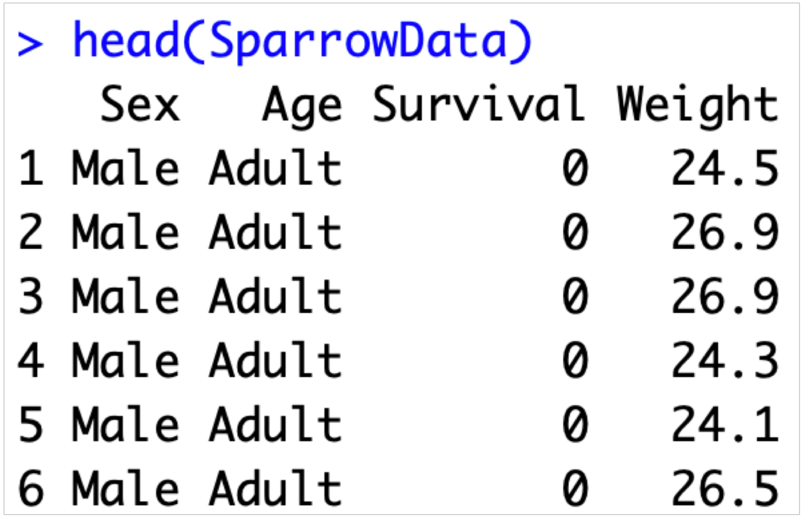
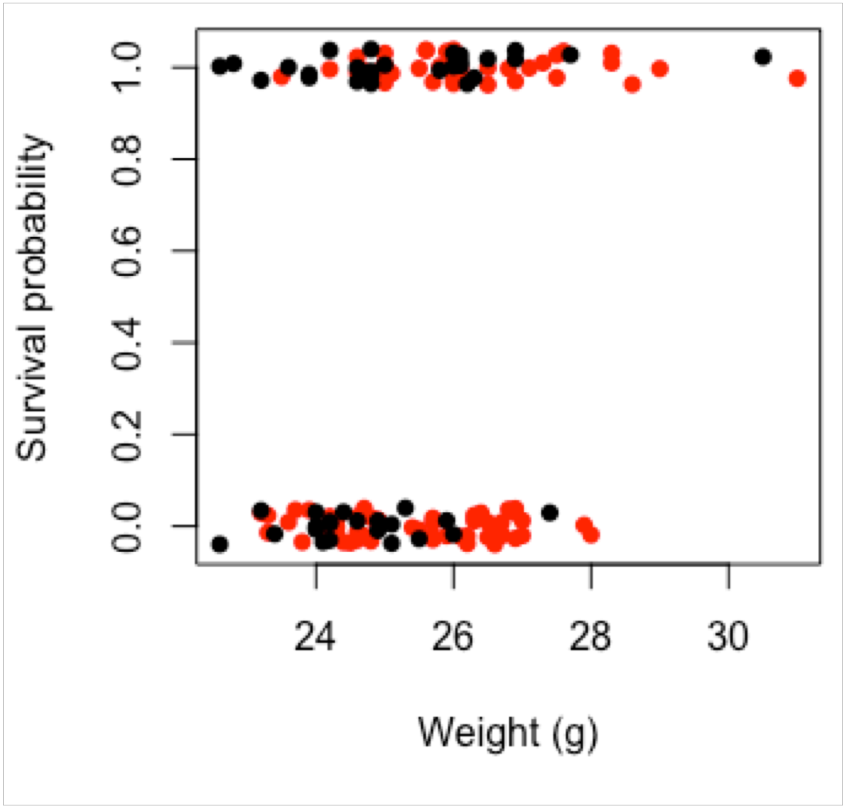
Data

Sex = Male and Female
Age = Adult and Juvenile
Survival = 0 and 1
Weight = g



```
> head(SparrowData)
   Sex   Age Survival Weight
1 Male Adult        0   24.5
2 Male Adult        0   26.9
3 Male Adult        0   26.9
4 Male Adult        0   24.3
5 Male Adult        0   24.1
6 Male Adult        0   26.5
```



Exercise 1: Key things to consider

- You have been presented with some data (or in reality you might have collected it)

- You now want to decide how to model it

- Things to think about:

- What is your biological question?
- What kind of data do you have: is it continuous or categorical? which is the response? is it counts?
- Will the data be normal?

- **See if you can answer all of these for today's data**

Exercise 1: ANSWER

What is your biological question?
There is no single correct answer, but should be something related to “What influences probability of survival in sparrows?” I chose: **“Does body weight and sex influence survival in sparrows?”,** you could also have a question relating to body weight e.g. “Does sex or age influence body weight in sparrows?”

What kind of data do you have: is it continuous or categorical? which is the response? is it counts?
Here we need to classify ALL variables: sex and age are categorical, survival is binary but could be considered categorical, weight is continuous. The response for me is survival. The only other option is weight. But sex and age cannot be caused by any of the others. None are counts.

Will the data be normal? No, it will follow a binomial distribution (survival), weight would be normal.

Exercise 2: Which model?

- Based on your answers in EX1, which model would you use for this week's data?

- What are you trying to find out?

- Why have you chosen this model? What are the parameters you will estimate with this model?

- How would you run this model in R? (one line of code)

Exercise 2: ANSWER

Based on your answers in EX1, which model would you use for this week's data?
Will continue assuming chose a question with survival as a response. **Then I would choose a binomial GLM with a logit link.**

What are you trying to find out? Whether sex and weight influence survival probability. So whether there is a difference in survival between two sexes and whether weight has a relationship with survival probability.

Why have you chosen this model? What are the parameters you will estimate with this model? I have chosen this model because it is the one I feel should represent how the data were generated. As it is binary they should come from a binomial distribution. The key parameters we will estimate are α and β from the following equation :

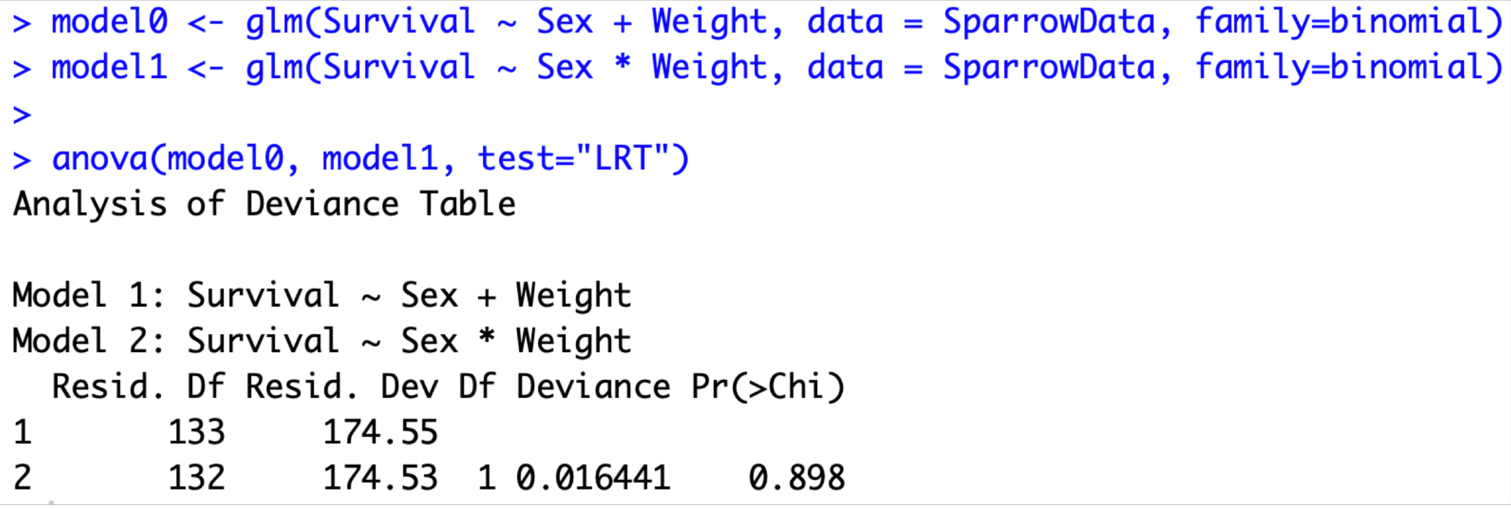
$$Y_i = \frac{1}{1 + e^{-(\alpha + \beta_{sex} X_{sex,i} + \beta_{weight} X_{weight,i})}}$$

β_{sex} represents the difference in intercept (α) caused by sex, β_{weight} represents the slope of the relationship between weight and survival (here the log odds of survival because of the link function).

How would you run this model in R? (one line of code)

```
glm(Survival ~ Sex + Weight, data = SparrowData, family = “binomial”(link=logit))
```

Exercise 3a: Interpreting

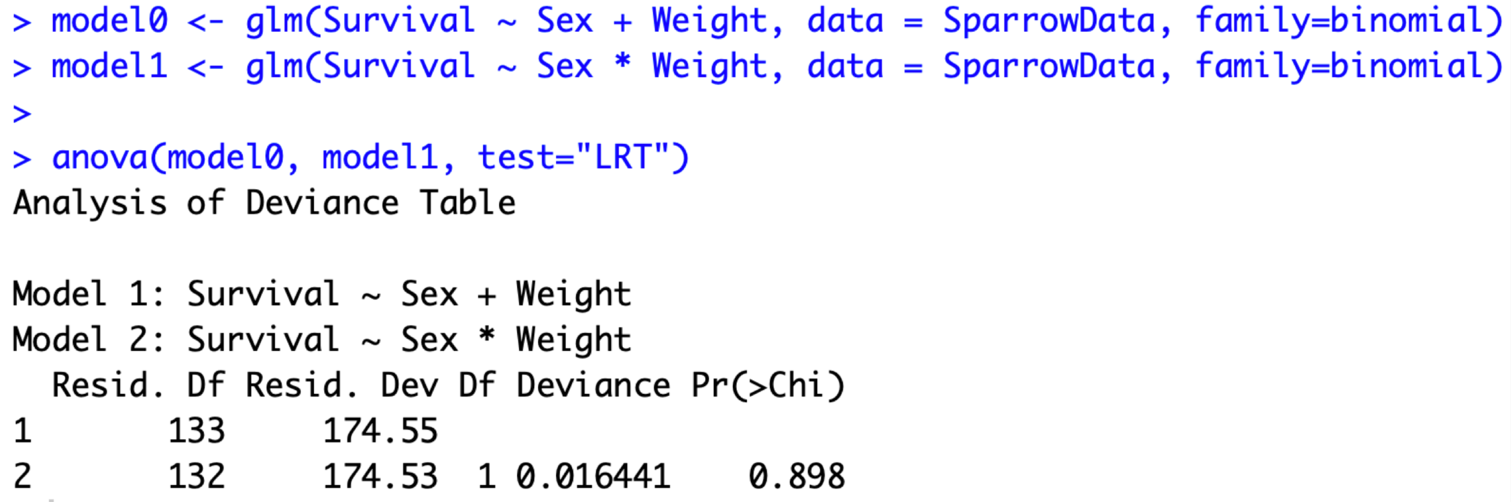


```
> model0 <- glm(Survival ~ Sex + Weight, data = SparrowData, family=binomial)
> model1 <- glm(Survival ~ Sex * Weight, data = SparrowData, family=binomial)
>
> anova(model0, model1, test="LRT")
Analysis of Deviance Table

Model 1: Survival ~ Sex + Weight
Model 2: Survival ~ Sex * Weight
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       133     174.55
2       132     174.53  1 0.016441    0.898
```

- Above is an output from R.
- What analysis has been conducted?
- What does the analysis aim to find out?
- What can you conclude from this output?

Exercise 3a: ANSWER



```
> model0 <- glm(Survival ~ Sex + Weight, data = SparrowData, family=binomial)
> model1 <- glm(Survival ~ Sex * Weight, data = SparrowData, family=binomial)
>
> anova(model0, model1, test="LRT")
Analysis of Deviance Table

Model 1: Survival ~ Sex + Weight
Model 2: Survival ~ Sex * Weight
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       133     174.55
2       132     174.53  1 0.016441    0.898
```

Above is an output from R.

