GLM

Questions

- Stormboard: ID: 664949 Key: yellow25

Binomial Regression

Logistic -> Binomial

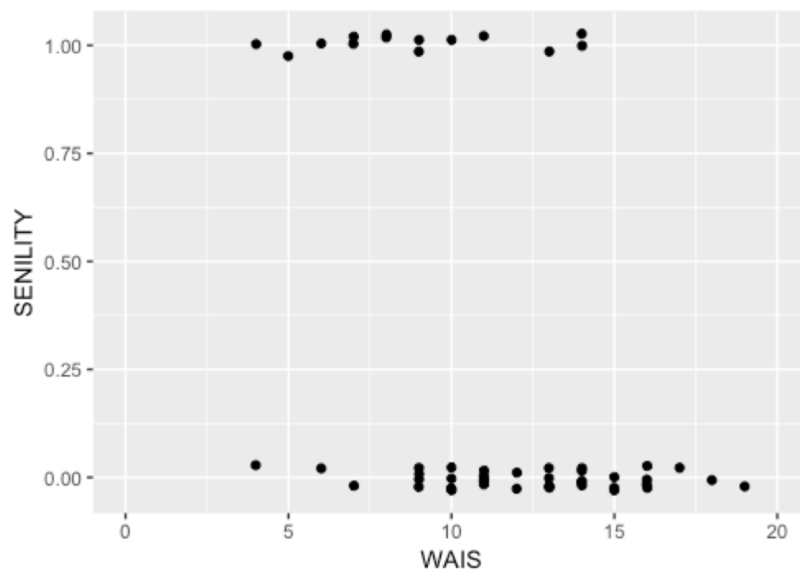
- Bernoulli (probability of one coin flip)
- Binomial (probability of n coin flips)

$$f(y_i, \mu) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

Senility and Wechsler Adult Intelligence Scale

- 54 elderly people
  - Psychiatric examination to determine signs of senility (0/1).
  - Also measured subset of Wechsler Adult Intelligence Scale (WAIS) (0-20)

	WAIS	SENILITY
1:	6	0
2:	8	1
3:	13	1
4:	15	0
5:	4	0
6:	12	0
7:	10	0
8:	15	0
9:	7	1
10:	18	0



Senility and Wechsler Adult Intelligence Scale

- 54 elderly people,
  - Psychiatric examination to determine signs of senility.
  - Also measured subset of Wechsler Adult Intelligence Scale (WAIS)

	WAIS	SENILITY
1:	6	0
2:	8	1
3:	13	1
4:	15	0
5:	4	0
6:	12	0
7:	10	0
8:	15	0
9:	7	1
10:	18	0

Logistic regression

```
glm(formula = SENILITY ~ WAIS, family = binomial,
data = swdt)
             coef.est coef.se
(Intercept)  2.40     1.19
WAIS        -0.32     0.11
---
  n = 54, k = 2
  residual deviance = 51.0, null deviance = 61.8
(difference = 10.8)
```

Aggregation

- n=54 elderly people can be grouped by G=17 unique WAIS groups

```
    WAIS SENILITY
 1:    6        0
 2:    8        1
 3:   13        1
 4:   15        0
 5:    4        0
 6:   12        0
 7:   10        0
 8:   15        0
 9:    7        1
10:   18        0
```

