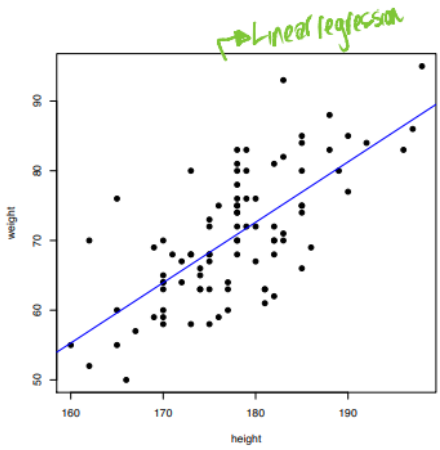
Ricard Garcia Isern
2023

# LOGISTIC REGRESSION

## 1. Introduction

- **CLASSICAL LINEAR REGRESSION:**




normal distribution variable

Theoretical model:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$ error

Usual assumptions:

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

Summarized:

$$Y_i | X_i \sim N(\underbrace{\beta_0 + \beta_1 X_i}_{\text{mean}}, \sigma^2)$$

- Logistic regression → **binary variable** →only 2 possible outcomes.

- **EXAMPLE:**

Example data set on Myocardial Infarction

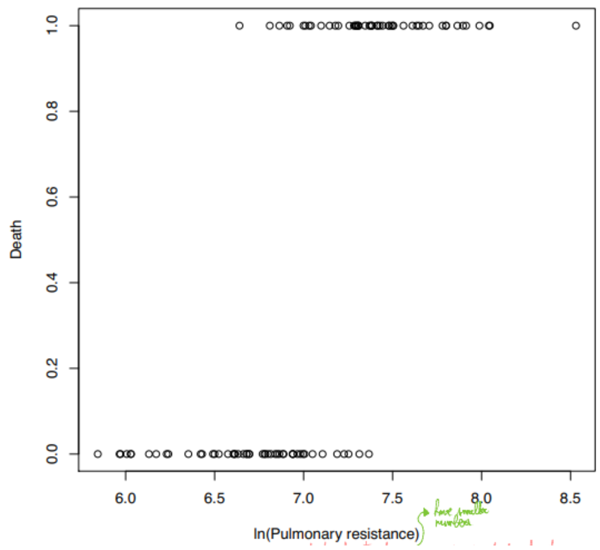
predictors

| # | 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 |

Death → binary ⇓ what we want to study → Model

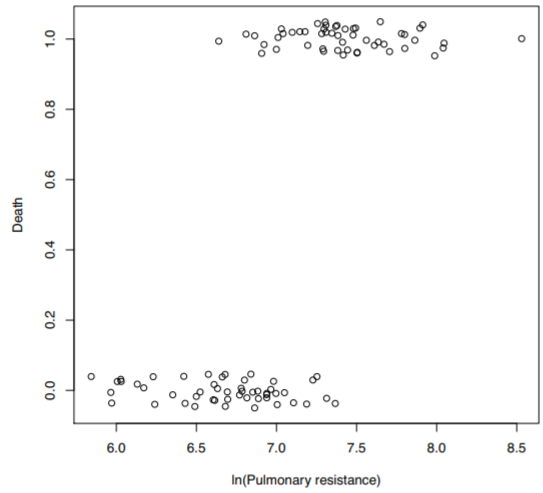
- **Jitter** → when data is very close (collapsed), the axes are opened and we can better visualize the data.

## Scatterplot



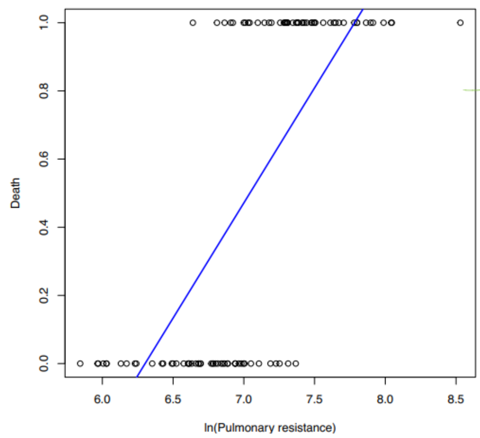

## Scatterplot with jitter

when the data is very together, collapsed, the axes are opened and we can better visualizing the data



