

# Introduction to Generalised Linear Models

Advanced Data Analysis

Matt Wand

1

## Example 1

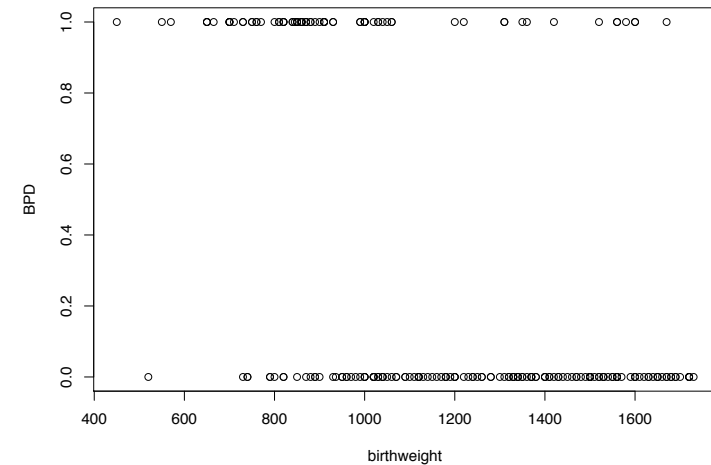
223 measurements available on birthweight and BPD where BPD is an indicator of baby having bronchopulmonary dysplasia (BPD)

3

## GLM versus LM

	response variable	examples
Linear Models (LM)	continuous, close to normal	weight, IQ
Generalised Linear Models (GLM)	binary; count; skewed continuous	presence of disease, number of loan defaults

2



4

## Regression Models

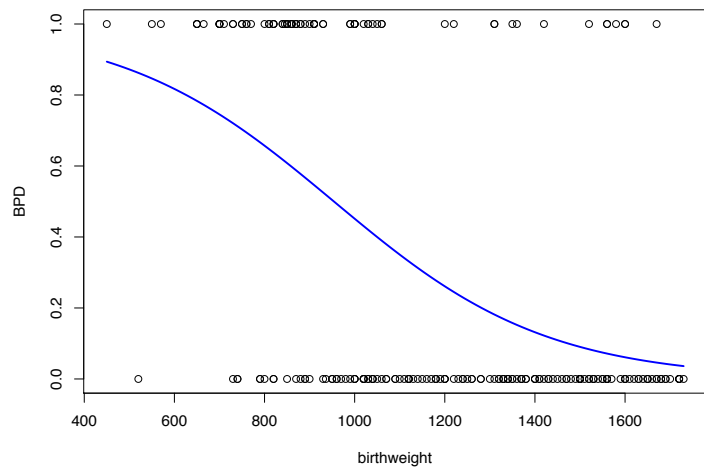
$$\text{BPD}_i = \beta_0 + \beta_1 \text{birthweight}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

is not appropriate.

Improvement is:

$$P(\text{BPD}_i = 1 | \text{birthweight}) = F(\beta_0 + \beta_1 \text{birthweight}_i)$$

where  $F: \mathbb{R} \rightarrow (0, 1)$ .



5

## R Regression Output

```
Call:
glm(formula = BPD ~ birthweight, family = "binomial")
```

```
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
```

7

6

8

## Residual Analysis

Any good regression analysis should examine the residuals.

However, residual graphics is more delicate in GLM than LM.

