Data
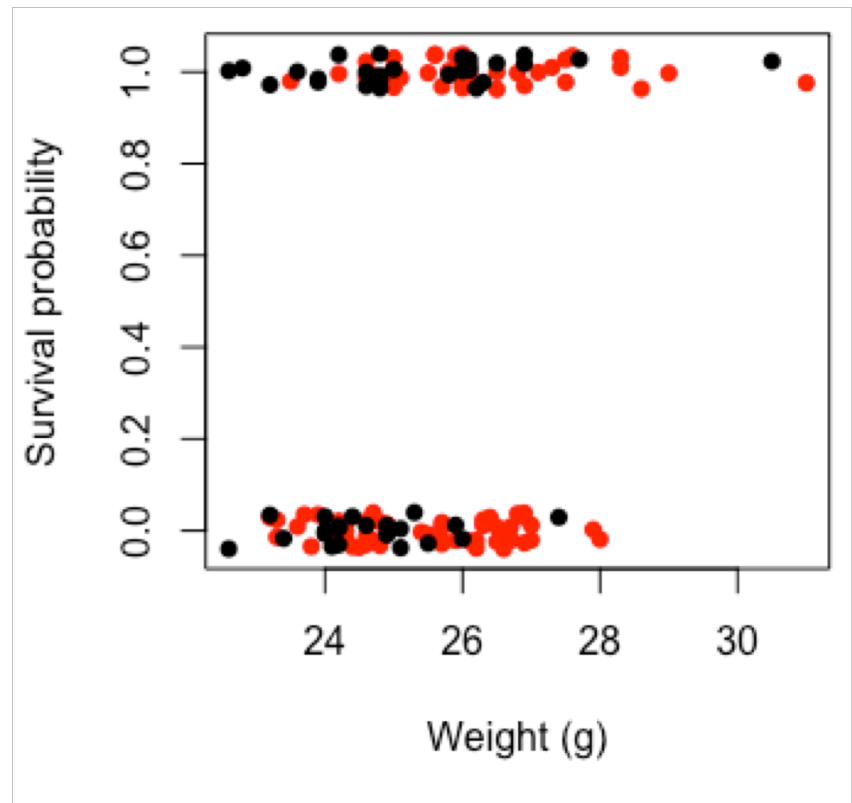

Sex = Male and Female
Age = Adult and Juvenile
Survival = 0 and 1
Weight = kg

```
> 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 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 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
```
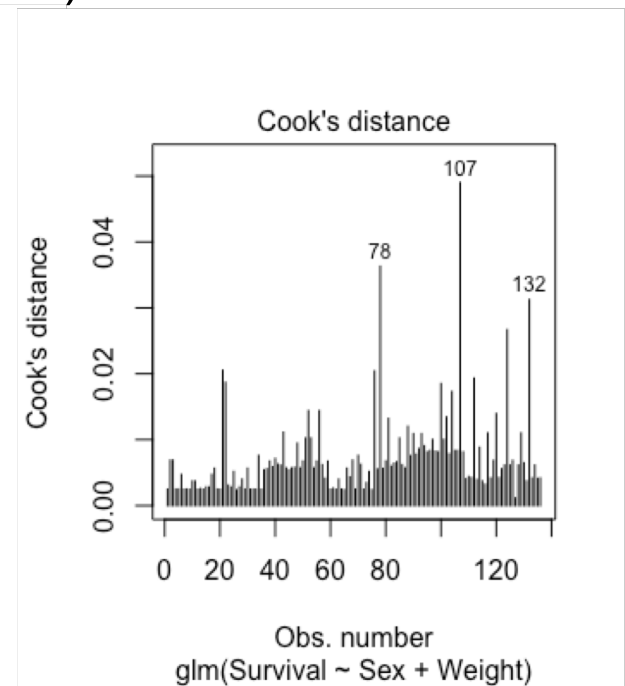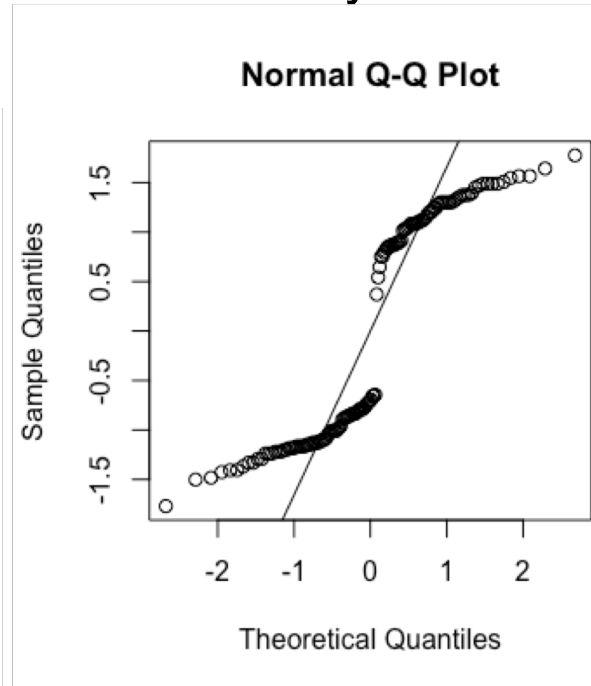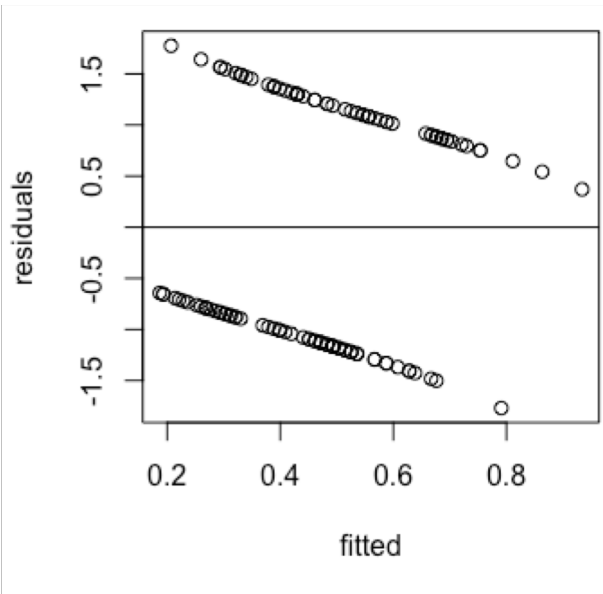
- Above is an output from R.