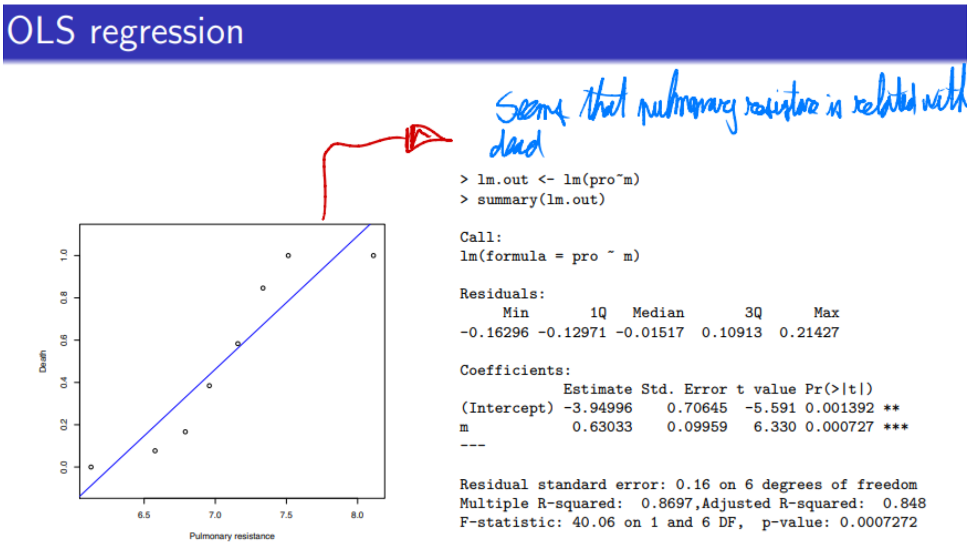

- We apply **logarithm transformation** → reduce number of outliers = we must undo it when explaining the final results.

## Scatterplot with OLS regression line

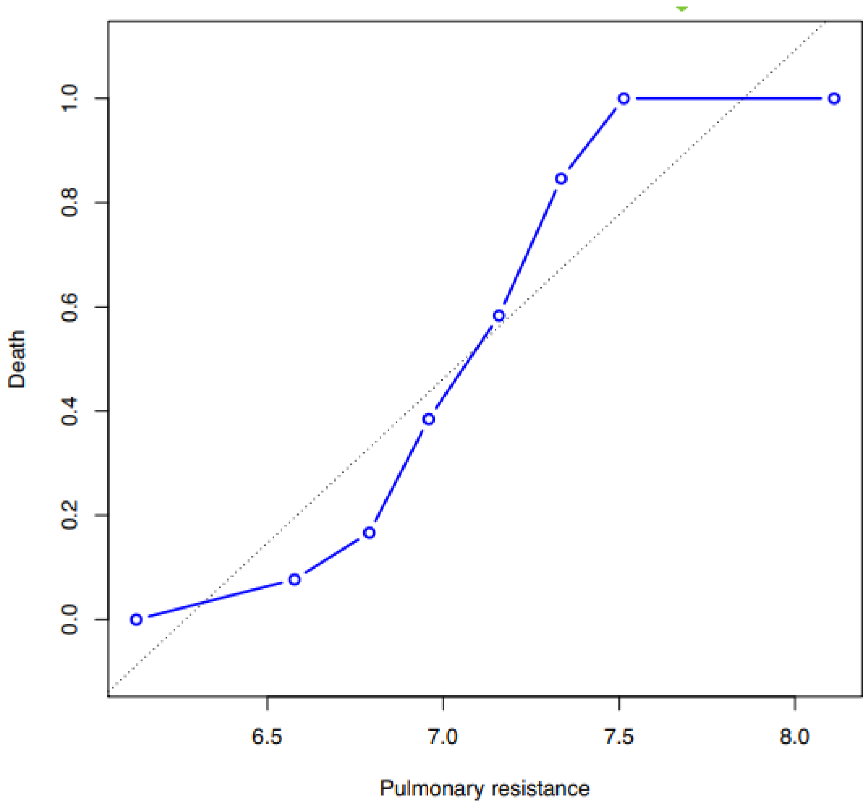



we don't get useful information.

- To get useful information we can do a range for the predictors → replot and see differences.

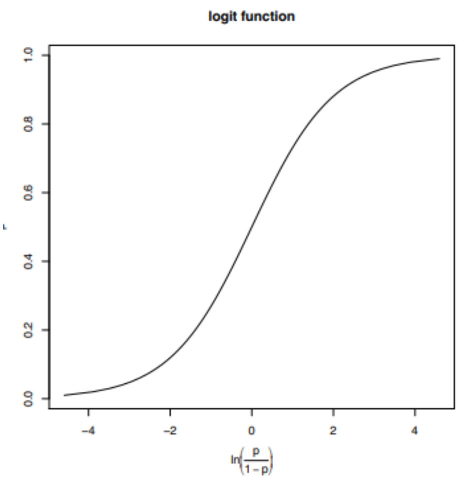



- We must join the points → **overfitting.**



## 2. FITTING THE LOGISTIC MODEL

### LOGIT FUNCTION




why!!?

$$\boxed{\text{Logit (or logistic) function:}}$$

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

Inverse of the logit function

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

to undo the transformation

Using logit($\pi$) as the response is the basis of logistic regression

prob. of x

expected value of y given value x

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

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

Error = residuals

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

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

expected values of error must be 0.

