Logistic regression

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

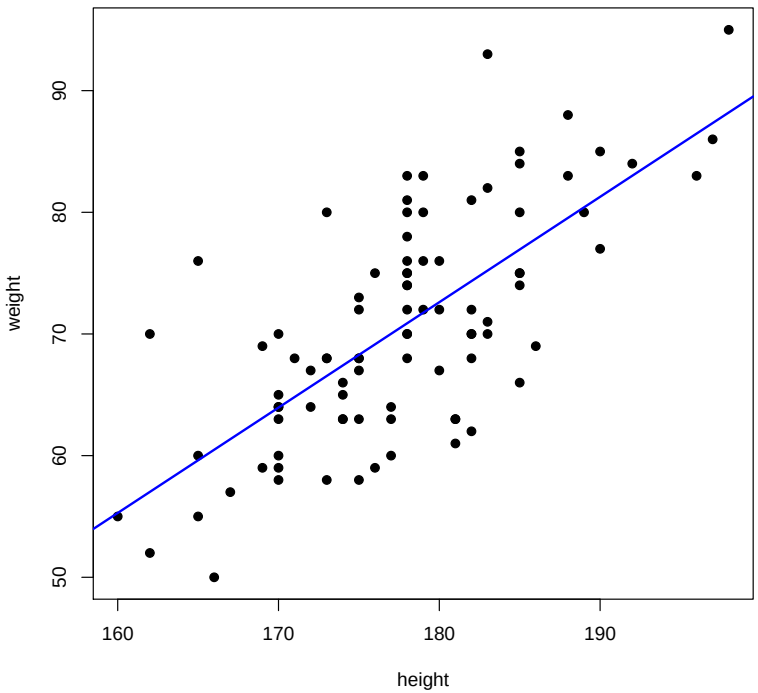
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

October 13, 2022

Contents

Classical linear regression



Theoretical model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Usual assumptions:

- $E(\varepsilon_i) = 0$.
- $V(\varepsilon_i) = \sigma^2$ (constant variance).
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ (independent observations).
- $\varepsilon_i \sim N(0, \sigma^2)$.

Summarized:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Binary response

What if the response is a binary variable?

Examples:

- Explaining disease (0/1) on the basis of biomarkers.
- Explaining admission (0/1) to a university or college using demographic variables and high school grades.
- Explaining low birth weight (0/1) on the basis of characteristics of the mother.
- ....

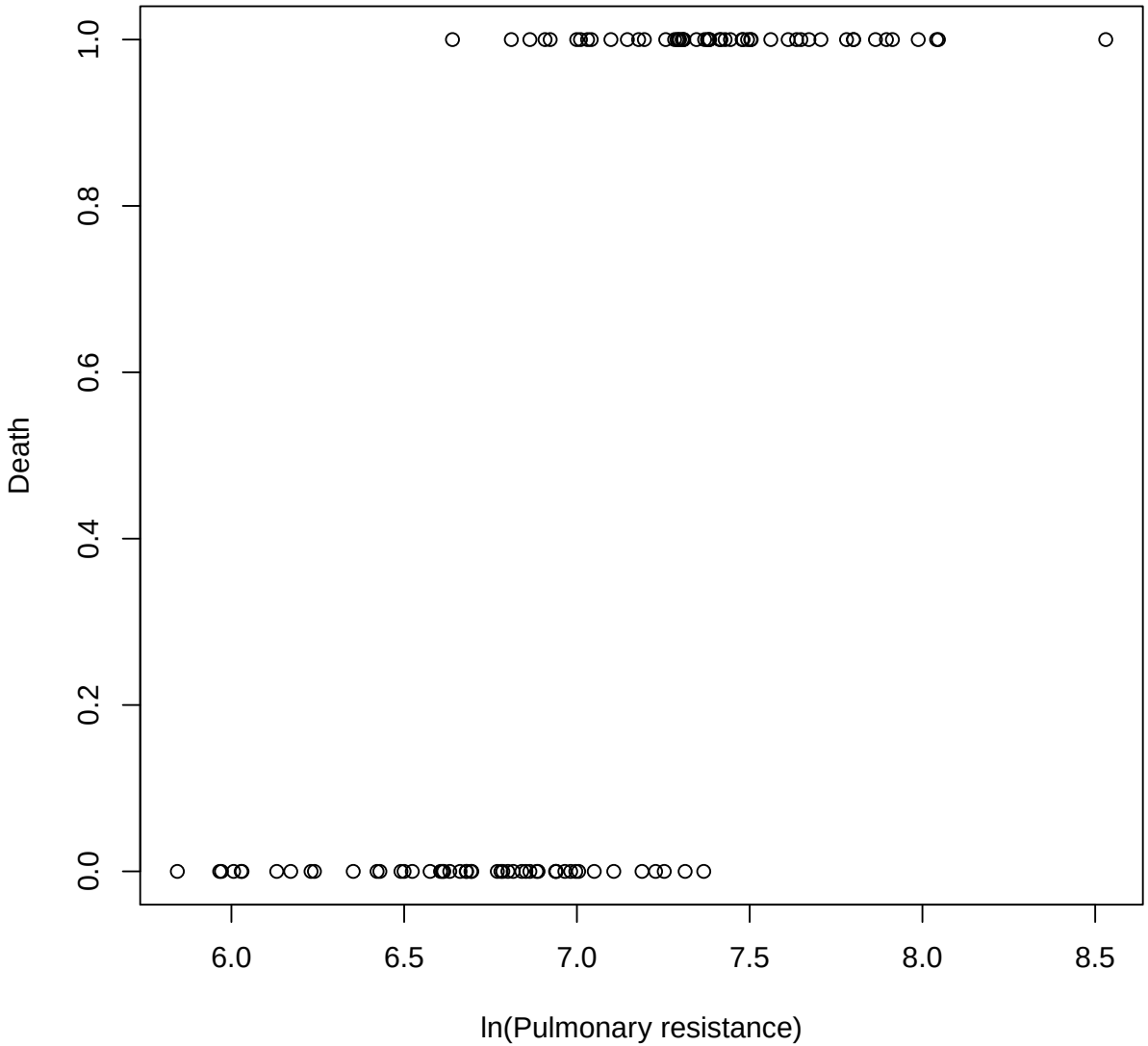
Example data set on Myocardial Infarction

#	Pulse	CI	SI	DBP	PA	VP	PR	Death
1	90	1.71	19.00	16.00	19.50	16.00	912	0
2	90	1.68	18.70	24.00	31.00	14.00	1476	1
3	120	1.40	11.70	23.00	29.00	8.00	1657	1
4	82	1.79	21.80	14.00	17.50	10.00	782	0
5	80	1.58	19.70	21.00	28.00	18.50	1418	1
6	80	1.13	14.10	18.00	23.50	9.00	1664	1
7	94	2.04	21.70	23.00	27.00	10.00	1059	0
8	80	1.19	14.90	16.00	21.00	16.50	1412	0
9	78	2.16	27.70	15.00	20.50	11.50	759	0
10	100	2.28	22.80	16.00	23.00	4.00	807	0
11	90	2.79	31.00	16.00	25.00	8.00	717	0
12	86	2.70	31.40	15.00	23.00	9.50	681	0
13	80	2.61	32.60	8.00	15.00	1.00	460	0
14	61	2.84	47.30	11.00	17.00	12.00	479	0
15	99	3.12	31.80	15.00	20.00	11.00	513	0
16	92	2.47	26.80	12.00	19.00	11.00	615	0
17	96	1.88	19.60	12.00	19.00	3.00	809	0
18	86	1.70	19.80	10.00	14.00	10.50	659	0
19	125	3.37	26.90	18.00	28.00	6.00	665	0
20	80	2.01	25.00	15.00	20.00	6.00	796	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
101	112	1.54	13.80	25.00	31.00	8.00	1610	1

