD, a lung disease that affects newborns) as a function of the baby's birth weight? (Note: our source says the unit of measurement is a 'gramme', but the numbers seem impossibly small. Possibly all babies in our sample are premature? Or maybe there is something weird about the measurements.)

Data from Pagano, M. and Gauvreau, K. (1993). *Principles of Biostatistics*. Duxbury Press.

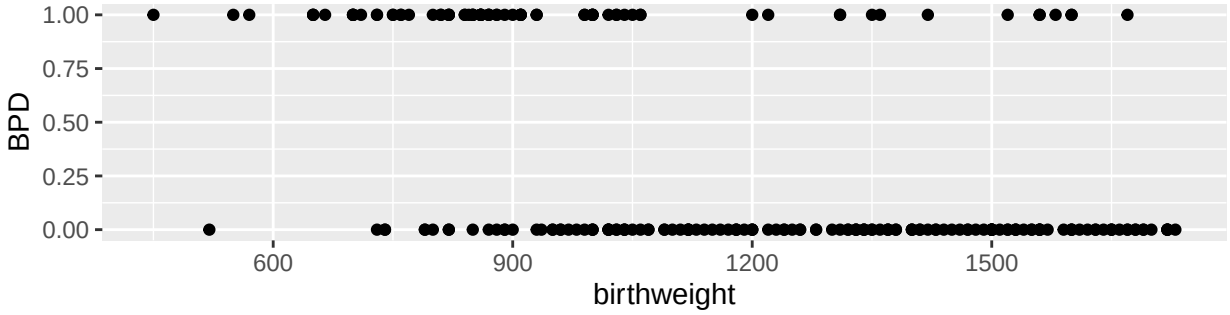
$$Y_i = \begin{cases} 1 & \text{if baby number } i \text{ has BPD} \\ 0 & \text{otherwise} \end{cases}$$

$$X_i = \text{ birth weight for baby number } i$$

```
head(bpd)
```

```
## # A tibble: 6 x 2
##   birthweight   BPD
##         <dbl> <dbl>
## 1         850     1
## 2        1500     0
## 3        1360     1
## 4         960     0
## 5        1560     0
## 6        1120     0
```

Here's a plot of the data:



Fitting model to data

```
logistic_fit <- glm(BPD ~ birthweight, data = bpd, family = binomial)
summary(logistic_fit)
```

```
##
## Call:
## glm(formula = BPD ~ birthweight, family = binomial, data = bpd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9916  -0.7993  -0.4096   0.9242   2.4802
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.0342913  0.6957121   5.799 6.68e-09 ***
## birthweight -0.0042291  0.0006408  -6.600 4.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 286.14  on 222  degrees of freedom
## Residual deviance: 223.72  on 221  degrees of freedom
## AIC: 227.72
##
## Number of Fisher Scoring iterations: 4
```

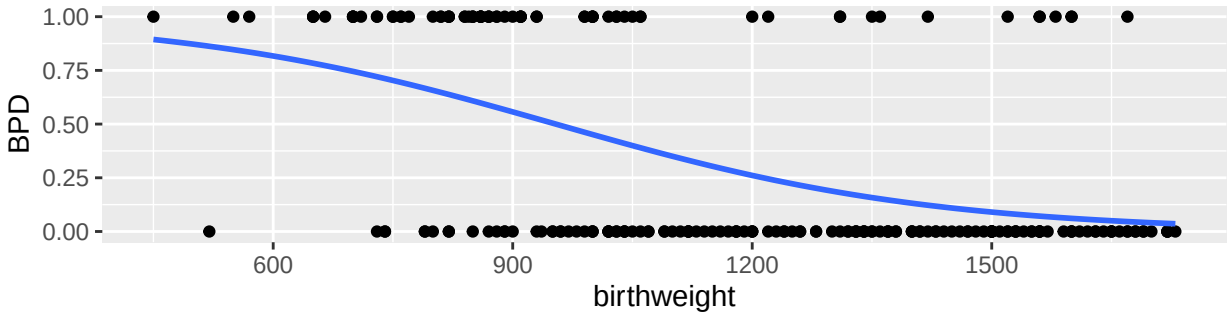
```
exp(-0.0042291)
```

```
## [1] 0.9957798
```

Here's a plot of the data and model fit

- The curve represents the estimated probability that a baby has BPD, as a function of birth weight.

```
ggplot(data = bpd, mapping = aes(x = birthweight, y = BPD)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE)
```



1. Based on this model, how could you calculate the estimated probability that a baby whose birth weight is 750 grams has BPD?

```
exp(4.0342913 - 0.0042291 * 750) / (1 + exp(4.0342913 - 0.0042291 * 750))
```

```
## [1] 0.7031757
```

. . . or. . .

```
predict_data <- data.frame(
  birthweight = 750
)
predict(logistic_fit, newdata = predict_data, type = "response")
```

```
##         1
## 0.7031695
```

2. What is the interpretation of $\hat{\beta}_1$ in terms of odds?

3. What is the estimated relationship between the odds that a baby with birth weight 1000 grams has BPD vs that a baby with birth weight 2000 grams has BPD?

```
exp(-0.0042291 * 1000)
```

```
## [1] 0.01456549
```