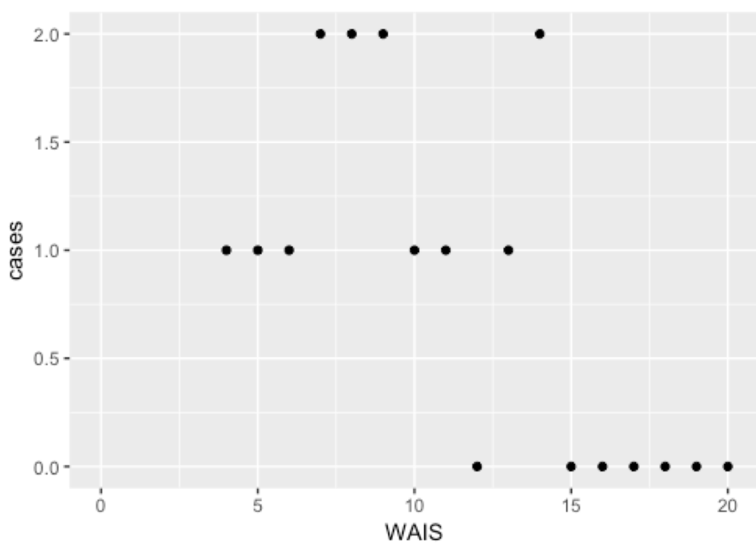
```
swdt%>%group_by(WAIS)%>%
summarise(cases=sum(SENILITY),n=length(SENILITY))
```

```
# A tibble: 17 x 3
    WAIS cases     n
   <int> <int> <int>
 1     4     1     2
 2     5     1     1
 3     6     1     2
 4     7     2     3
 5     8     2     2
 6     9     2     6
 7    10     1     6
 8    11     1     6
 9    12     0     2
10    13     1     6
```

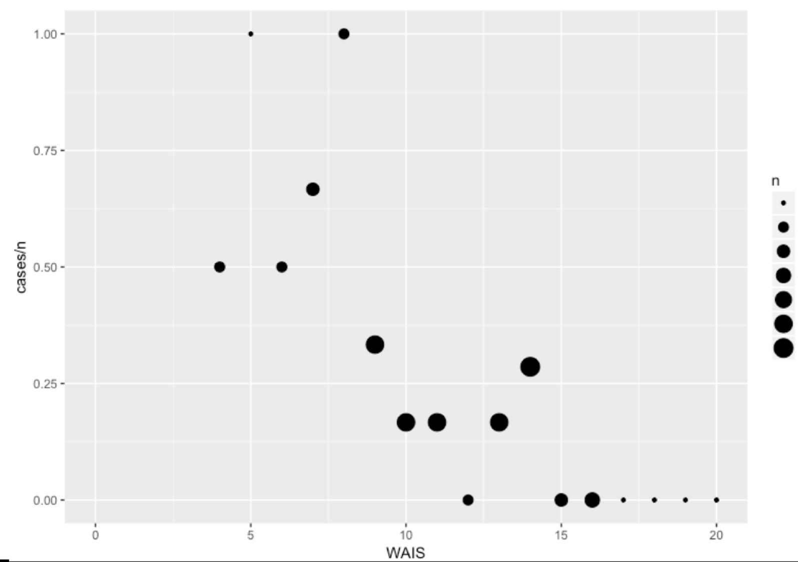
Count or proportion?

- Key distinction is are these numbers capped?

count



Count/n



Aggregation

- n=54 elderly people can be grouped by G=17 unique WAIS groups

```
# A tibble: 17 x 3
    WAIS cases     n
   <int> <int> <int>
 1     4     1     2
 2     5     1     1
 3     6     1     2
 4     7     2     3
 5     8     2     2
 6     9     2     6
 7    10     1     6
 8    11     1     6
 9    12     0     2
10    13     1     6
```

Binomial logistic regression

```
glm(formula = cbind(cases, n - cases) ~ WAIS,
family = binomial,
    data = swdtsum)
            coef.est coef.se
(Intercept)  2.40     1.19
WAIS        -0.32     0.11
---
  n = 17, k = 2
  residual deviance = 9.4, null deviance = 20.2
(difference = 10.8)
```