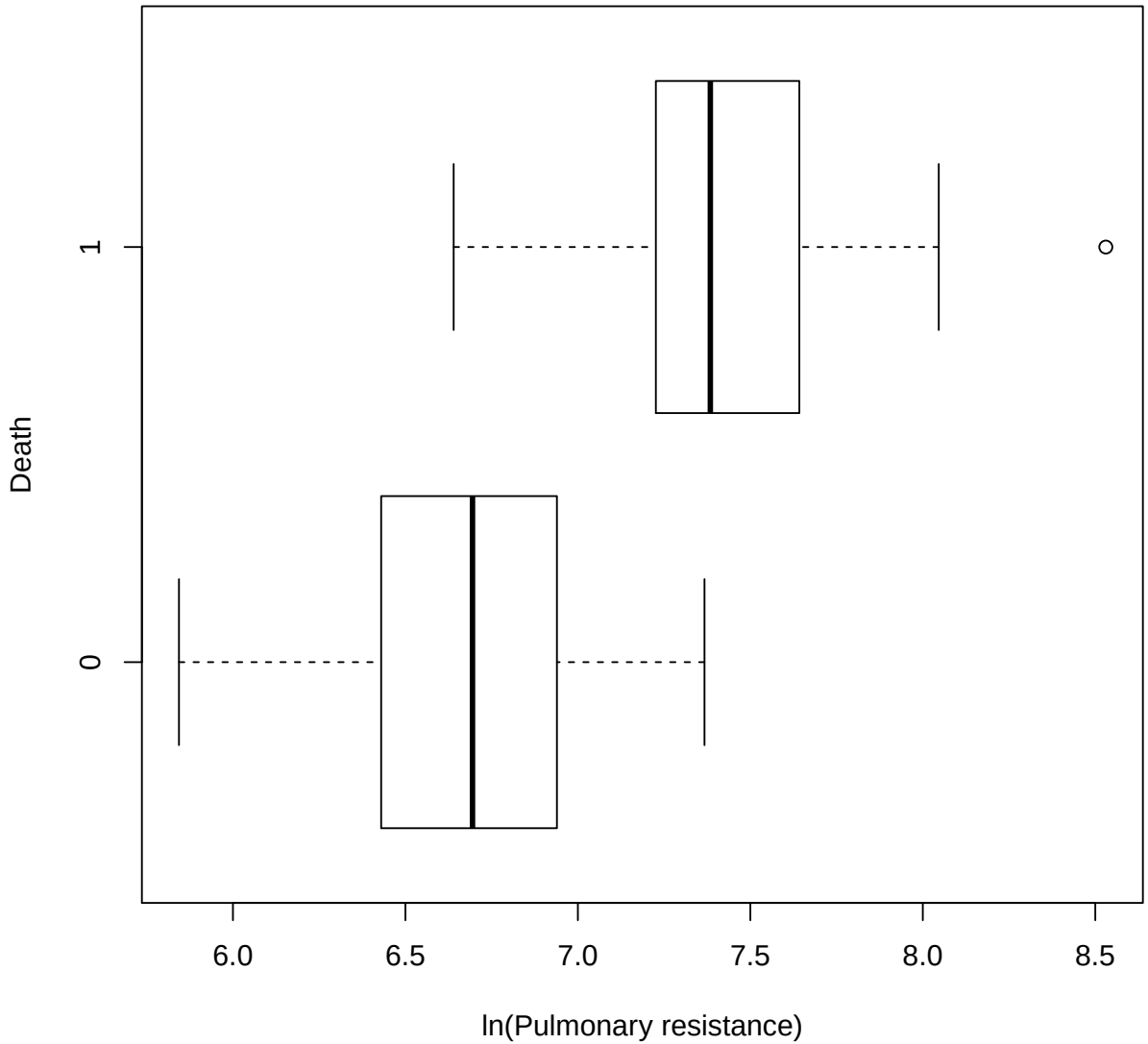
CI = Cardiac index; SI = Systolic index; DBP = Diastolic blood pressure
PA = Pulmonary artery pressure; VP = Ventricular pressure; PR = Pulmonary resistance

Saporta, G. (2006) Probabilités, Analyse des Données et Statistique. Editions Technip, Paris.