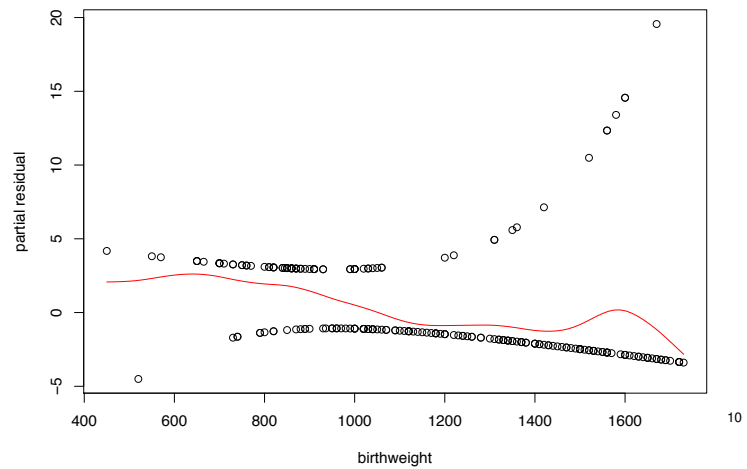
## Comments of Previous Graphic

The red curve is roughly linear, so no strong departure from linearity assumption inherent in

$$P(\text{BPD}_i = 1 | \text{birthweight}) = F(\beta_0 + \beta_1 \text{birthweight}_i)$$

9

11

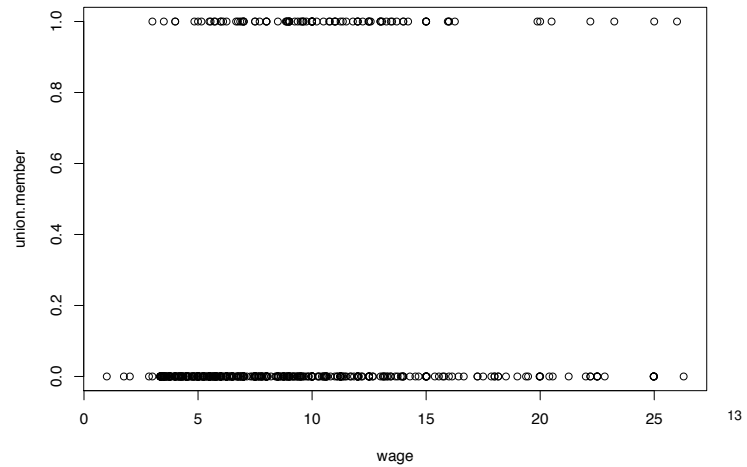


## Example 2

533 measurements available on  
wage and trade union membership

12

## R Regression Output

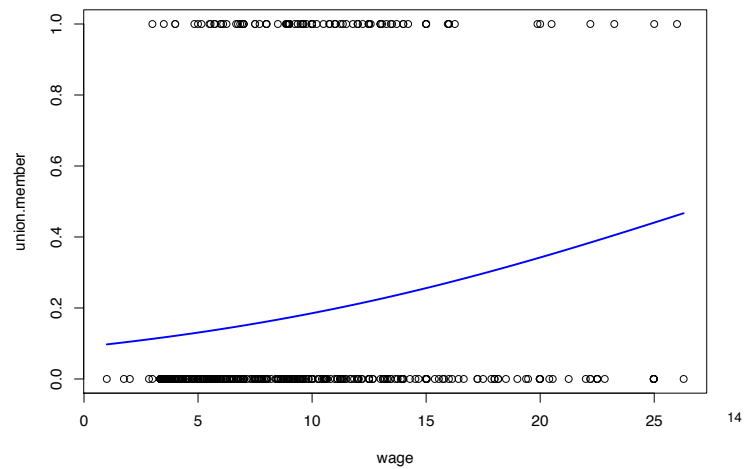


```
Call:
glm(formula = union.member ~ wage, family = "binomial")
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1217 -0.6341 -0.5500 -0.5022  2.0884
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.30923     0.24221  -9.534 < 2e-16 ***
wage          0.08280     0.02098   3.947  7.9e-05 ***
---
```

15



```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 502.69  on 532  degrees of freedom
Residual deviance: 487.61  on 531  degrees of freedom
AIC: 491.61
```

```
Number of Fisher Scoring iterations: 4
```