Compare the result

- Logistic

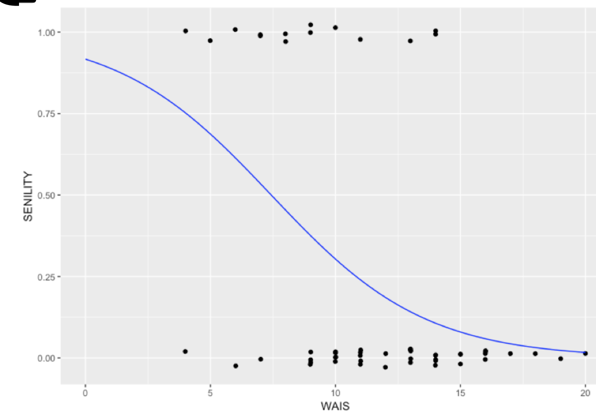
```
glm(formula = SENILITY ~ WAIS, family = binomial,
data = swdt)
              coef.est coef.se
(Intercept)    2.40     1.19
WAIS          -0.32     0.11
---
  n = 54, k = 2
  residual deviance = 51.0, null deviance = 61.8
(difference = 10.8)
```

- Binomial

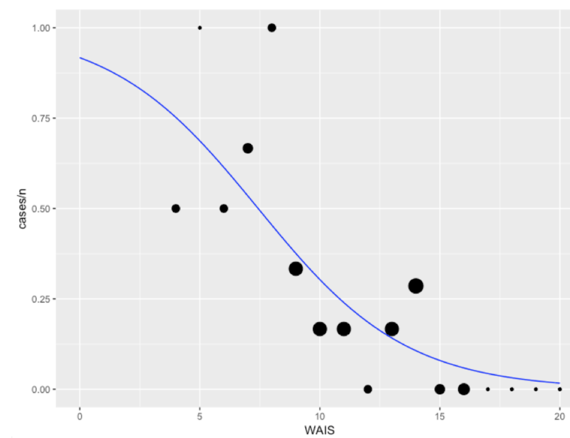
```
glm(formula = cbind(cases, n - cases) ~ WAIS, family
= binomial, data = swdtsum)
              coef.est coef.se
(Intercept)    2.40     1.19
WAIS          -0.32     0.11
---
  n = 17, k = 2
  residual deviance = 9.4, null deviance = 20.2
(difference = 10.8)
```

Displaying the fitted curve

```
glm(formula = SENILITY ~ WAIS, family = binomial,
data = swdt)
             coef.est coef.se
(Intercept)  2.40     1.19
WAIS        -0.32     0.11
---
  n = 54, k = 2
  residual deviance = 51.0, null deviance = 61.8
(difference = 10.8)
```




```
glm(formula = cbind(cases, n - cases) ~ WAIS, family
= binomial, data = swdtsum)
             coef.est coef.se
(Intercept)  2.40     1.19
WAIS        -0.32     0.11
---
  n = 17, k = 2
  residual deviance = 9.4, null deviance = 20.2
(difference = 10.8)
```