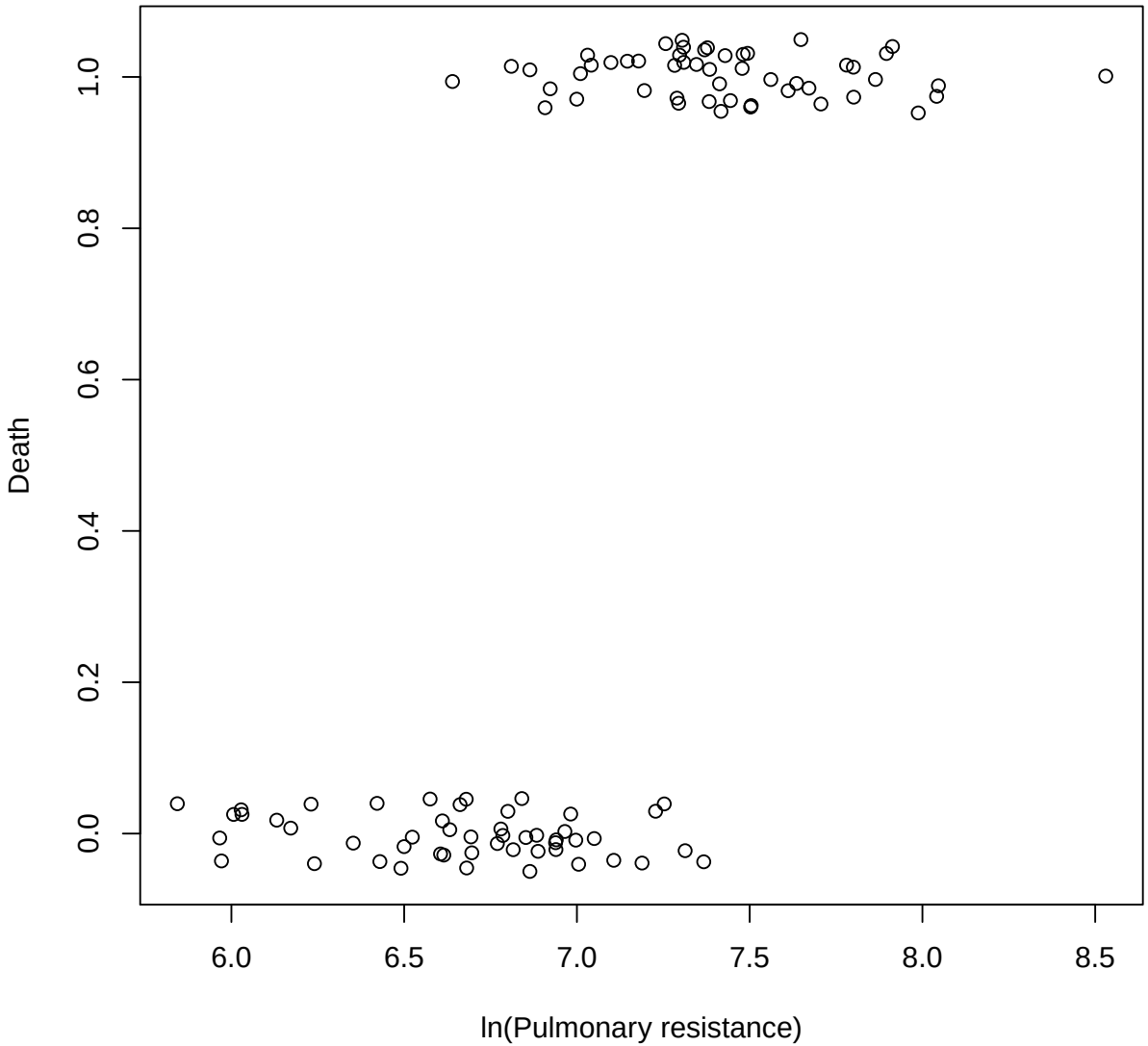
Scatterplot



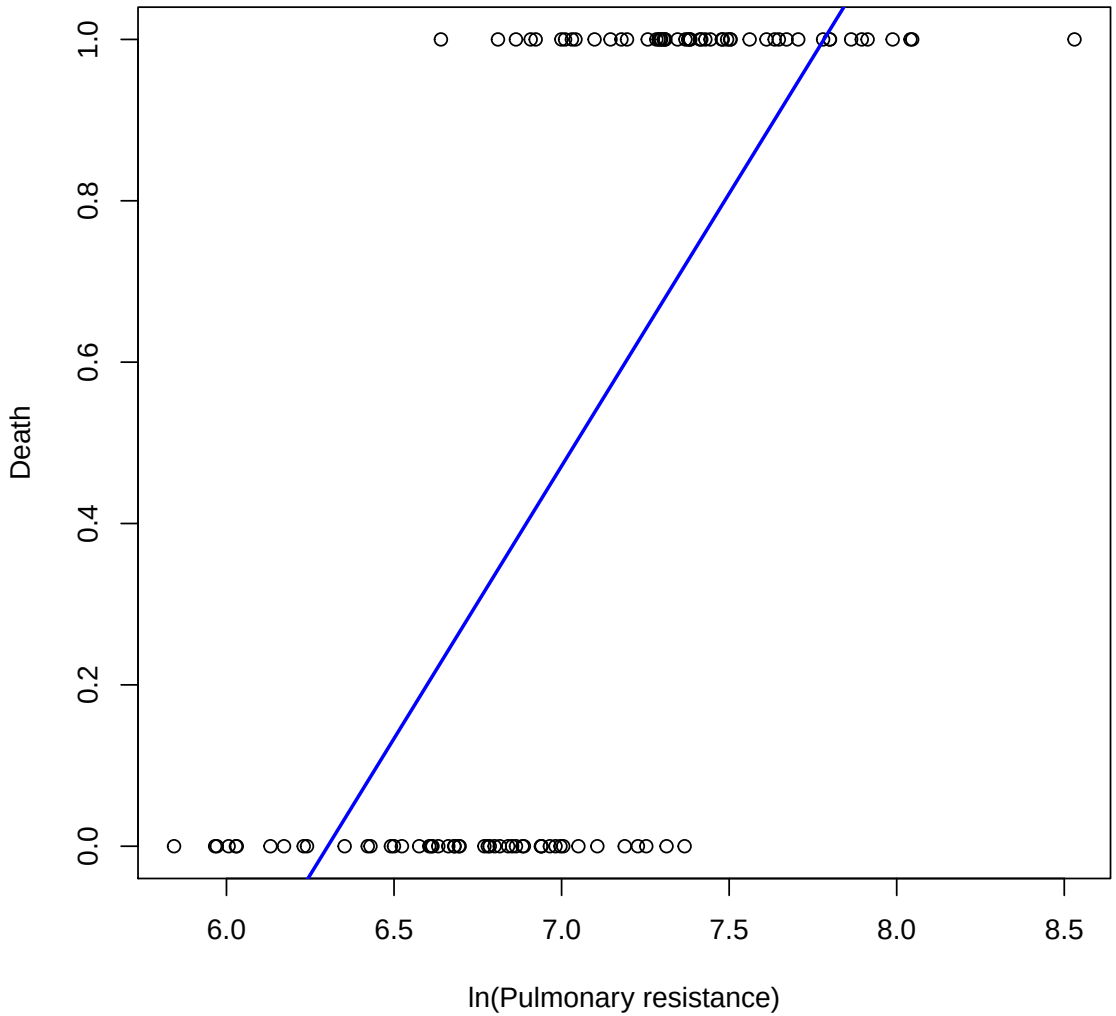
Boxplot



Scatterplot with jitter



Scatterplot with OLS regression line

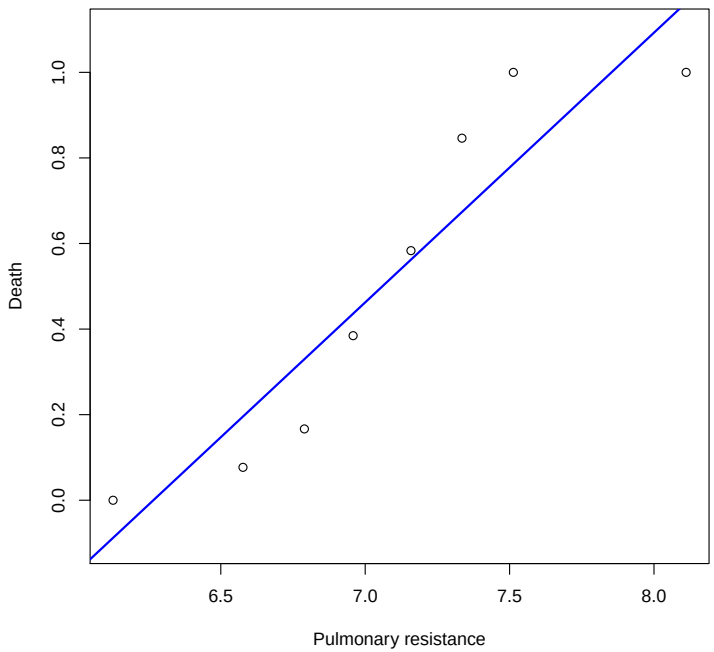


Any problem?