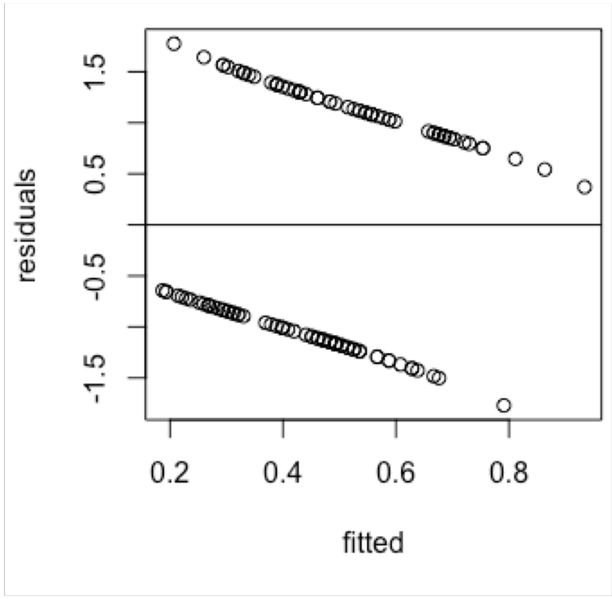
What analysis has been conducted? Confirmatory model selection using an analysis of deviance

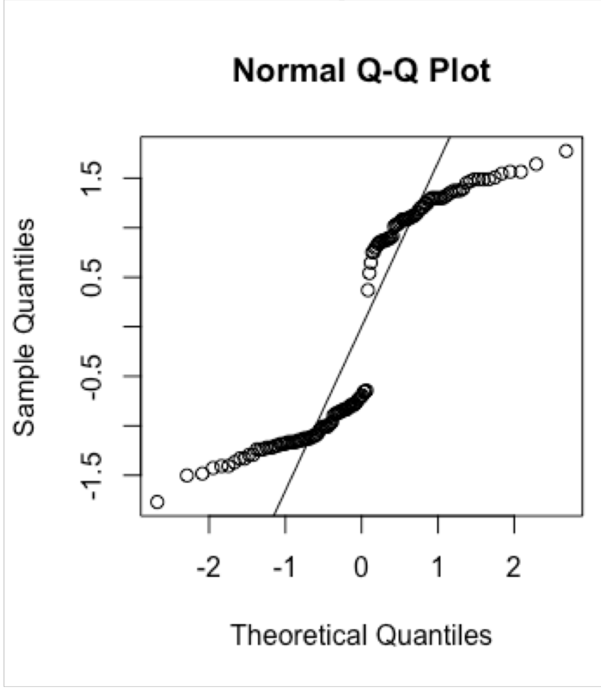
What does the analysis aim to find out? It tests the hypothesis that there is an interaction between weight and sex

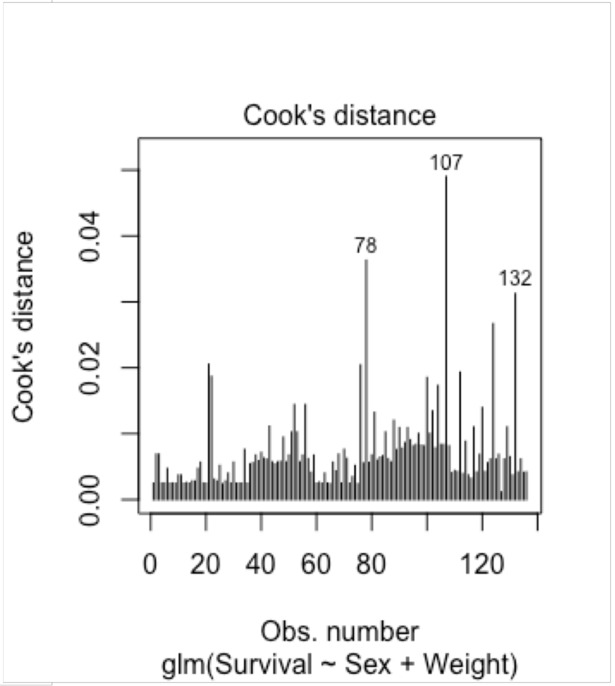
What can you conclude from this output? The probability Pr(>Chi) value for our test statistic (deviance) suggests we have a 90% chance of seeing our deviance or higher, if the null hypothesis was true. **Therefore we do not reject the null**, as we are very likely to see our result if the null is true. (null = no interaction)

Exercise 3b: Interpreting

- Below are some model fitting plots for model0 (previous slides)
- What do you think of the fit of this model? (include statement of what each plot tests AND what you think of it)







Exercise 3b: ANSWER

Below are some model fitting plots for model0 (previous slides)

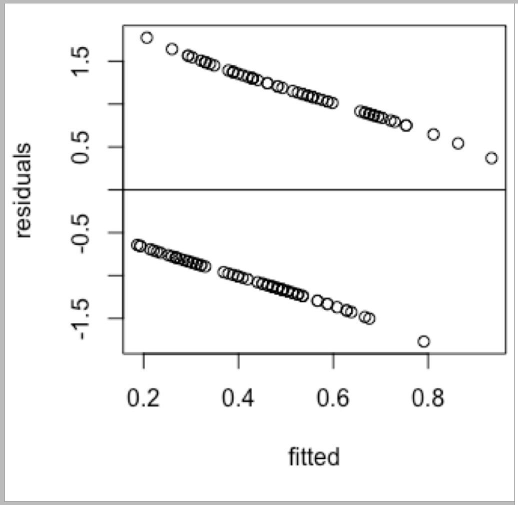
What do you think of the fit of this model? (include statement of what each plot tests AND what you think of it)

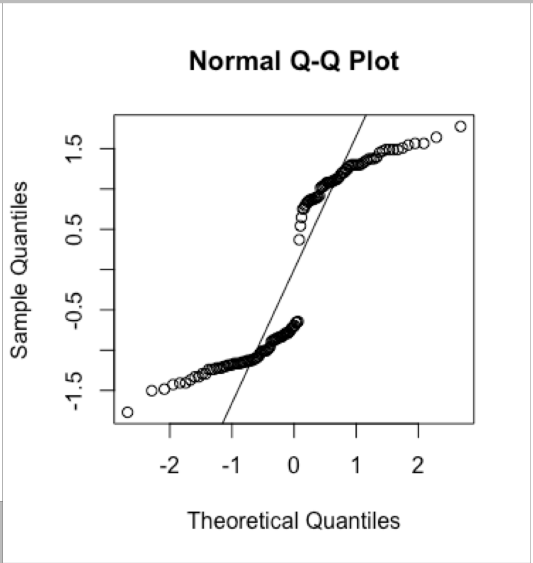
The residuals vs fitted tests equal variance and linearity. It is hard to assess these from this plot. They all look bad! But possibly the variance does remain equal.

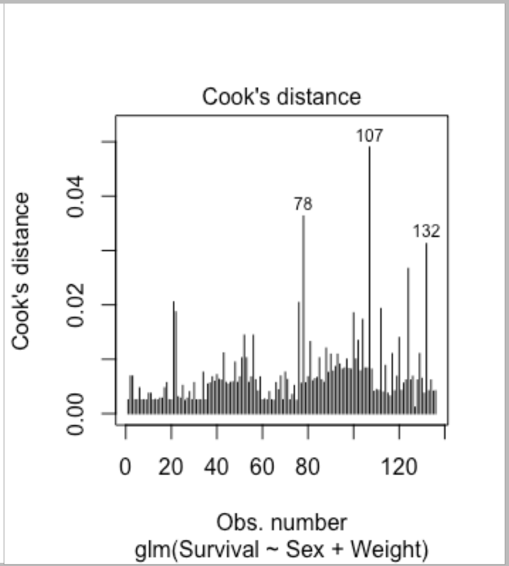
The normal QQ tests normality of the residuals. We would not expect it to be perfect but because we use deviance residuals it could be close. I seems to roughly follow a normal but with skew at high and low theoretical quantiles.

Cook’s distance tests for outliers, it has identified 3 but the distance they produce is very low.

Overall the fit is not great, but it seems to be roughly ok. We might want to improve normality but as this is a non-normal model, we have a bit more tolerance for deviation from normality.





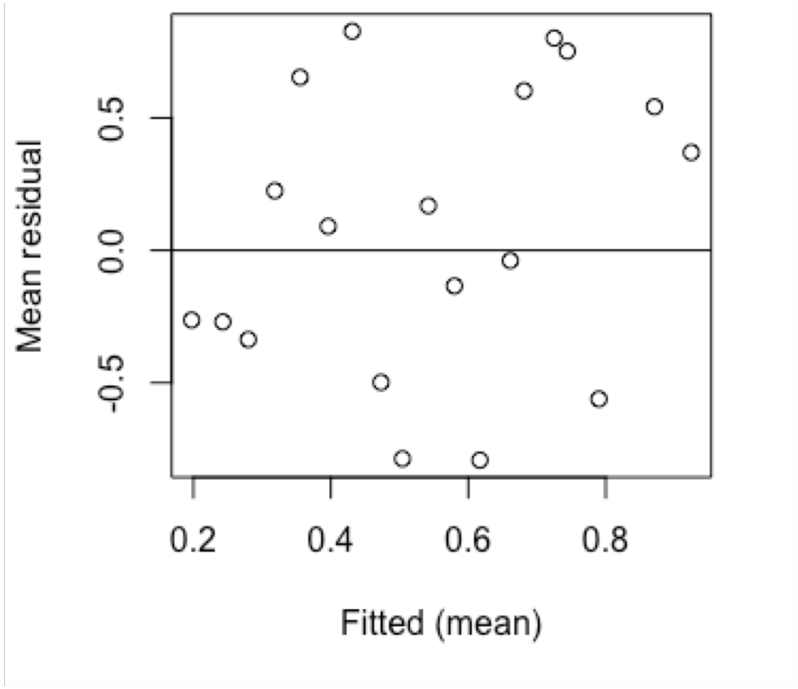


Exercise 3b: Interpreting

- Residual plots for binomial data can be very hard to interpret

- It can be easier to group the residuals and take mean values i.e. take all residuals for fitted values between 0.2 and 0.21 and take the mean

- This has been done below – does it change your interpretation of model fit?

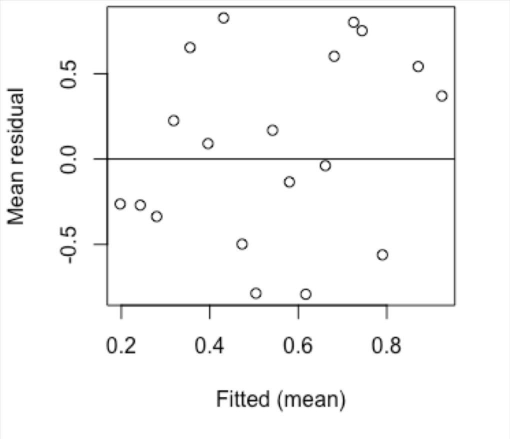


Exercise 3b: ANSWER

Residual plots for binomial data can be very hard to interpret

It can be easier to group the residuals and take mean values i.e. take all residuals for fitted values between 0.2 and 0.21 and take the mean

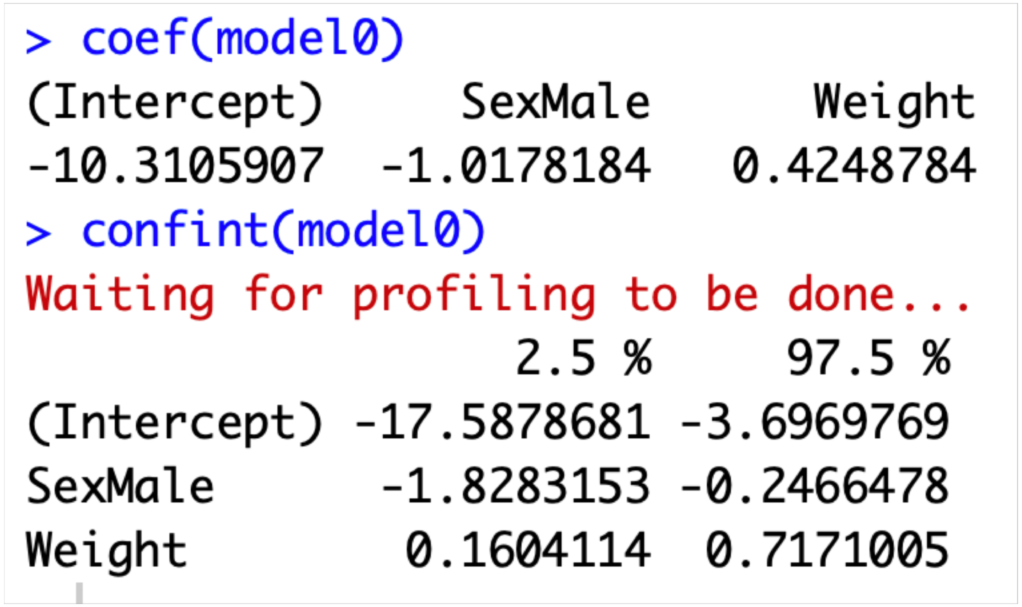
This has been done below – does it change your interpretation of model fit?
Now we can see a bit more clearly how the residuals change (on average) with fitted values. The variance seems equal across the fitted values. I am not more confidence that equal variance assumption is met.