Chisquare goodness of fit test

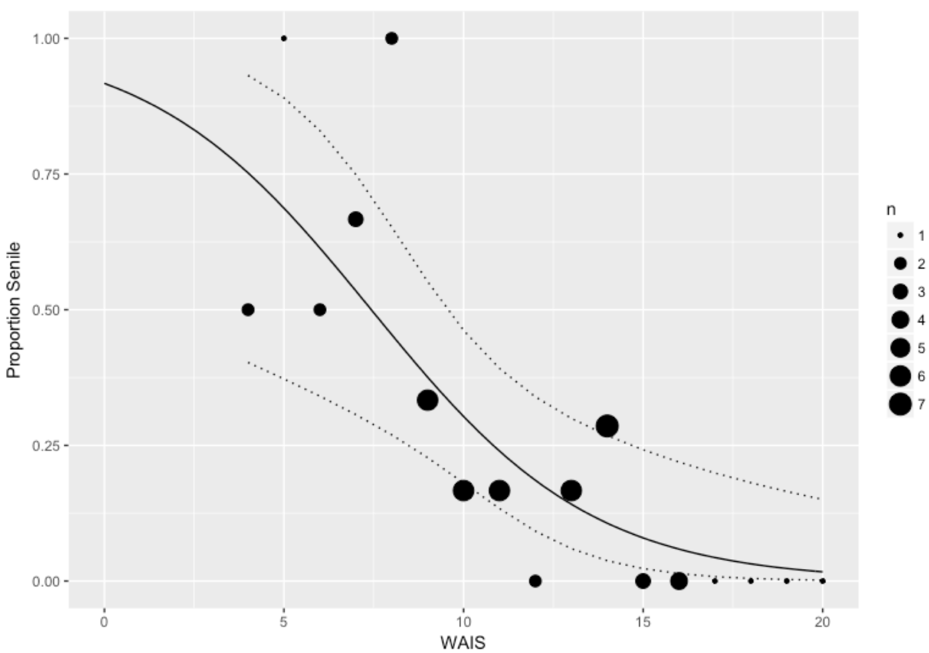
- Should be identical for the two

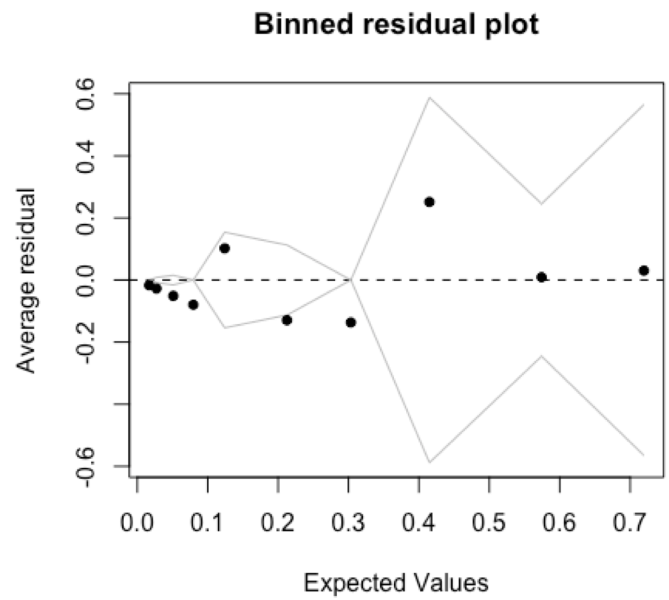
```
> with(bernlogitfit, pchisq(null.deviance - deviance,
df.null - df.residual, lower.tail = FALSE))
[1] 0.001021086
> with(binlogitfit, pchisq(null.deviance - deviance,
df.null - df.residual, lower.tail = FALSE))
[1] 0.001021086
```