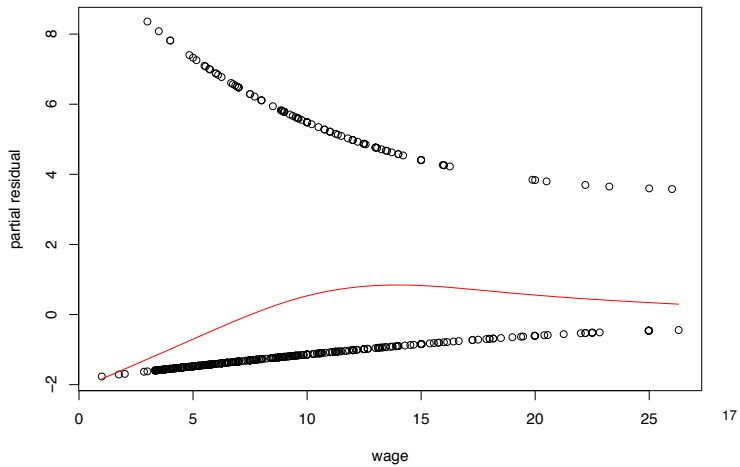
16

### Example 3

365 measurements available on  
daily number of respiratory deaths

and total suspended particles (TSP)

in Milan, Italy, for the year 1989.

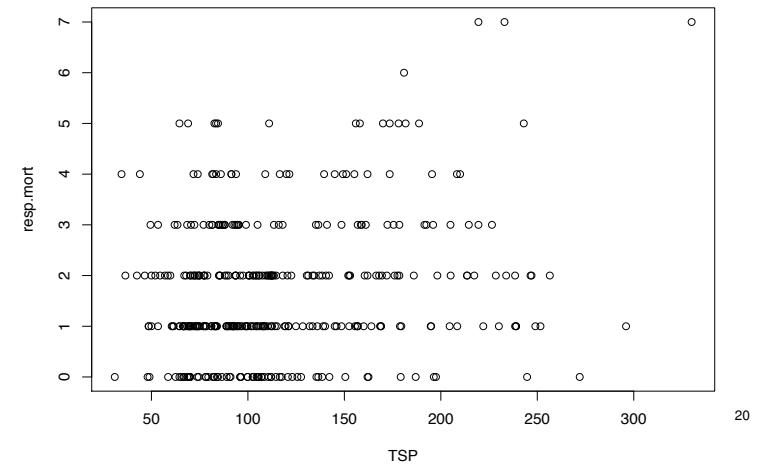


19

### Comments of Previous Graphic

The red curve has pronounced non-linear (quadratic) shape so linearity assumption is probably not reasonable.

18



20

## Count Data

In this example the response variable is a count:

i.e. response  $\in \{0, 1, 2, \dots\}$ . ; so again

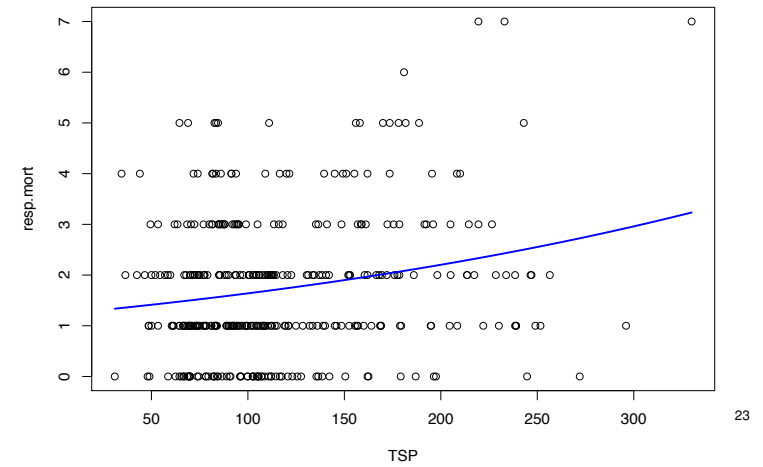
$$\text{respMort}_i = \beta_0 + \beta_1 \text{TSP}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

is not appropriate.

Improvement is:

$$\text{respMort}_i \sim \text{Poisson}[\exp(\beta_0 + \beta_1 \text{TSP}_i)].$$

21



22