Categorized pulmonary resistance

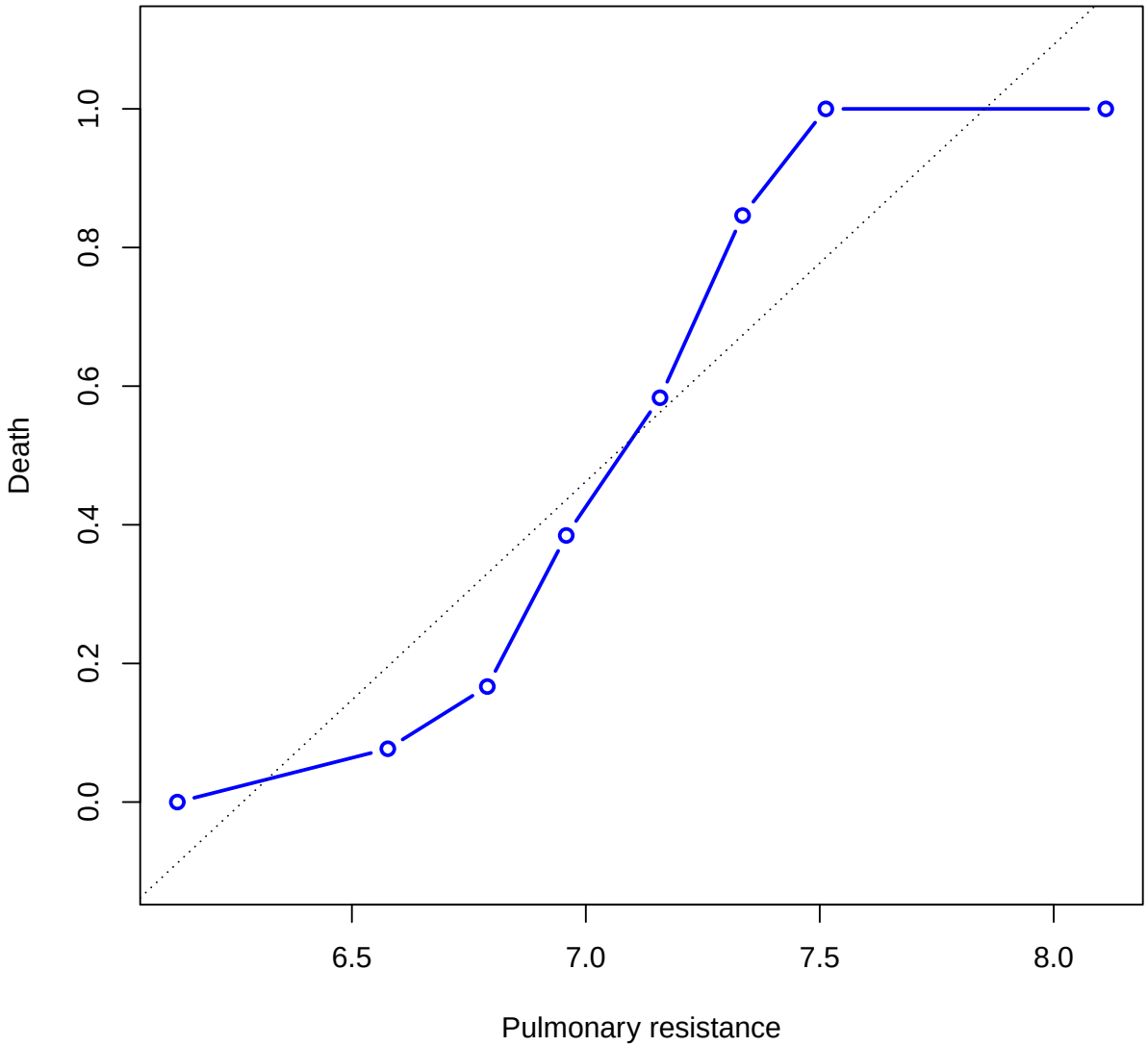
range	n	0	1	proportion
(5.79,6.46]	13	13	0	0.00
(6.46,6.69]	13	12	1	0.08
(6.69,6.89]	12	10	2	0.17
(6.89,7.03]	13	8	5	0.38
(7.03,7.29]	12	5	7	0.58
(7.29,7.38]	13	2	11	0.85
(7.38,7.64]	12	0	12	1.00
(7.64,8.58]	13	0	13	1.00
All	101	50	51	0.50

OLS regression

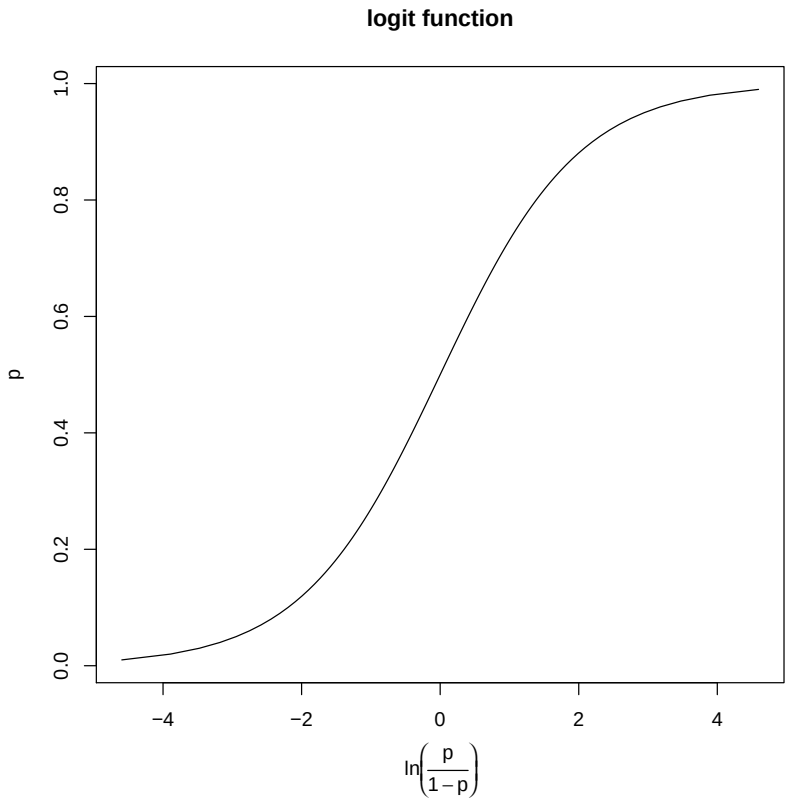



```
> lm.out <- lm(pro~m)
> summary(lm.out)

Call:
lm(formula = pro ~ m)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16296 -0.12971 -0.01517  0.10913  0.21427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.94996    0.70645  -5.591 0.001392 **
m            0.63033    0.09959   6.330 0.000727 ***
---

Residual standard error: 0.16 on 6 degrees of freedom
Multiple R-squared:  0.8697,Adjusted R-squared:  0.848
F-statistic: 40.06 on 1 and 6 DF,  p-value: 0.0007272

>
```

Observed and fitted pattern



The logit function



Logit (or logistic) function:

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

Inverse of the logit function

$$logit^{-1}(\pi) = \frac{e^{\pi}}{e^{\pi}+1}$$

Using logit(π) as the response is the basis of logistic regression

The logistic regression model

$$\pi(x) = E(Y|x) = P(Y = 1|x)$$

$$\text{Model: } y = \pi(x) + \varepsilon \quad y|x \sim Bin(n = 1, \pi(x))$$

$$\varepsilon = \begin{cases} 1 - \pi(x) & \text{if } y = 1 \quad \text{with prob.} \quad \pi(x) \\ -\pi(x) & \text{if } y = 0 \quad \text{with prob.} \quad 1 - \pi(x) \end{cases}$$

$$E(\varepsilon) = (1 - \pi(x))\,\pi(x) - \pi(x)\,(1 - \pi(x)) = 0$$

$$V(\varepsilon) = \pi(x)(1 - \pi(x))$$

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

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{e^{\beta_0 + \beta_1 x} + 1}$$

Note that

- $0 \leq \pi(x) \leq 1$
- $-\infty \leq g(x) \leq +\infty$

Model and likelihood

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i \ln \left[\pi(x_i)\right] + (1 - y_i) \ln \left[1 - \pi(x_i)\right]\}$$

We maximize $\ell(\beta_0, \beta_1)$ by numerical methods

Fitting a logistic model regression in R

```
model <- glm(Death~lPR, family = binomial(link = 'logit'), trace=FALSE)
summary(model)

Call:
glm(formula = death ~ lPR, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.09196  -0.41945   0.01073   0.46258   2.36750

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -46.651      9.231  -5.054 4.33e-07
lPR            6.613      1.307   5.059 4.22e-07

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.006  on 100  degrees of freedom
Residual deviance:  64.529  on  99  degrees of freedom
AIC: 68.529

Number of Fisher Scoring iterations: 6

>
```

Iteration history

Iteration	log-likelihood	Deviance
1	-37.30	74.60
2	-33.08	66.17
3	-32.30	64.61
4	-32.26	64.53
5	-32.26	64.53
6	-32.26	64.53

Writing the fitted model

In OLS regression we used:

$$\hat{y}_i = b_0 + b_1 x_i$$

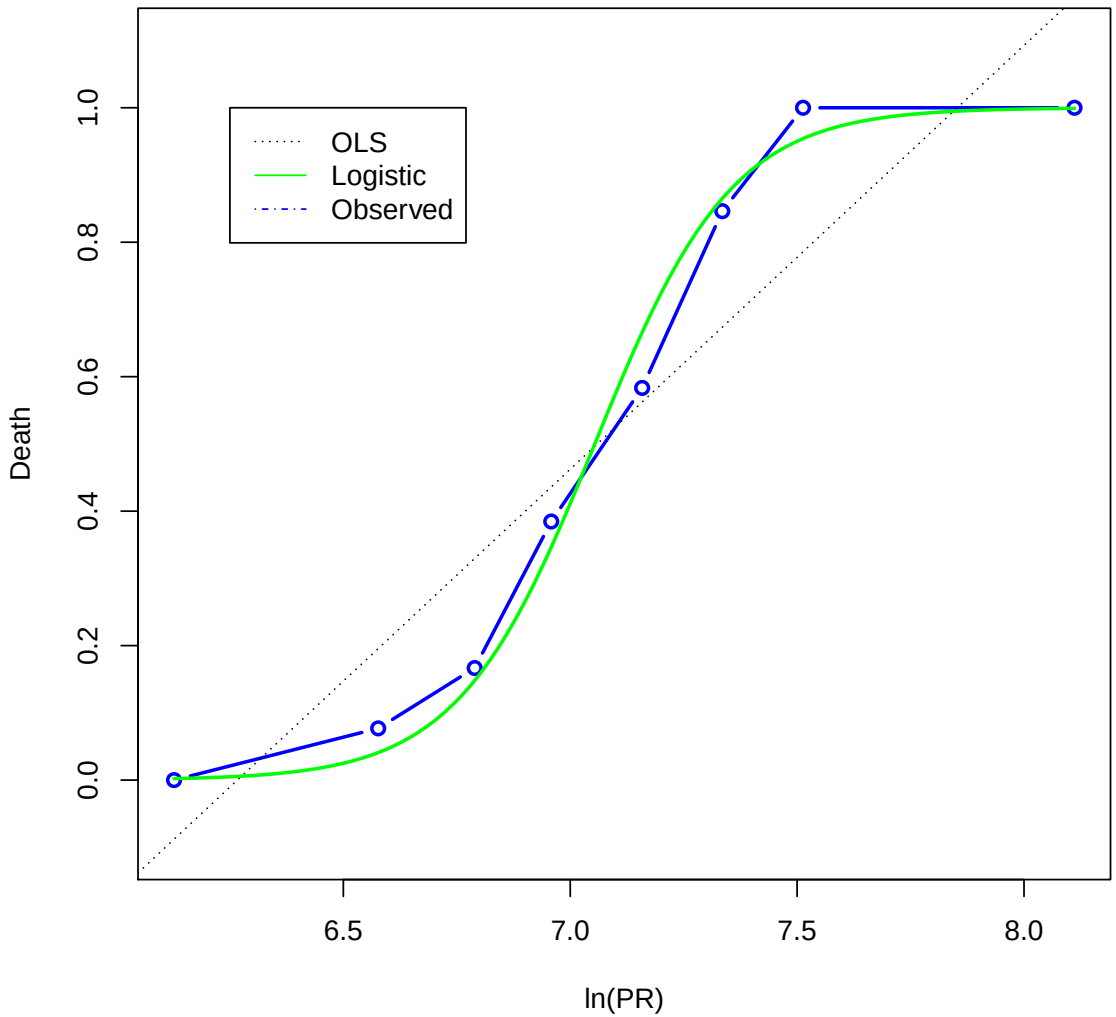
In logistic regression we have the fitted values:

$$\hat{\pi}(x) = \frac{e^{-46.651+6.613 Pul.Res}}{1 + e^{-46.651+6.613 Pul.Res}}$$