AIC

```
> AIC(bernlogitfit)
[1] 55.01738
> AIC(binlogitfit)
[1] 27.79186
```

Checking the fit

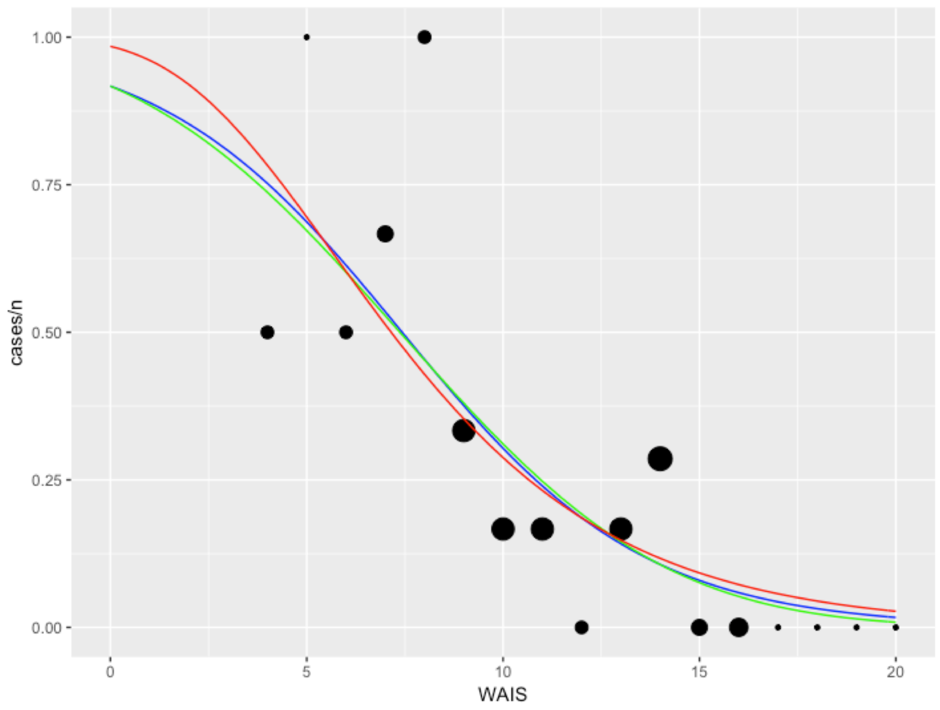




Different link functions

- Blue logistic
- Green: probit
- Red: c-log-log

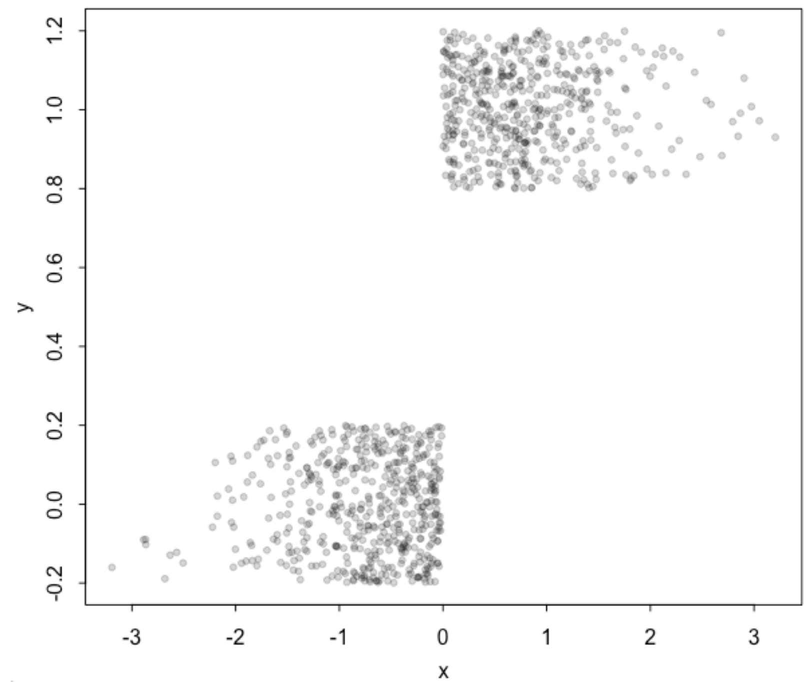
```
> AIC(binlogitfit)
[1] 27.79186
> AIC(binprobitfit)
[1] 27.75807
> AIC(bincloglogfit)
[1] 28.08519
```



Be careful with the warnings!!

- When you have linear separation

```
> x<-rnorm(1000)
> y<-1*(x>0)
>
> binlogitfit<-glm(y~x,
data=data.frame(x,y),family=binomial)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities
numerically 0 or 1 occurred
```



```
> summary(binlogitfit)

Call:
glm(formula = y ~ x, family = binomial, data = data.frame(x, y))

Deviance Residuals:
      Min          1Q      Median          3Q         Max
-0.003301   0.000000   0.000000   0.000000   0.003318

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    10.47     241.68   0.043    0.965
x            3081.20   46607.71   0.066    0.947

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.3862e+03  on 999  degrees of freedom
Residual deviance: 2.1910e-05  on 998  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

|Logistic regression coefficient| > 5 means

- One unit difference can make the probability go from 50:50 to 99%

```
> invlogit(0)
[1] 0.5
> invlogit(0+5)
[1] 0.9933071
> invlogit(0-5)
[1] 0.006692851
```

- If you have such a predictor, why bother with logistic regression?

Logistic or Poisson

- Number of elderly person that was senile out of particular study group.
- Number of senile elderly person that was counted in elderly care facility.

- Number of mouse that survived in toxicity experiment.
- Number of mouse that was counted dead after exterminator sprayed pesticides.

- Number of successful mating after certain number of trials for elephants.
- Number of mating for an elephant.

Poisson Regression