Exercise 3b: Interpreting

- Now we have the output, both coefficients and confidence intervals for model0

- Interpret the output (work out what all of the numbers mean, then draw a biological conclusion)



```
> coef(model0)
(Intercept)      SexMale       Weight 
-10.3105907   -1.0178184    0.4248784 
> confint(model0)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -17.5878681 -3.6969769
SexMale      -1.8283153 -0.2466478
Weight        0.1604114  0.7171005
```

$$\mu = \log(\frac{p}{1-p})$$

The inverse is

$$p = \frac{e^{\mu}}{1+e^{\mu}}$$

Exercise 3b: ANSWER

Interpret the output (work out what all of the numbers mean, then draw a biological conclusion)

From the output below, we can see the coefficient (parameter estimates) for the linear predictor on the logit scale. As we have one continuous (weight) and one categorical (sex) variable as explanatory variables and NO interaction, we know that we expect to get out a single slope value (for weight) and an intercept for females and a difference in intercept for males.

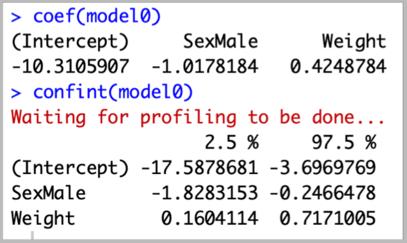
The intercept for females seems counter intuitive at -10.3 but it is a log odds so we need to use the inverse link to get back to survival probability ($\frac{e^{-10.3}}{1+e^{-10.3}}$ = 0.00003).

We can see the difference in intercept for males is -1 log odds or ($\frac{e^{-10.3-1}}{1+e^{-10.3-1}}$ = 0.00001 = **intercept males**) **Males have lower survival than females**

The effect of weight can also be seen to be positive at 0.42 log odds per g. Bigger birds have higher survival probability.

Both the effect of sex and weight have confidence intervals that do not cross 0, therefore we would be unlikely to see these effects if the null (no effect) was true. Even with uncertainty – we see the same direction of effect.

The biological reasons for this could be that bigger birds have greater reserves or are older so can better survive a disturbance (in this case a storm). Males having lower survival could also be explained by maybe they are more bold so more exposed. No single answer here, but anything biologically sensible!



```
> coef(model0)
(Intercept)     SexMale      Weight
-10.3105907  -1.0178184   0.4248784
> confint(model0)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -17.5878681 -3.6969769
SexMale      -1.8283153 -0.2466478
Weight        0.1604114  0.7171005
```

$$\mu = \log(\frac{p}{1-p})$$

The inverse is

$$p = \frac{e^{\mu}}{1+e^{\mu}}$$

HINTS

***Hints**

Think about what kind of data went into the model (particularly the explanatory variables)

What was the biological question?

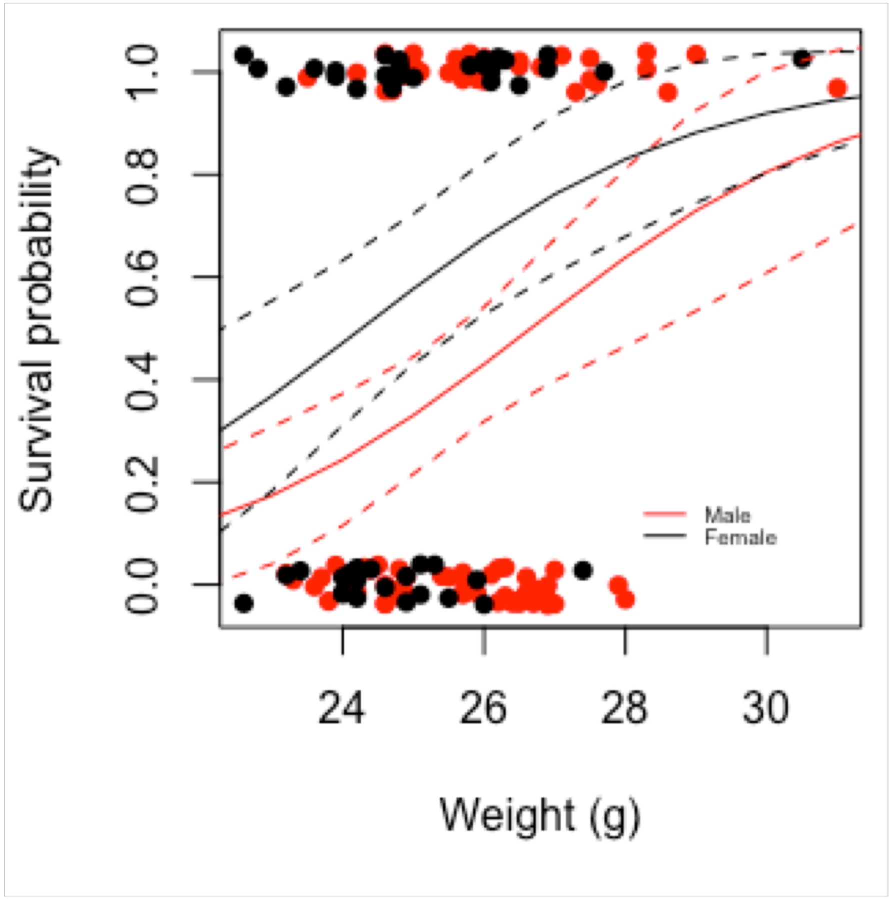
How do the coefficients fit the linear equation $\alpha + \beta X_i$?

Remember the link function

Exercise 4: Reading plots

- Below you have a plot of the results of model0

- What you can interpret from each of this? (include effect of Sex and Weight and the uncertainty)



Exercise 4: Reading plots

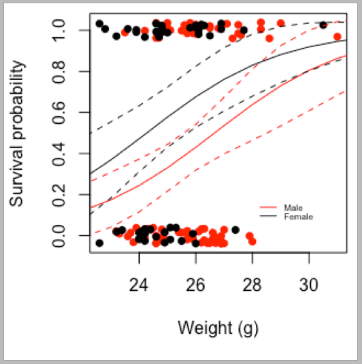
Below you have a plot of the results of model0

What you can interpret from this? (include effect of Sex and Weight and the uncertainty)

This plot should support the interpretation you did already from the coefficients. We can see that females indeed do have higher survival than males (the female predicted line is higher than males) and that survival probability increases with weight (positive slope/curve).

Both lines have the same slope, but because the line is curved they are not quite parallel. The uncertainty around these relationships is quite wide. There is also some overlap between the two sexes, especially as the uncertainty increases at higher and lower values of X (weight). Despite this uncertainty, we can still see clearly that males have lower survival probability and that survival probability increases with weight. The exact difference and exact slope are uncertain but the directions seem robust even with uncertainty.

But generally this shows the same thing we already discussed from the coefficients. Could note it is back on the original scale.



Binomial/logistic GLM: Part 2

Lecture Outline

Recap of yesterday

Introduction to the data

- EX1: Key things to consider
- EX2: Which model?
- EX3a: Interpreting output from a model

Mini-lecture 1 = Other links

Mini-lecture 2 = Categorical and continuous

Mini-lecture 3 = Overdispersion

- EX3b: Interpreting output from a model
- EX4: Reading plots

Exam style

Recap of yesterday

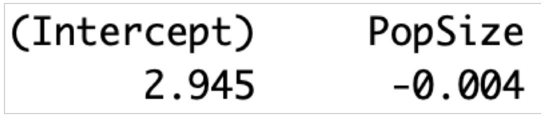
Logit link

$$\mu = \log(\frac{p}{1-p})$$

The inverse is

$$p = \frac{e^{\mu}}{1+e^{\mu}} = \frac{1}{1+e^{-\mu}}$$

Interpreting the logit



```
(Intercept)      PopSize
      2.945       -0.004
```

$$\alpha + \beta X_i$$

$$\frac{1}{1+e^{-\mu}} = p$$

E.g.

For X (PopSize) = 300

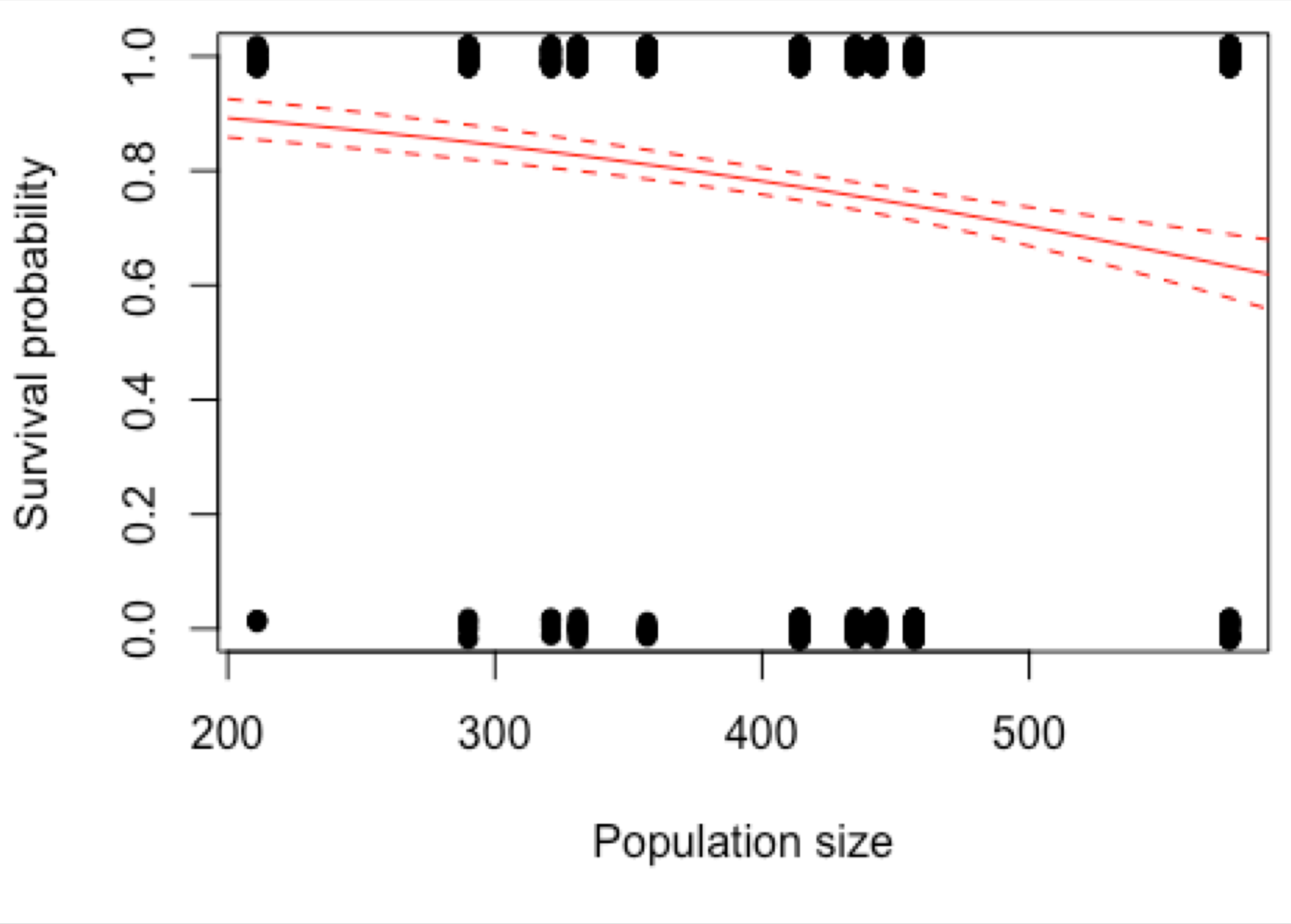
$$\frac{1}{1+e^{-(2.945+(-0.004*300))}} = 0.85$$