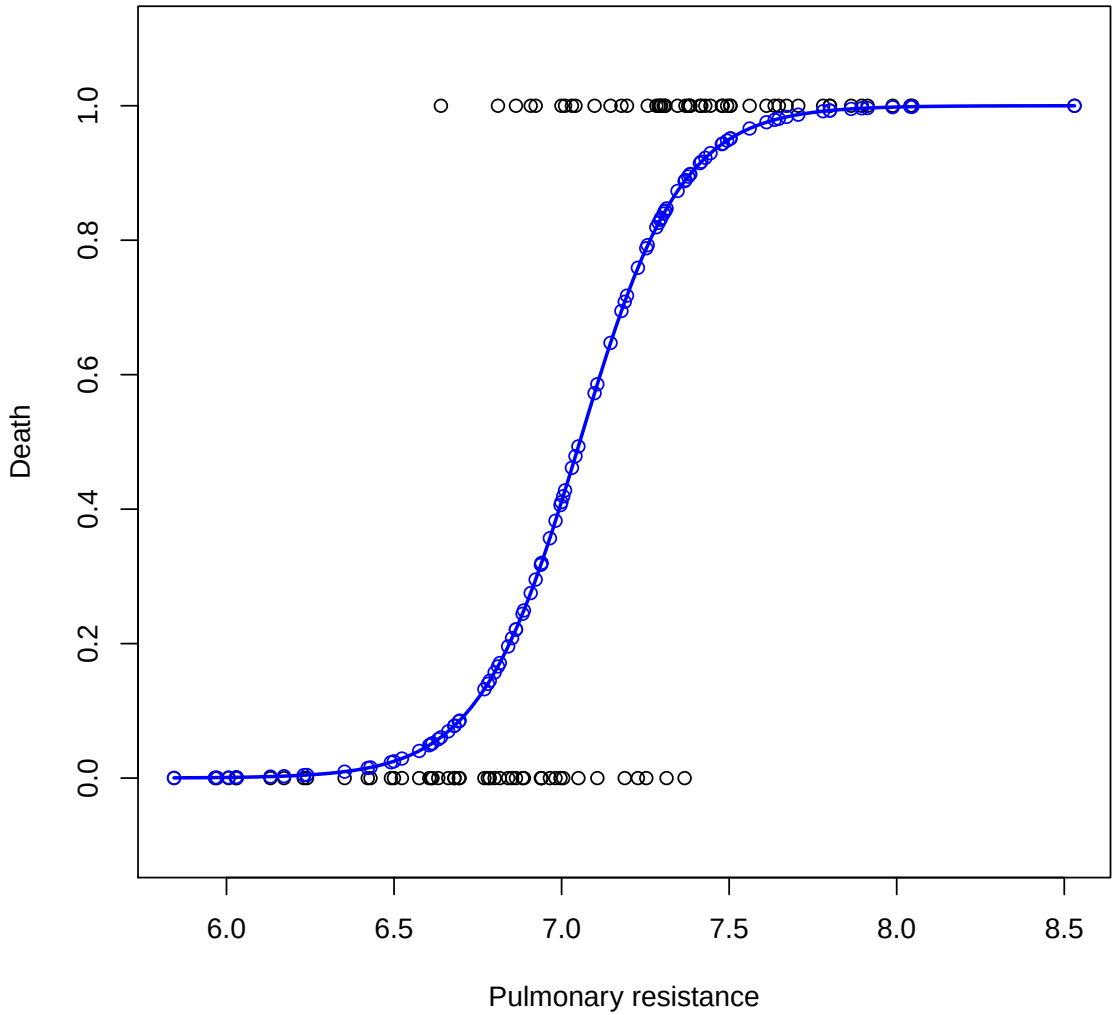
or the estimated logit:

$$\hat{g}(x) = -46.651 + 6.613 Pul.Res$$

Plotting the fitted logistic model



Plotting the fitted logistic model (usual representation)



Some R code

```
model <- glm(Death~lPR, family = binomial(link = 'logit'), trace=TRUE)
summary(model)
plot(lPR,Death,xlab="Pulmonary resistance",ylab="Death", ylim=c(-0.1,1.1))
curve(predict(model,data.frame(lPR=x),type="resp"),add=TRUE,col="blue",lwd=2)
points(lPR,fitted(model),pch=1,col="blue")
```

Likelihood ratio test for comparing models (1/2)

- We first compare the fitted model with a saturated model:

$$D = -2 \ln \left(\frac{\text{Likelihood fitted model}}{\text{Likelihood saturated model}} \right)$$

- A saturated model is a model with as many data points as parameters.
- D is usually called the deviance, and is analogous to the sum-of-squares of the residuals.
- The likelihood of the saturated model is

$$\prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} = \prod_{i=1}^{n} {y_i}^{y_i} \left[1 - y_1\right]^{1-y_i} = 1$$

- The deviance simplifies to

$$D = -2 \ln \left(\text{Likelihood fitted model}\right)$$

- The null deviance is the deviance of a model containing only the intercept.

Likelihood ratio test for comparing models (2/2)

- We wish to compare the model with and without the predictor (pulmonary resistance)
- $$\begin{aligned} G &= -2 \ln \left(\frac{\text{Likelihood without predictor}}{\text{Likelihood with predictor}} \right) \\ &= -2 \left[\ln (\text{Likelihood without predictor}) - \ln (\text{Likelihood with predictor}) \right] \\ &= D(\text{without predictor}) - D(\text{with predictor}) \end{aligned}$$
- The reduction in deviance determines if the predictor is relevant.

Assessing over all model fit (1/2)

```
> anova(model,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Death

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   100    140.006
lPR   1   75.477        99     64.529 < 2.2e-16 ***
>
```

$$G = 140.006 - 64.529 = 75.477$$

$$P\left(\chi^2_{(1)} > 75.47\right) \approx 0$$

Assessing over all model fit (2/2)

- Some programs report McFadden's pseudo R^2 for assessing model fit.
- $$R^2_{\text{McFadden}} = 1 - \frac{\text{Likelihood model considered}}{\text{Likelihood null model}}$$
- $0 \leq R^2_{\text{McFadden}} \leq 1$
- For the example at hand

$$R^2_{McFadden} = 1 - \frac{-32.26456}{-70.00291} = 0.539$$

- Interpretation different from R^2 in standard linear regression

Testing individual predictors

In logistic regression three procedures are in use to test predictors for significance

1. Likelihood ratio test (LRT)
2. Wald test
3. Score test

The Wald test: $H_0 : \beta_i = 0 \qquad \beta_i \neq 0$

$$Z = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} \sim N(0,1) \text{ under } H_0$$

Wald confidence interval

$$CI(\beta_i) = \hat{\beta}_i \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_i)$$

E.g. for Pulm. Res.

$$Z = \frac{6.613}{1.307} = 5.059 \qquad \text{p-value} = 2P\left(Z > 5.059\right) = 4.22e-07$$

$$CI(\beta_{PM}) = 6.613 \pm 1.96 \cdot 1.307 = (4.05, 9.18)$$

```
> confint(model)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -67.486679 -30.879875
lPR           4.380653   9.566105
>
```

Logistic regression with binary predictor

- $\pi(x) = E(Y|x) = P(Y = 1|x)$
- The odds of $Y = 1$ with $x = 1$ is

$$\text{odds} = \frac{\pi(1)}{1 - \pi(1)}$$

- Note the logit is just the logarithm of the odds.
- The odds ratio is a widely used measure of association.
-

$$\text{OR} = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}$$

- For the logistic model $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{e^{\beta_0+\beta_1 x}+1}$

	$x = 1$	$x = 0$
$y = 1$	$\pi(1)$	$\pi(0)$
$y = 0$	$1 - \pi(1)$	$1 - \pi(0)$

-

$$\text{OR} = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} = e^{\beta_1}$$

