Estimation for Logistic Regression Models

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?

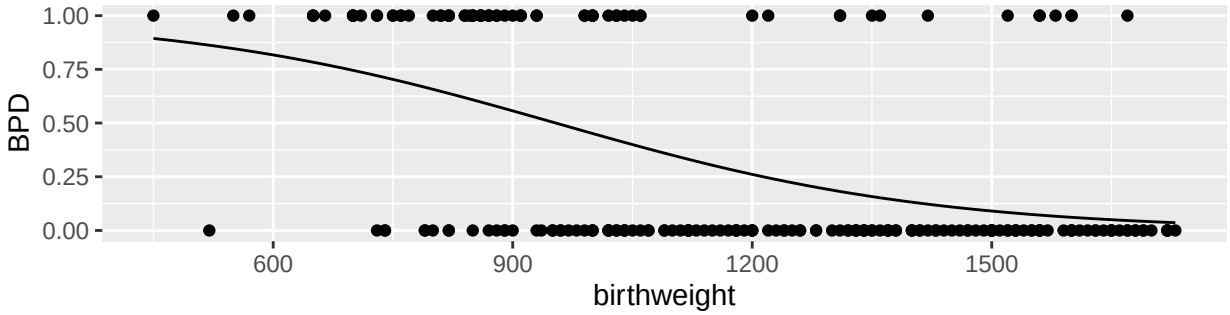
Data from Pagano, M. and Gauvreau, K. (1993). *Principles of Biostatistics*. Duxbury Press.

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

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

```
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
```



The parameter estimates for our model fit are $\hat{\beta}_0 = 4.03429128$ and $\hat{\beta}_1 = -0.00422914$.

Joint Probability of Observed Data

For a fixed value of β_0 and β_1, the probability assigned to the observed data $y_1, \ldots, y_n$ is:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | x_1, \ldots, x_n) = P(Y_1 = y_1|x_1)P(Y_2 = y_2|x_2) \cdots P(Y_n = y_n|x_n)$$
$$= \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_i}}$$

Based on the parameter estimates for our model ($\hat{\beta}_0 = 4.03429128$ and $\hat{\beta}_1 = -0.00422914$), the joint probability assigned to the data is:

```
bpd_augmented <- bpd %>%
  mutate(
    est_prob_Y_eq_1 =
      exp(4.03429128 - 0.00422914 * birthweight) / (1 + exp(4.03429128 - 0.00422914 * birthweight)),
    est_prob_Y_eq_y = ifelse(BPD == 1, est_prob_Y_eq_1, 1 - est_prob_Y_eq_1)
  )
```

```r
head(bpd_augmented, 3)
```

```
## # A tibble: 3 x 4
##   birthweight   BPD est_prob_Y_eq_1 est_prob_Y_eq_y
##         <dbl> <dbl>           <dbl>           <dbl>
## 1         850     1          0.608           0.608
## 2        1500     0          0.0903          0.910
## 3        1360     1          0.152           0.152
```

```r
nrow(bpd_augmented)
```

```
## [1] 223
```

```r
prod(bpd_augmented$est_prob_Y_eq_y)
```

```
## [1] 2.628358e-49
```

Maximum likelihood estimation

The best choice of β_0 and β_1 assigns highest probability to the observed data.

$\max_{\beta_0,\beta_1}$ Likelihood$(\beta_0, \beta_1) = \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_i}}$