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

transformation

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

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

y [illegible], probability

Note that

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

# 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
```

# Writing the fitted model

ordinary list square

In OLS regression we used:

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

In logistic regression we have the fitted values:

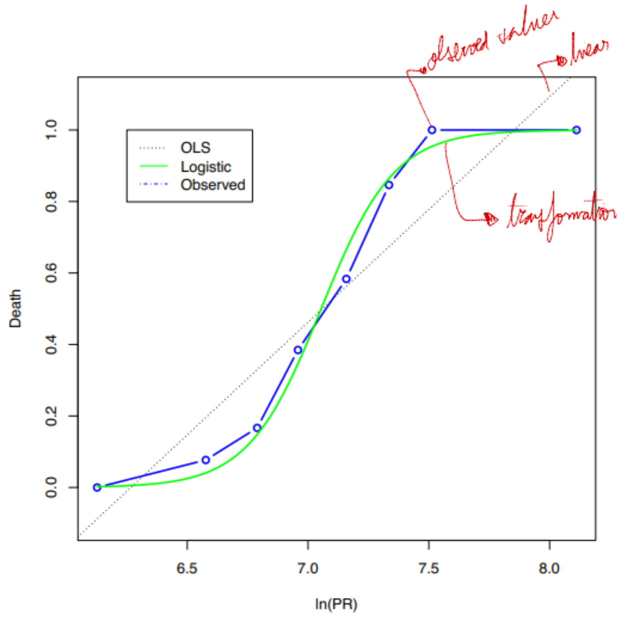
$$\boxed{\hat{\pi}(x) = \frac{e^{-46.651+6.613 Pul.Res}}{1+e^{-46.651+6.613 Pul.Res}}}$$

must do transformation

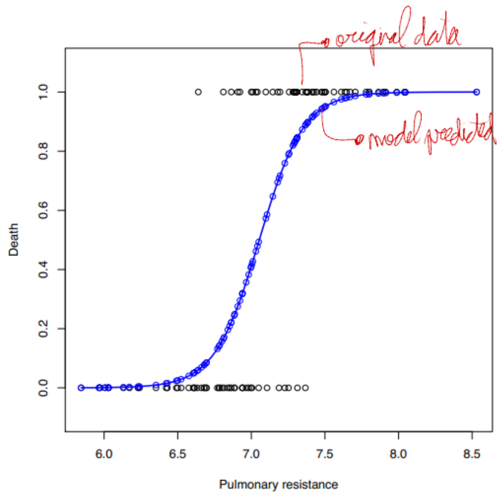
or the estimated logit:

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

## Plotting the fitted logistic model




## Plotting the fitted logistic model (usual representation




## Likelihood ratio test for comparing models (1/2)

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

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

has as many parameters as observations ↳ follow the data with line

- 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

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

- The deviance simplifies to

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

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

- We wish to compare the model with and without the predictor (pulmonary resistance)
-

$$\begin{aligned} G &= -2\ln\left(\frac{\text{Likelihood without predictor}}{\text{Likelihood with predictor}}\right) \\ &= -2\left[\ln(\text{Likelihood without predictor}) - \ln(\text{Likelihood with predictor})\right] \\ &= D(\text{without predictor}) - D(\text{with predictor}) \end{aligned}$$

$\beta_0, \beta_1$

- The reduction in deviance determines if the predictor is relevant.

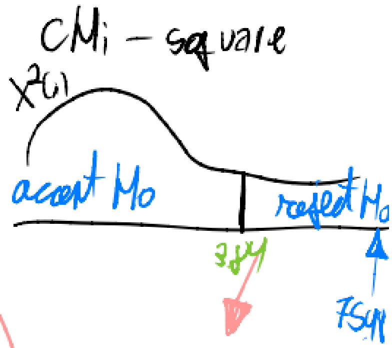
```
> anova(model,test="Chisq")
Analysis of Deviance Table

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 ***
>
```




$$\boxed{\begin{array}{c} G = 140.006 - 64.529 = 75.477 \\ P\left(\chi^2_{(1)} > 75.47\right) \approx 0 \end{array}}$$

compares amb valor chi-square i determina si a significant model parametre

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

## 3. Hypothesis testing

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: $H_0 : \beta_i = 0$ $H_A : \beta_i \neq 0$

→ follow normal distribution

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

Wald confidence interval

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

E.g. for Pulm. Res.

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

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

→ we can only interpret $g(x)$ = logit

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

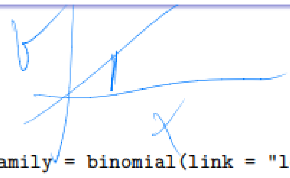
---

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

| | Death | |
|---|---|---|
| CID | 0 | 1 |
| 0 | 42 | 9 |
| 1 | 8 | 42 |

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

→ odd ratio



```
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
```

OR

$a + b \times age$

$e$

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

→ used when we have binary data

$a + b \times \underset{\text{OR}}{age}$ → how much increase for each time we increase x.

## Confidence interval for the odds ratio

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

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

## 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
```

- We transform data to be linear
- We undo changes to study results.

model??

- 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 $e^{10\times 0.02991} = 1.35$

# 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

>
```

None predictor is significant

No * * *

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

# 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 $\phi$ is the overdispersion parameter
- $\phi$ can be estimated as $\boxed{\hat{\phi} = \frac{X^2}{df}}$.
- This can be done by quasi-binomial regression.

OVERDISPERSION

```
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
```

when higher than 1.2 we must consider quasi binomial