Example

$$\text{CID} = I_{CI<Me}$$

	Death	
CID	0	1
0	42	9
1	8	42

$$\text{OR} = \frac{42 \times 42}{8 \times 9} = 24.5$$

```
Call:
glm(formula = Death ~ CId, family = binomial(link = "logit"),
    trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9145  -0.6231   0.5905   0.5905   1.8626

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5404     0.3673  -4.194 2.74e-05
CIdTRUE       3.1987     0.5327   6.005 1.91e-09

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.006  on 100  degrees of freedom
Residual deviance:  91.499  on  99  degrees of freedom
AIC: 95.499

Number of Fisher Scoring iterations: 4
```

$$\text{OR} = e^{3.1987} = 24.5$$

Confidence interval for the odds ratio

$$CI(\beta_i) = \hat{\beta}_i \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_i)$$

$$CI(OR) = e^{\hat{\beta}_i \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_i)}$$

For the data at hand:

$$CI(\beta_i) = (2.15, 4.24)$$

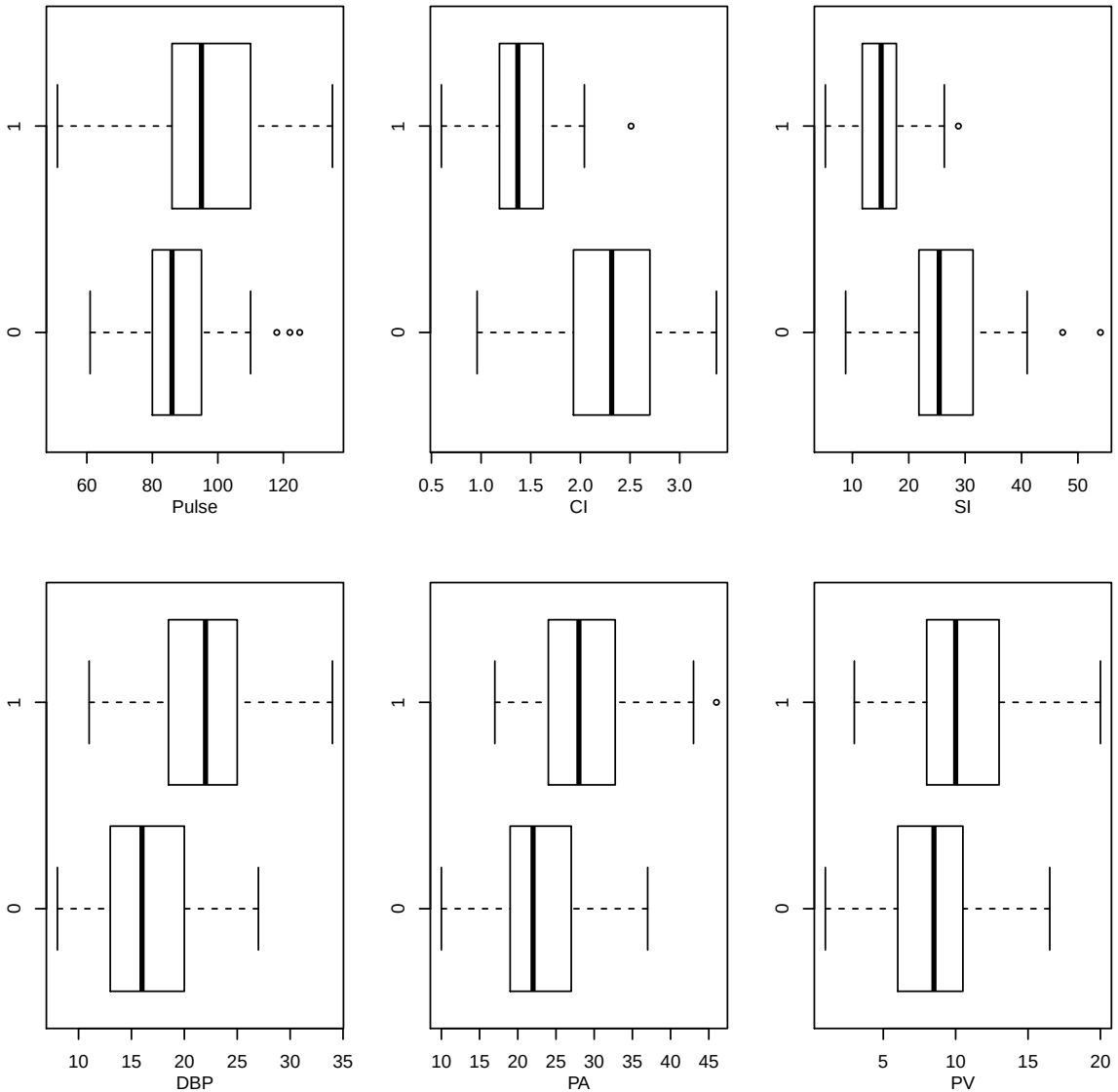
$$CI(OR) = (8.63, 69.59)$$

Interpretation with continuous predictor

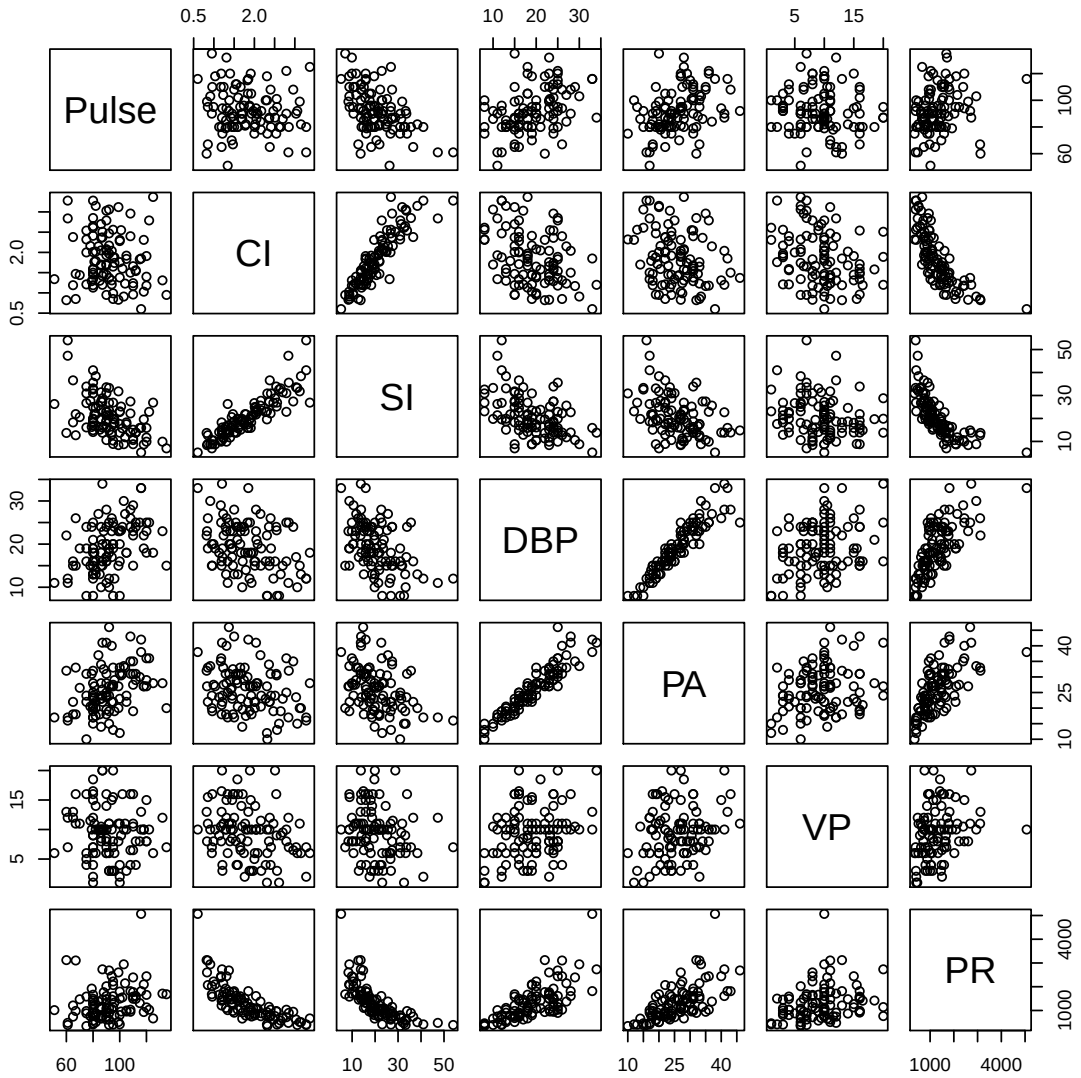
```
Call:
glm(formula = Death ~ Pulse, family = binomial(link = "logit"),
    trace = TRUE)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.73307    1.23132  -2.220   0.0264
Pulse        0.02991    0.01321   2.263   0.0236

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.01  on 100  degrees of freedom
Residual deviance: 134.45  on  99  degrees of freedom
AIC: 138.45

Number of Fisher Scoring iterations: 4
```