For X (PopSize) = 400

$$\frac{1}{1+e^{-(2.945+(-0.004*400))}} = 0.79$$

Interpreting the logit

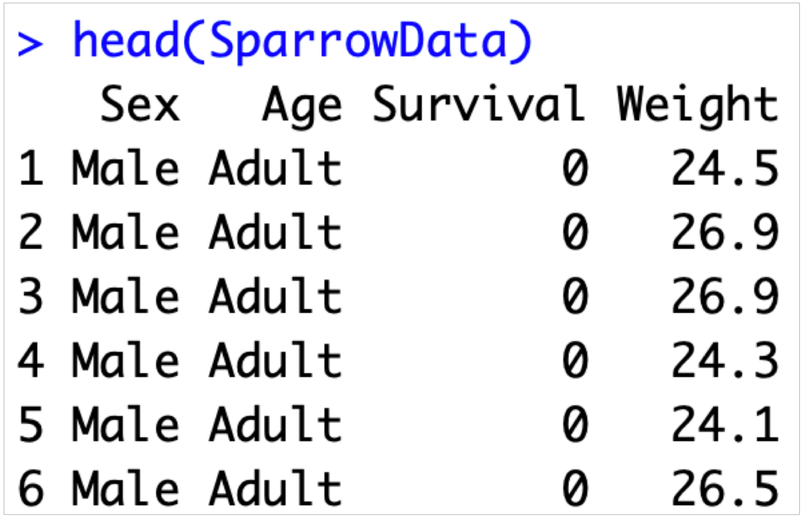
or plot it!



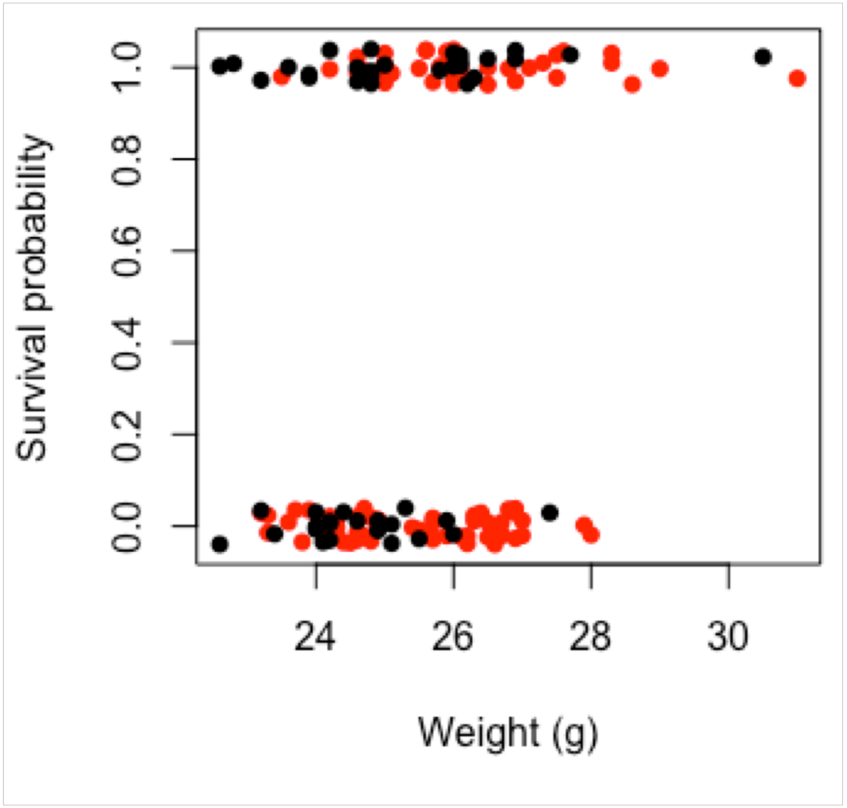
Introduction to the data

Data

Sex = Male and Female
Age = Adult and Juvenile
Survival = 0 and 1
Weight = g



```
> head(SparrowData)
   Sex   Age Survival Weight
1 Male Adult        0   24.5
2 Male Adult        0   26.9
3 Male Adult        0   26.9
4 Male Adult        0   24.3
5 Male Adult        0   24.1
6 Male Adult        0   26.5
```



Other link functions

Other links

So far we have used the logit link for Binomial GLM

This is the default (canonical) link in R

But you can use others too

Other links

So far we have used the logit link for Binomial GLM

This is the default (canonical) link in R

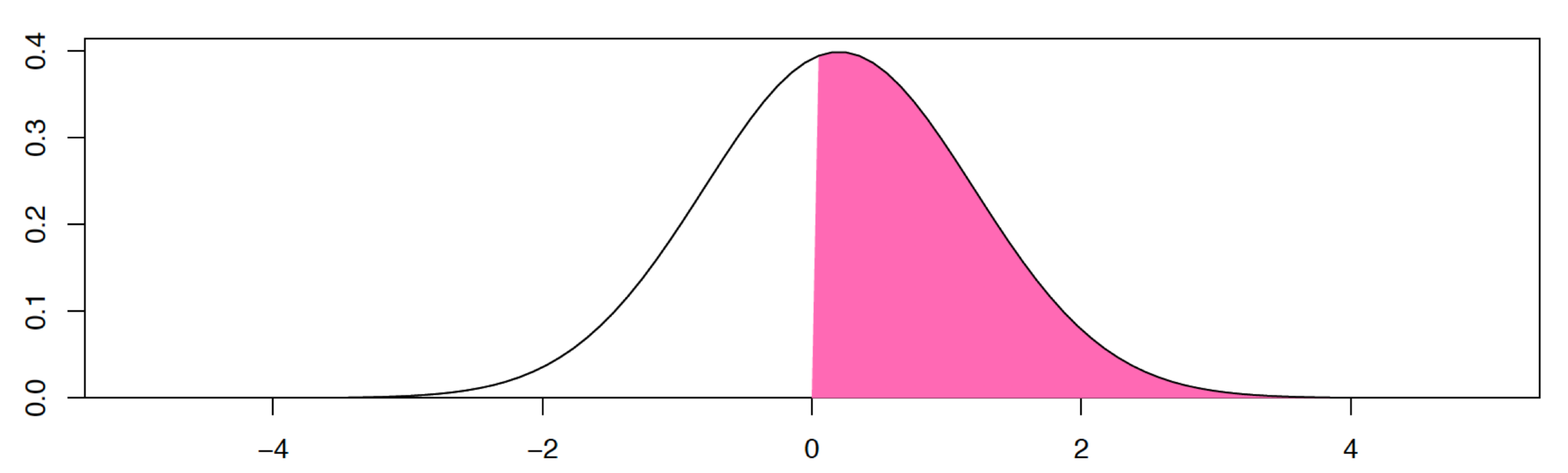
But you can use others too:

Probit

cloglog

Probit

Is a threshold model e.g. >0 = success, <0 = failure



Uses and inverse normal link function

Higher mean = higher probability of success

cloglog

Links to count data

Useful when 0 and 1 come from counts and you want to link to abundance

So when 0 and 1 really come from a Poisson distribution

$$\boxed{\log(\lambda)} = \boxed{\log(-\log(1-p))}$$

Poisson
(log link)

cloglog link
(Binomial)

Other links

Can all be used in Binomial GLM

Logit = default (most common)

Probit = can be easier to understand

cloglog = if you have Poisson-like counts

Other links

Can all be used in Binomial GLM

Logit = default (most common)

Probit = can be easier to understand

cloglog = if you have Poisson-like counts

Categorical and continuous

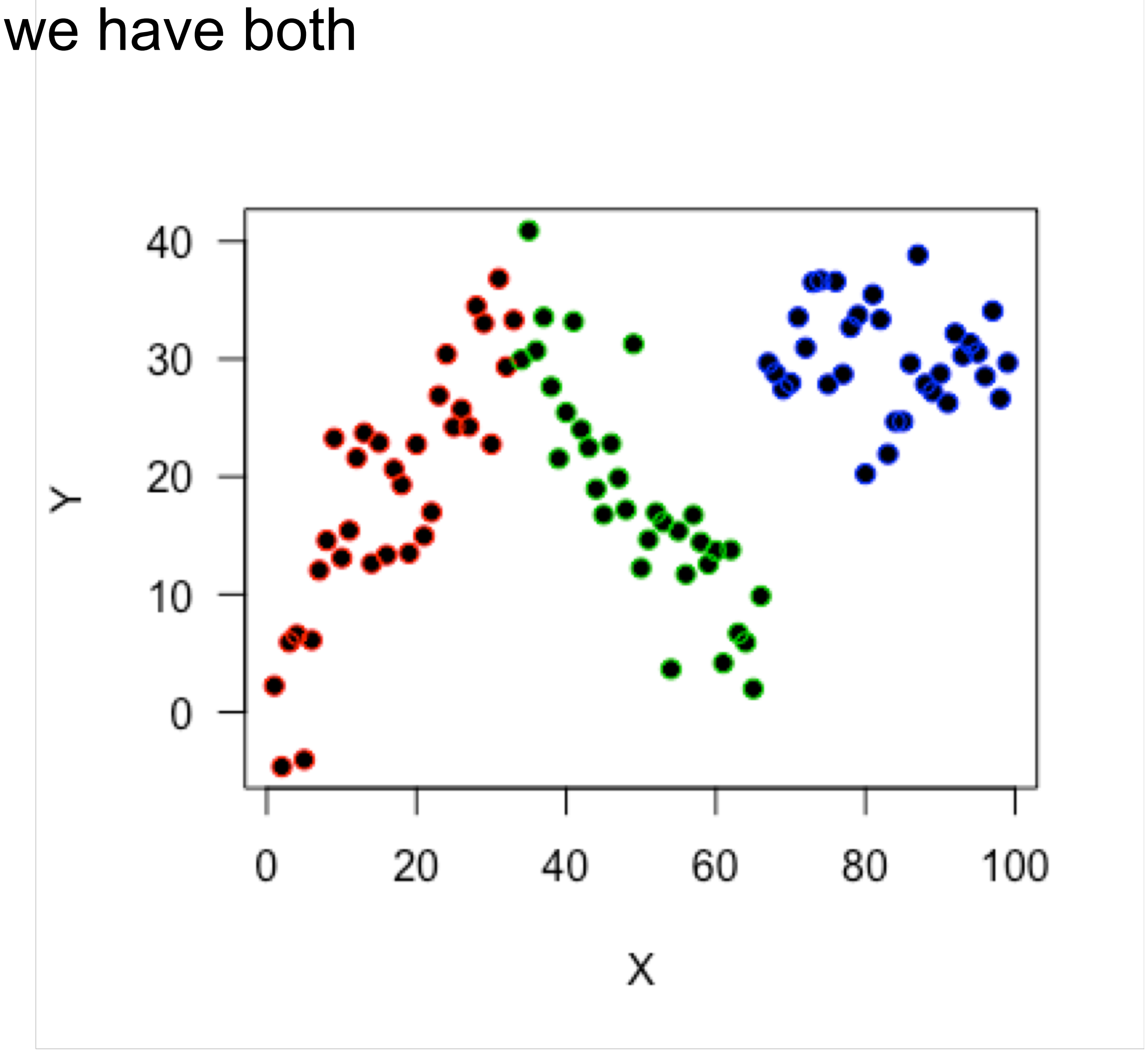
Definitions

Categorical = in groups

Continuous = every value can exist

When you combine them

Here we have both



When you combine them

Several ways we can model this

Y ~ X	Separately
Y ~ Groups	
Y ~ X + Groups	Additively
Y ~ X * Groups	Interaction

When you combine them

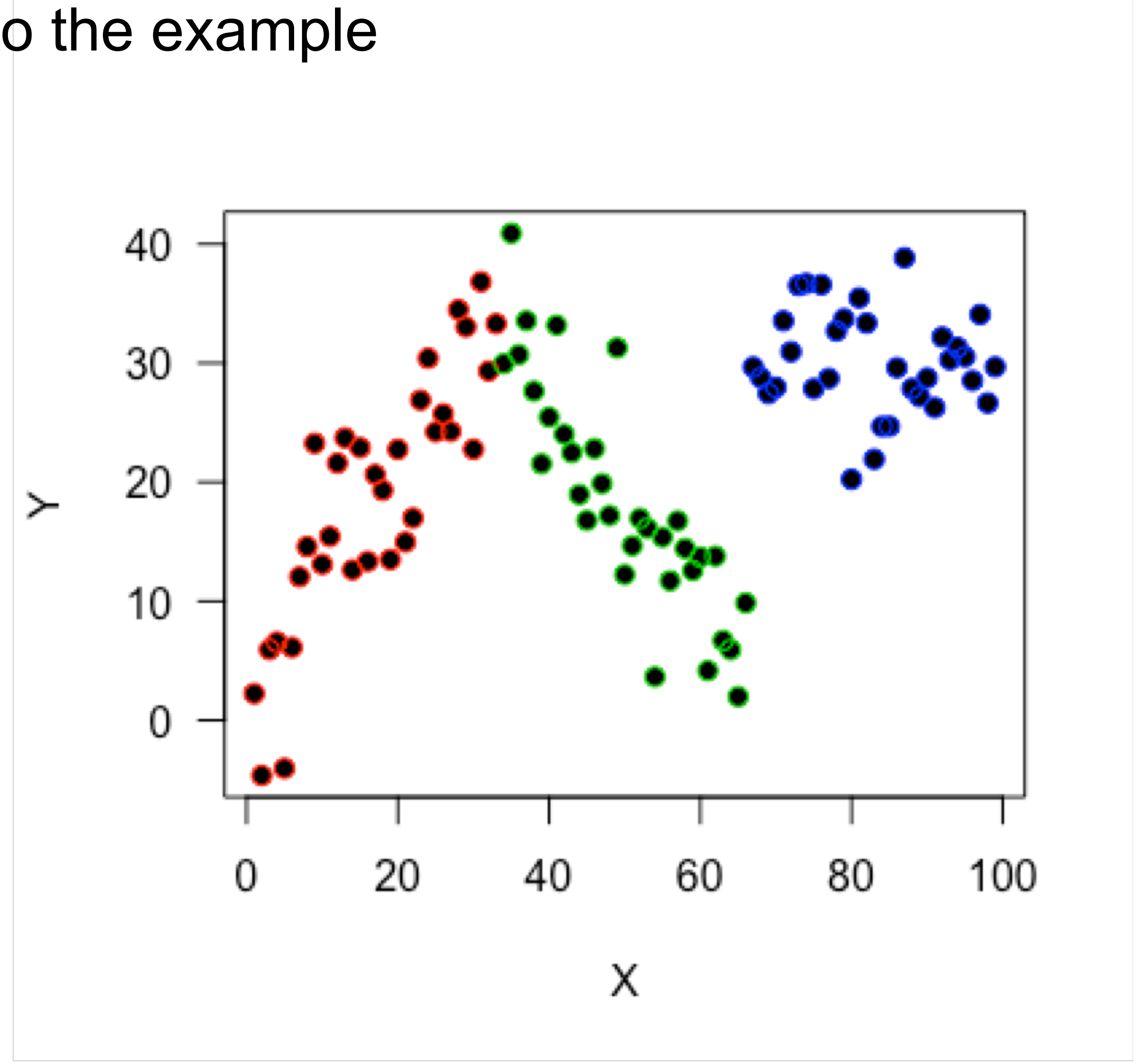
Several ways we can model this

Y ~ X	Separately
Y ~ Groups	
Y ~ X + Groups	Additively
Y ~ X * Groups	Interaction

Will depend on the effect of each

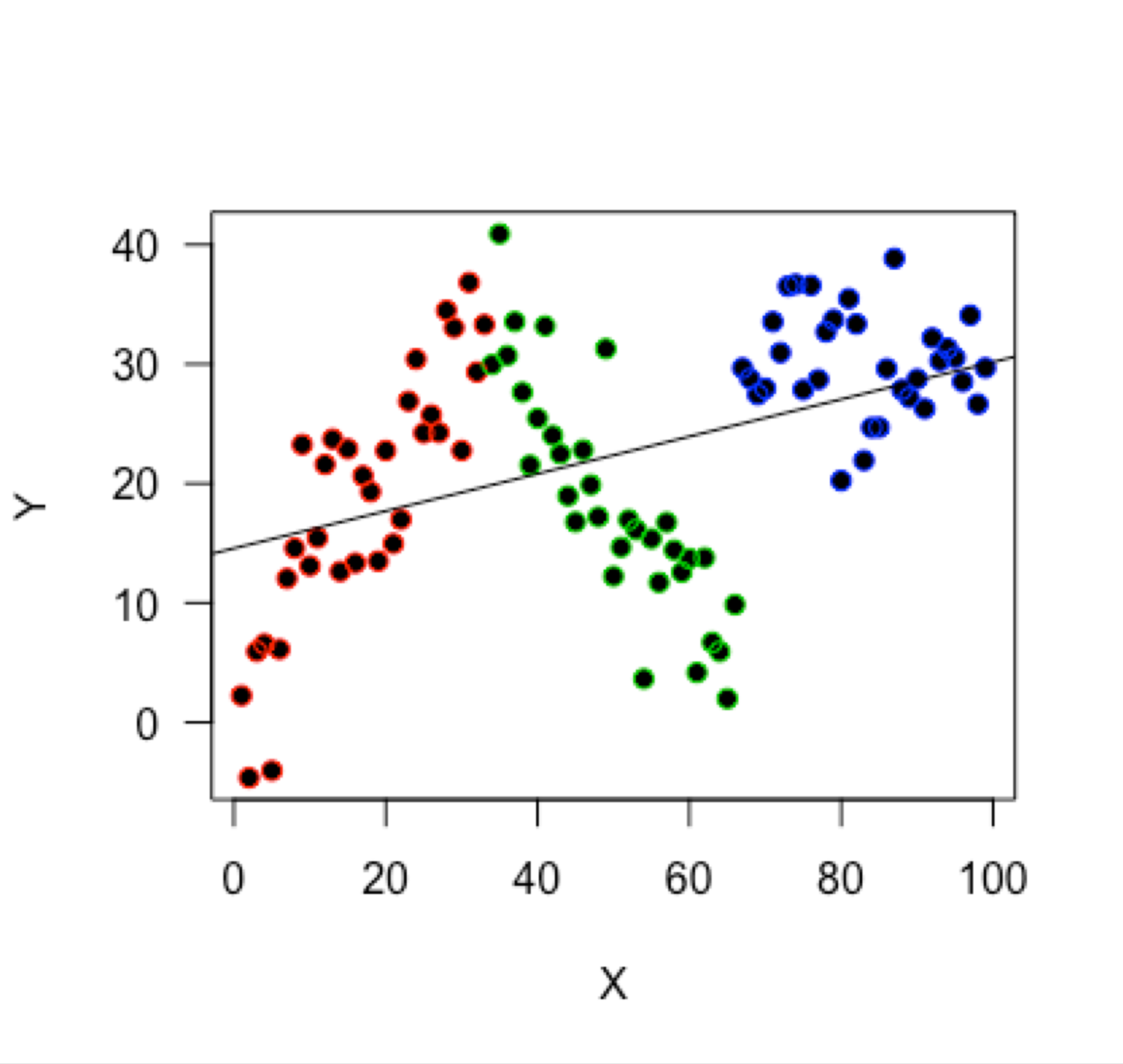
When you combine them

Back to the example

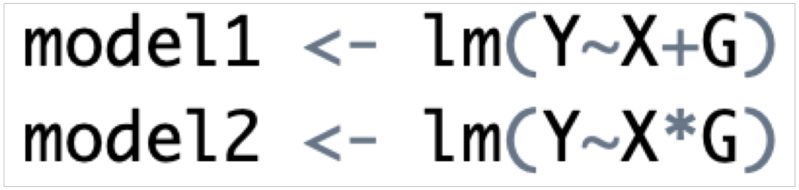


When you combine them

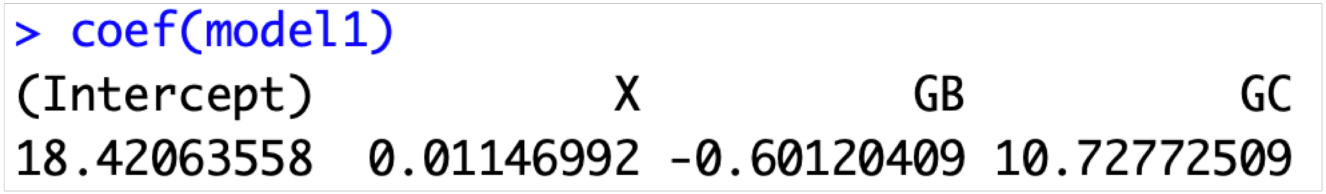
Back to the example



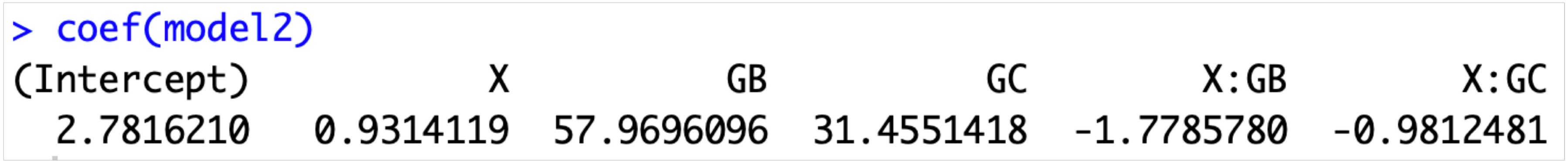
Interpreting



```
model1 <- lm(Y~X+G)
model2 <- lm(Y~X*G)
```

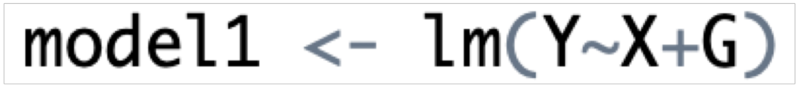


```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```

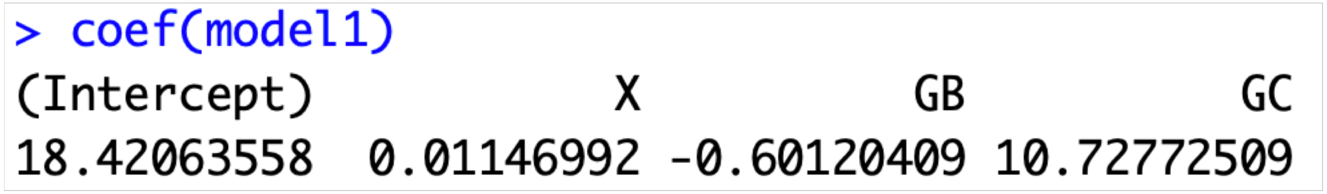