- Estimated logit $\hat{g}(x) = -2.73307 + 0.02991 Pulse$
- The slope gives the change in the logit for a one-unit change in Pulse.
- With a one-unit change in Pulse, the odds for death is multiplied by $e^{0.02991} = 1.03$
- With a 10-unit change in Pulse, the odds for death is multiplied by $e^{10 \times 0.02991} = 1.35$

Boxplots for other predictors



Relationships between predictors



Multiple predictors

```
> summary(model)

Call:
glm(formula = Death ~ Pulse + CI + SI + DBP + PA + VP + lPR,
    family = binomial(link = "logit"), trace = TRUE)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.59039  -0.40158   0.02522   0.39452   2.66587

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.43452   52.39826   0.352    0.725
Pulse        0.04705    0.08874   0.530    0.596
CI          -7.35661    6.15306  -1.196    0.232
SI           0.10457    0.39514   0.265    0.791
DBP          0.05335    0.20022   0.266    0.790
PA           0.25157    0.30728   0.819    0.413
VP           0.05218    0.07913   0.659    0.510
lPR         -2.79126    7.29886  -0.382    0.702

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.006  on 100  degrees of freedom
Residual deviance:  58.497  on  93  degrees of freedom
AIC: 74.497

Number of Fisher Scoring iterations: 7

>
```

But note that $G = 140.006 - 58.497 = 81.509$ and $P\left(\chi^2_7 \geq 81.509\right) = 6.779506e - 15$

After backward elimination

```
Call:
glm(formula = Death ~ CI + PA, family = binomial(link = "logit"),
    trace = TRUE)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.22793  -0.39632   0.02777   0.44453   2.71096

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9331     1.7855   1.643   0.1004
CI           -4.5491     0.9402  -4.838 1.31e-06
PA            0.2015     0.0622   3.239   0.0012

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.01  on 100  degrees of freedom
Residual deviance:  60.27  on  98  degrees of freedom
AIC: 66.27

Number of Fisher Scoring iterations: 6

>
```

$$G = 60.270 - 58.497 = 1.773 \qquad P\left(\chi^2_5 \geq 1.773\right) = 0.880$$

Overdispersion

- In logistic regression, overdispersion sometimes occurs.
- Overdispersion refers to the fact that the variance exceeds the theoretical binomial variance.
- With overdispersion, standard errors are typically too small.
- Overdispersion can be modelled with $V(Y_i) = \phi E(Y_i)$, where ϕ is the overdispersion parameter
- ϕ can be estimated as $\hat{\phi} = \frac{X^2}{df}$.
- This can be done by quasi-binomial regression.

Example

```
model <- glm(Death~VP, family = binomial(link = 'logit'))
> summary(model)

Call:
glm(formula = Death ~ VP, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6022  -1.1319   0.6562   1.1386   1.5492

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.24190    0.52391  -2.370  0.01777 *
VP           0.13340    0.05124   2.603  0.00923 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.01  on 100  degrees of freedom
Residual deviance: 132.48  on  99  degrees of freedom
AIC: 136.48

Number of Fisher Scoring iterations: 4
```

```
model <- glm(Death~VP, family = quasibinomial(link = 'logit'))
> summary(model)

Call:
glm(formula = Death ~ VP, family = quasibinomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6022  -1.1319   0.6562   1.1386   1.5492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.24190    0.52801  -2.352   0.0206 *
VP           0.13340    0.05164   2.583   0.0113 *
---

(Dispersion parameter for quasibinomial family taken to be 1.0156

    Null deviance: 140.01  on 100  degrees of freedom
Residual deviance: 132.48  on  99  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

References

- Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013) Applied Logistic Regression. 3rd Edition. Wiley, New York.

Exercise

Low birth weight. In a study on low birth weight (LBW) of 189 babies, the following variables were registered: LOW ($0 = \geq 2500$ g; $1 = < 2500$ g), AGE (of the mother), LWT (Weight of mother at last menstrual period), RACE (1 = White; 2 = Black; 3 = Other), SMOKE (Smoking status during pregnancy 0 = No; 1 = Yes), PTL (History of premature labor 0 = None; 1 = One; 2 = Two, etc.), HT (History of hypertension 0 = No; 1 = Yes), UI (Presence of Uuterine irritability 0 = No; 1 = Yes), FTV (Number of physician visits during the first trimester 0 = None; 1 = One; 2 = Two etc.) and BWT (Recorded birth weight). The file `lowbwt.dat` contains the data.

- Load the data in the R environment with the `read.table` instruction.
- Is LBW related to the smoking status of the mother? Make the corresponding two-way table and do a chi-square test to test for independence.
- Do a logistic regression of LBW on SMOKE. Report the estimated logit. Is there a significant association?
- Calculate the odds of LBW for smokers and non-smokers. Calculate the odds ratio (OR). Construct a 95% confidence interval for the OR.
- Investigate the effect of LWT on the risk of LBW. What relation do you expect? Does the data present evidence for the relation you expected? What's the effect of LWT on the odds of a LBW child? Is that effect significant?
- Calculate the predicted probabilities of LBW according to the last model, and plot the logistic regression curve.
- Is smoking still relevant if we control for the effect of LWT? Calculate the *adjusted OR* of the effect of smoking. Give a 95% confidence interval for the adjusted OR and compare this with the unadjusted OR.
- Investigate the effect of RACE on the odds of a LBW child. Take care to use indicator variables for its categories. Is RACE a relevant predictor of LBW?
- Build a model with all available predictors in the database. If possible, try to simplify the model, and suggest what you would think that is a good model for the data.