```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```

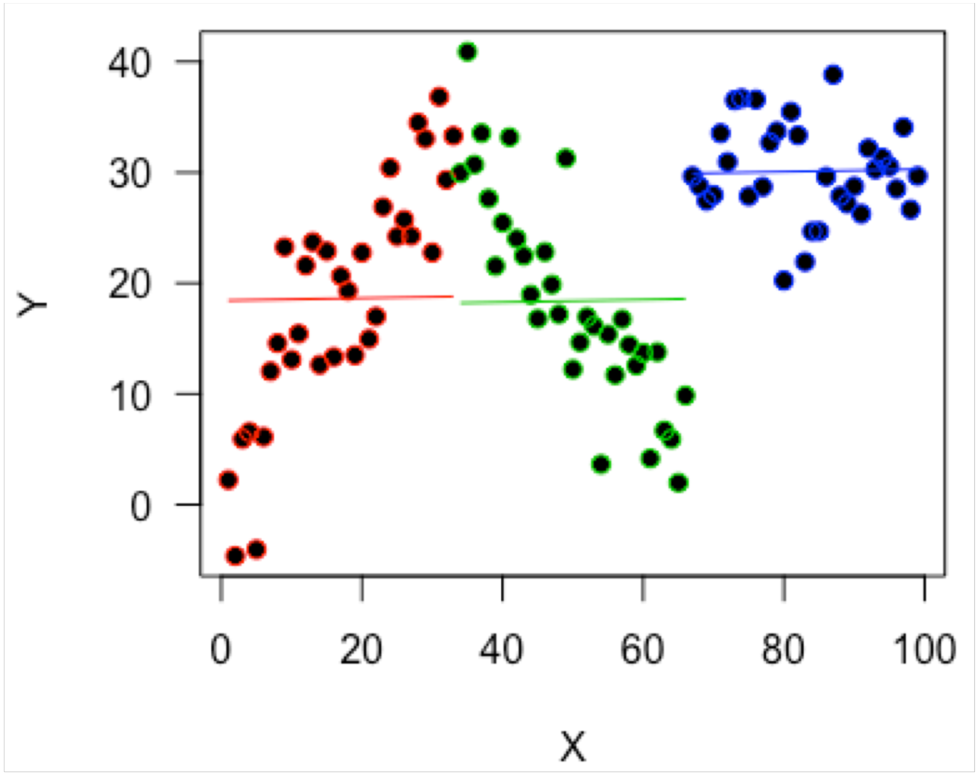
Interpreting



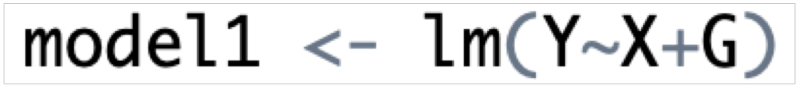
```
model1 <- lm(Y~X+G)
```



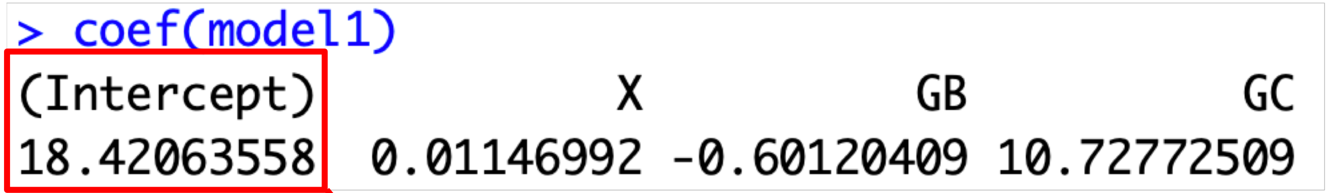
```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```



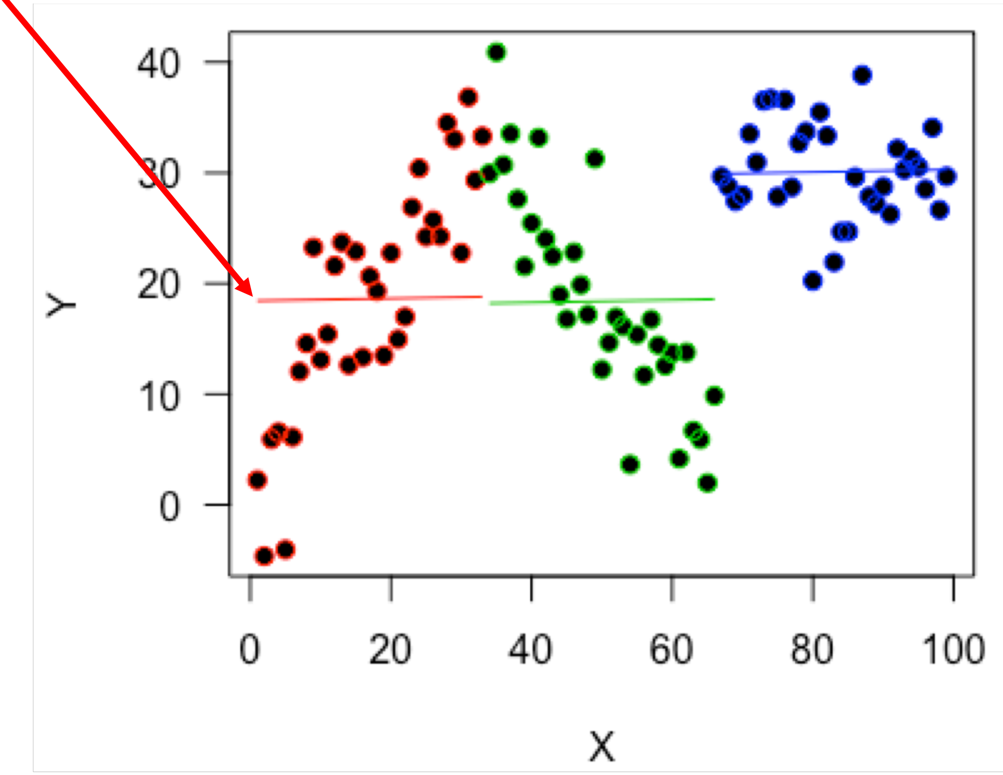
Interpreting



```
model1 <- lm(Y~X+G)
```

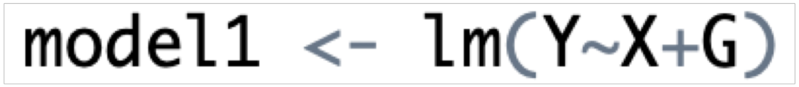


```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```

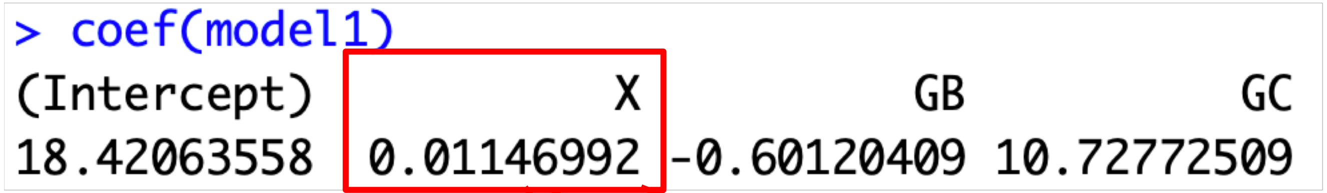


Intercept
of line of
Group A

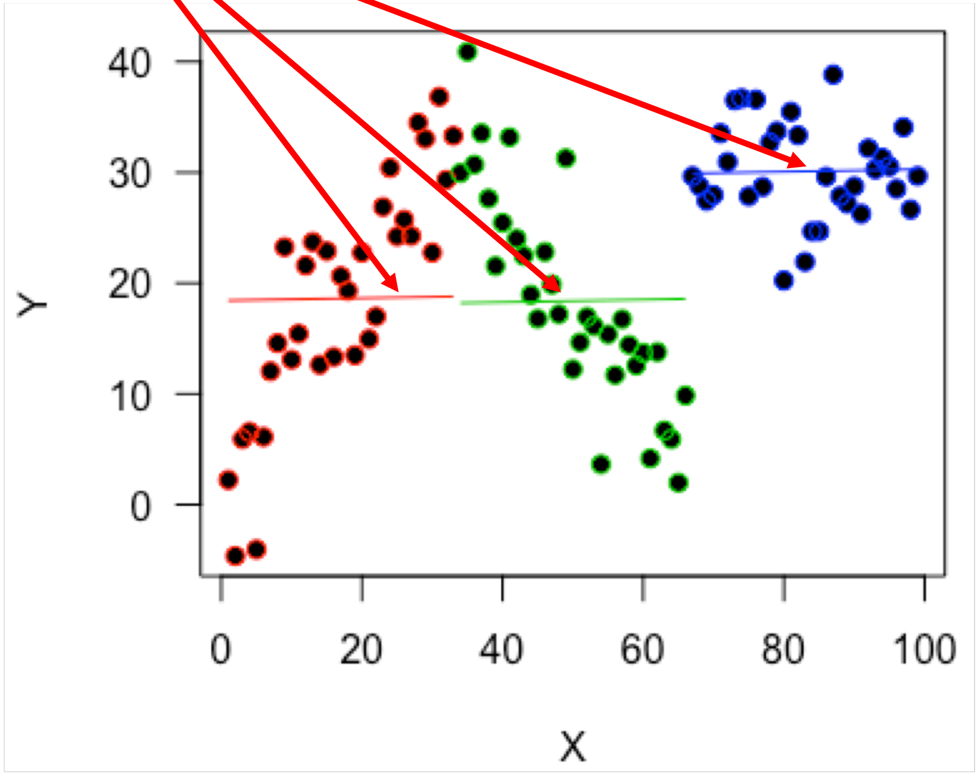
Interpreting



```
model1 <- lm(Y~X+G)
```

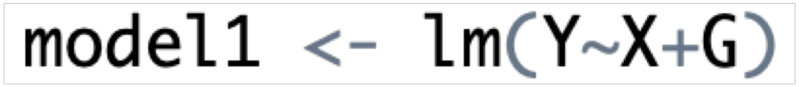


```
> coef(model1)
(Intercept)           X          GB          GC
18.42063558  0.01146992 -0.60120409 10.72772509
```

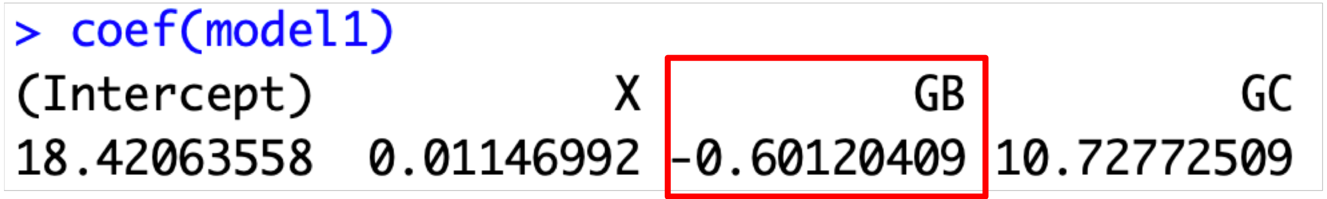


Slope value for all groups (same)

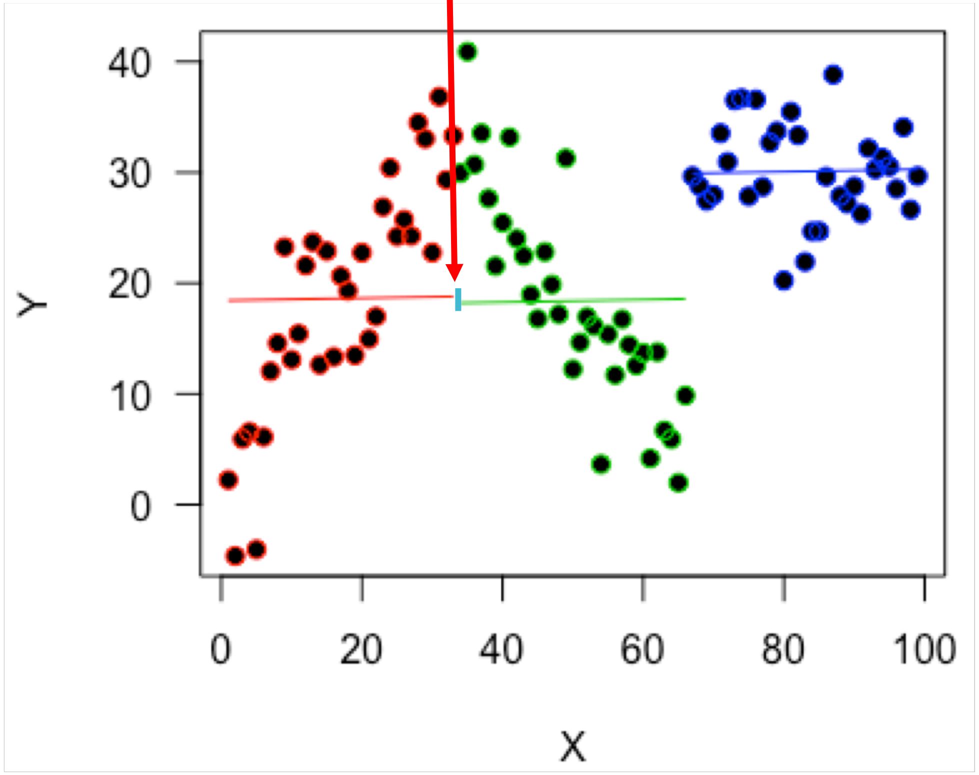
Interpreting



```
model1 <- lm(Y~X+G)
```

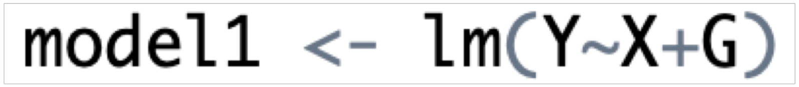


```
> coef(model1)
(Intercept)           X          GB          GC 
18.42063558  0.01146992 -0.60120409 10.72772509 
```

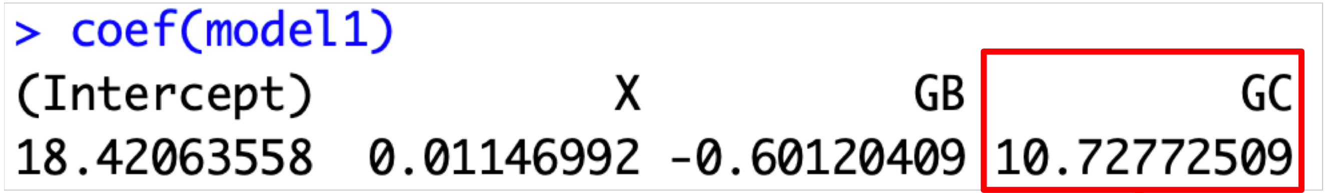


Difference in intercept from Group A to Group B

Interpreting

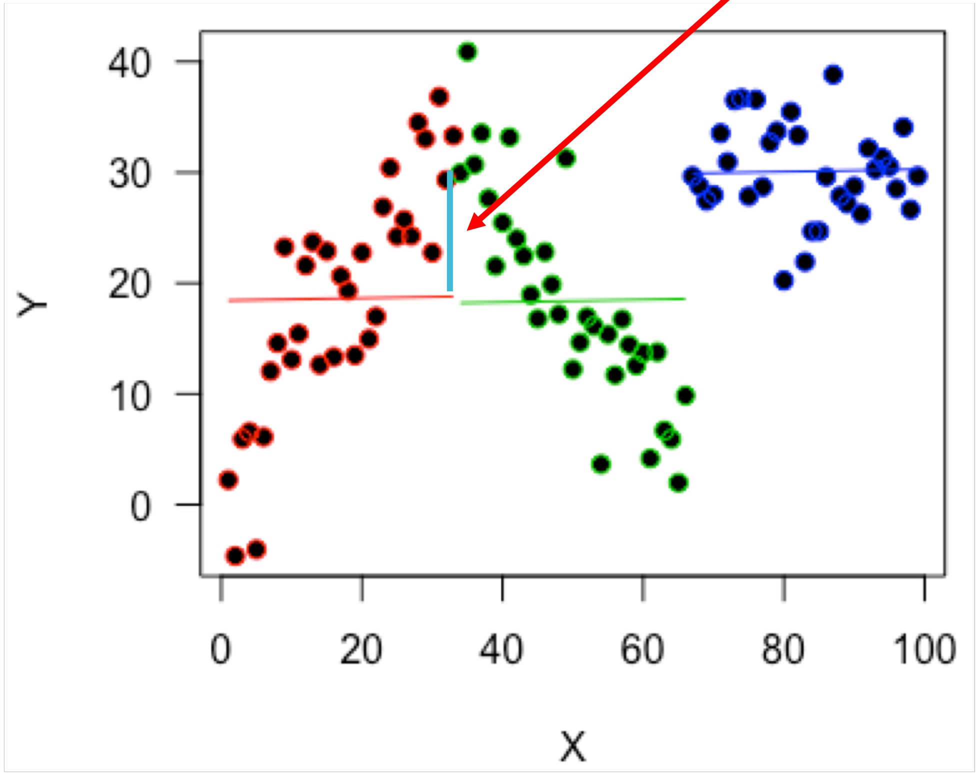


```
model1 <- lm(Y~X+G)
```

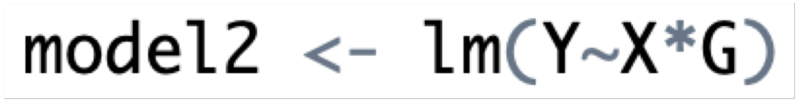


```
> coef(model1)
(Intercept)           X          GB          GC
18.42063558  0.01146992 -0.60120409 10.72772509
```

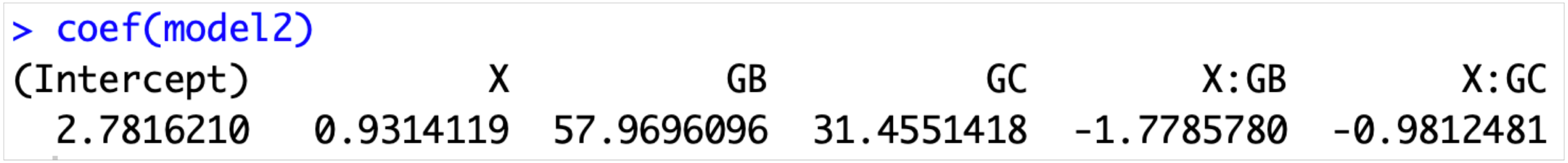
Difference in intercept from Group A to Group C



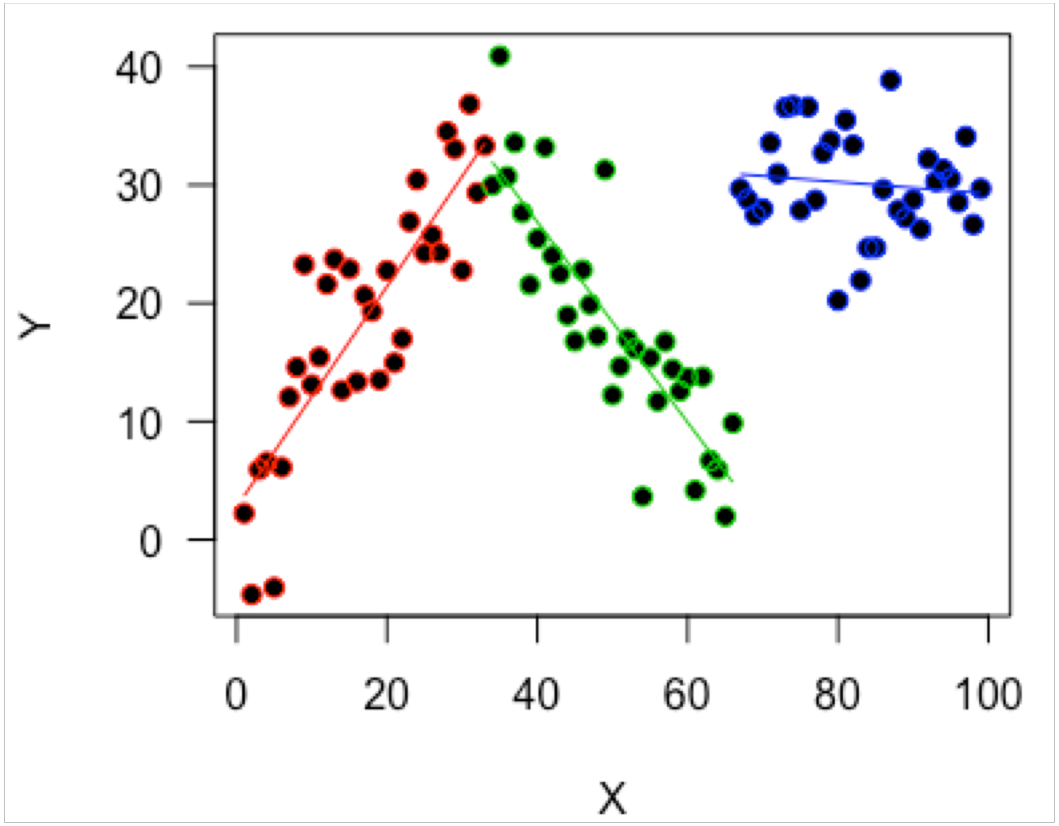
Interpreting



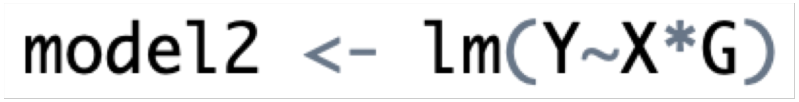
```
model2 <- lm(Y~X*G)
```



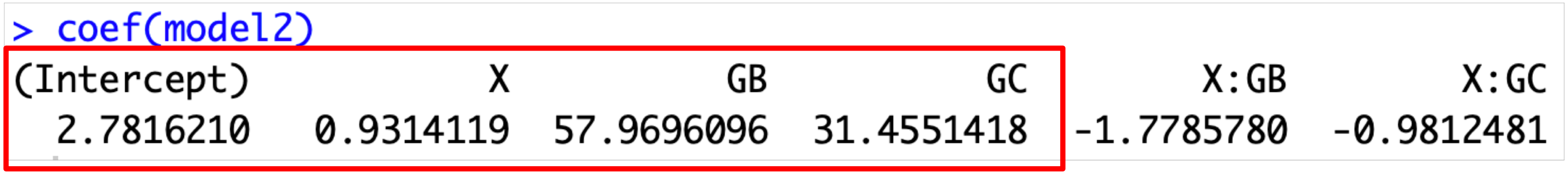
```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```



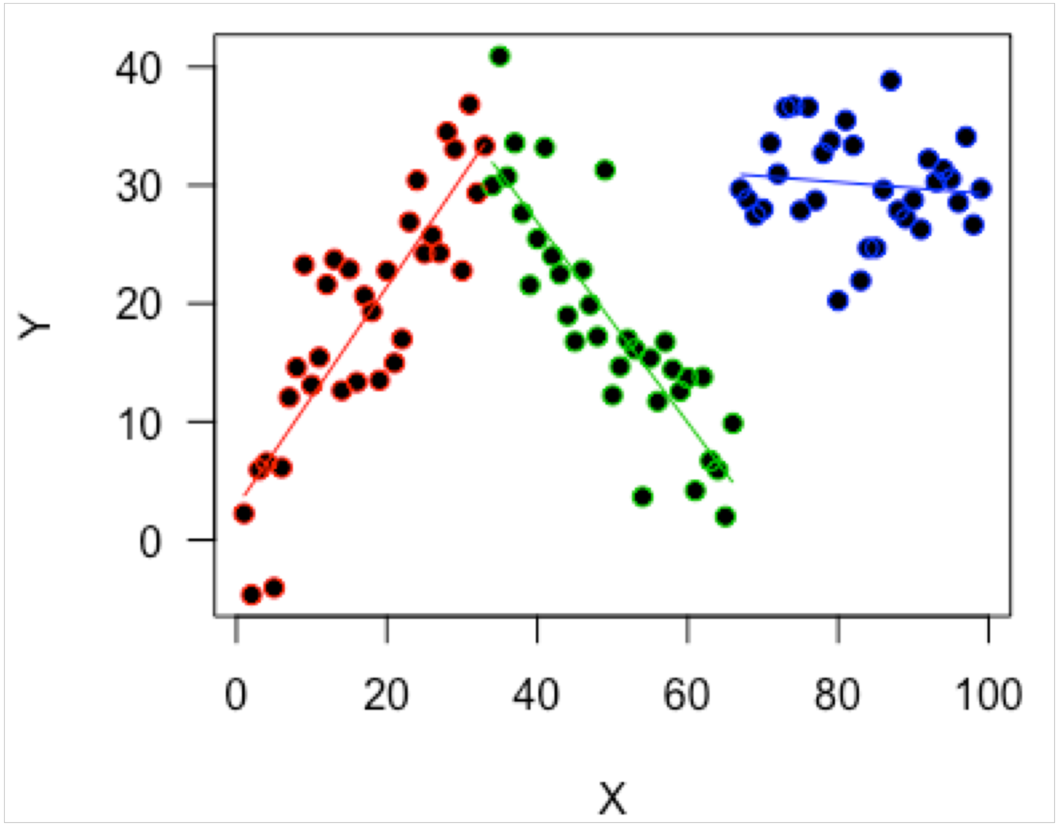
Interpreting



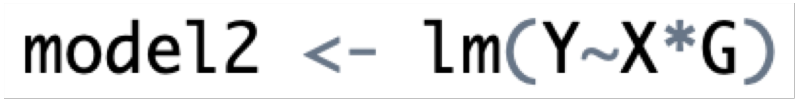
```
model2 <- lm(Y~X*G)
```



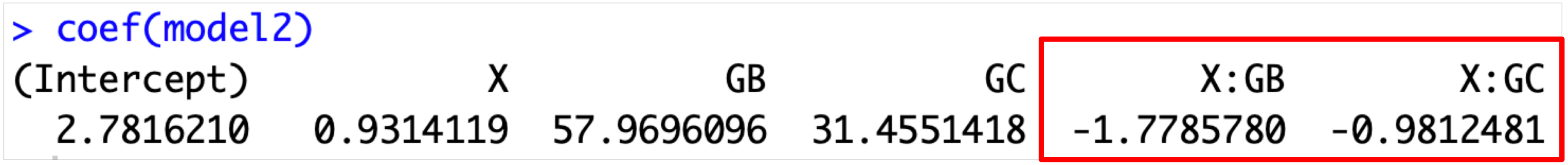
```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```



Interpreting



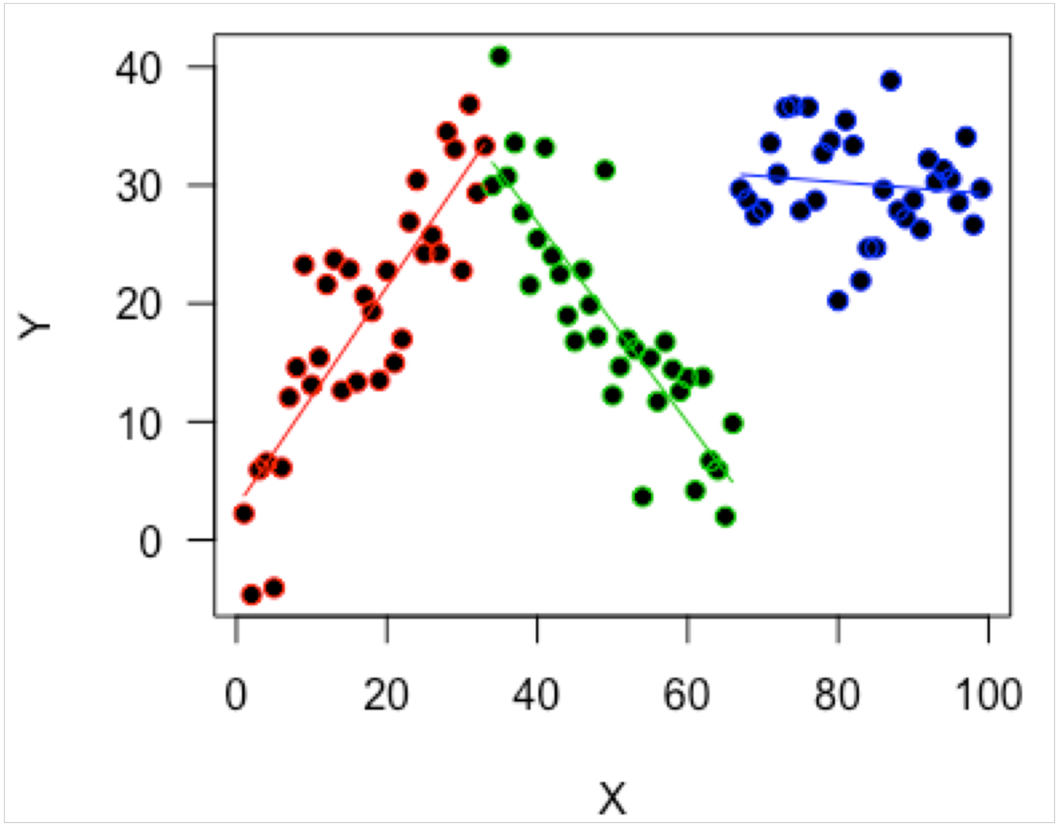
```
model2 <- lm(Y~X*G)
```



```
> coef(model2)
(Intercept)           X          GB          GC        X:GB        X:GC 
  2.7816210   0.9314119  57.9696096  31.4551418  -1.7785780  -0.9812481 
```

Differences in slopes

Interaction!



Overdispersion

Overdisperion

Variance is controlled by the mean (assumption)

Not always true

We could get **overdispersion** (more variation than we expect)

Can check!

Overdisperion

If the variance is controlled by the mean – should also control the **residual deviance**

Can estimate the overdispersion from deviance

Take the ratio of residual deviance and residual degrees of freedom

Can find these in summary() e.g.

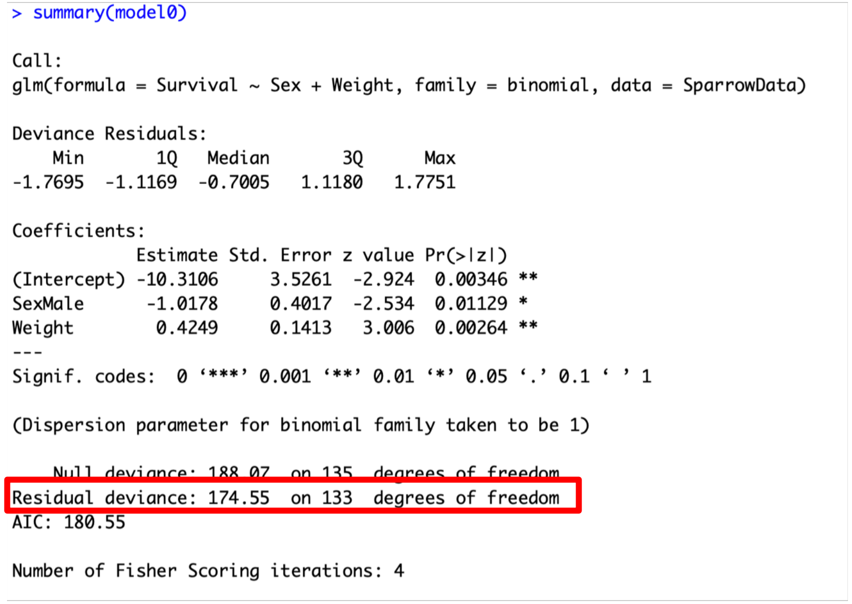
Overdisperion

If the variance is controlled by the mean – should also control the **residual deviance**

Can estimate the overdispersion from deviance

Take the ratio of residual deviance/residual degrees of freedom

Can find these in summary() e.g.



```
> summary(model0)

Call:
glm(formula = Survival ~ Sex + Weight, family = binomial, data = SparrowData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.7695  -1.1169  -0.7005   1.1180   1.7751  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept) -10.3106     3.5261  -2.924  0.00346 **
SexMale      -1.0178     0.4017  -2.534  0.01129 * 
Weight        0.4249     0.1413   3.006  0.00264 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.07  on 135  degrees of freedom
Residual deviance: 174.55  on 133  degrees of freedom
AIC: 180.55

Number of Fisher Scoring iterations: 4
```

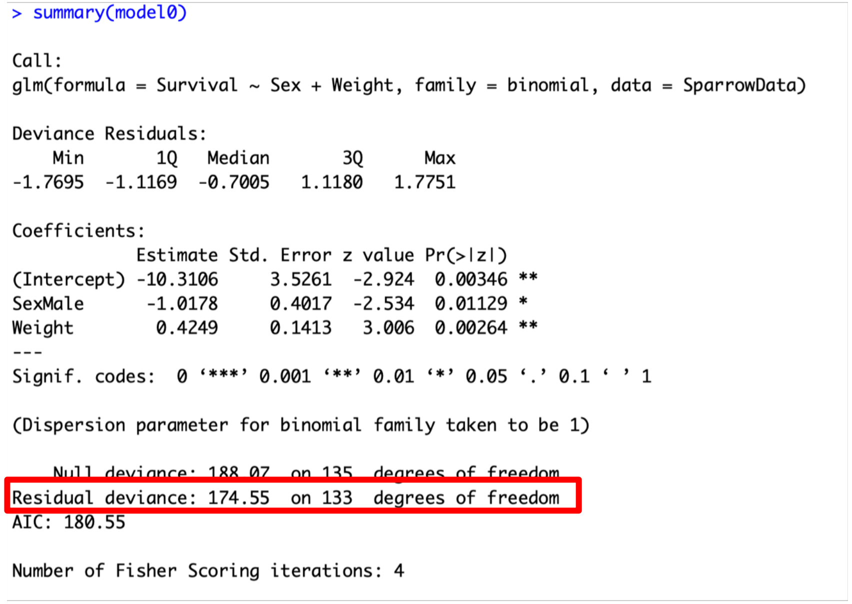
Overdisperion

If the variance is controlled by the mean – should also control the **residual deviance**

Can estimate the overdispersion from deviance

Take the ratio of residual deviance/residual degrees of freedom

Can find these in summary() e.g.



```
> summary(model0)

Call:
glm(formula = Survival ~ Sex + Weight, family = binomial, data = SparrowData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7695  -1.1169  -0.7005   1.1180   1.7751

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.3106     3.5261  -2.924  0.00346 **
SexMale      -1.0178     0.4017  -2.534  0.01129 *
Weight        0.4249     0.1413   3.006  0.00264 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.07  on 135  degrees of freedom
Residual deviance: 174.55  on 133  degrees of freedom
AIC: 180.55

Number of Fisher Scoring iterations: 4
```

Deviance ratio = 174.55/133 = 1.31

With no overdispersion should be 1, quite close here!

Exercise this week

A different example – but more sparrows

Practice interpreting

More exam style questions (but still some coding)

Lecture Summary

Mini-lecture 1 = Other links

Logit, Probit and cloglog for Binomial

Mini-lecture 2 = Categorical and continuous

Influences how we interpret the output of our models

Mini-lecture 3 = Overdispersion

Can check by looking a deviance ratio