- What analysis has been conducted?

- What does the analysis aim to find out?

- What can you conclude from this output?

Exercise 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: 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?
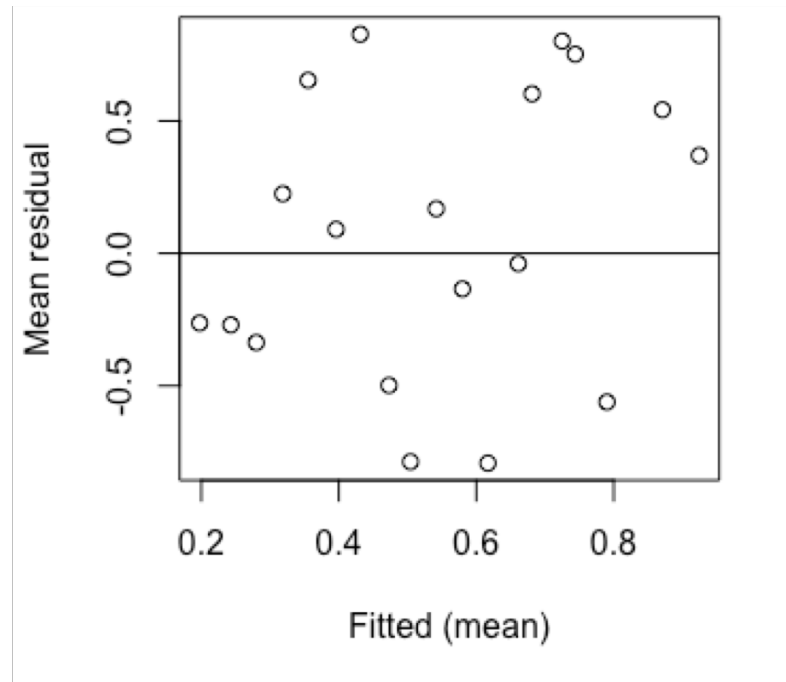
Exercise 3b: Interpreting

- 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

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

HINTS

*Hints

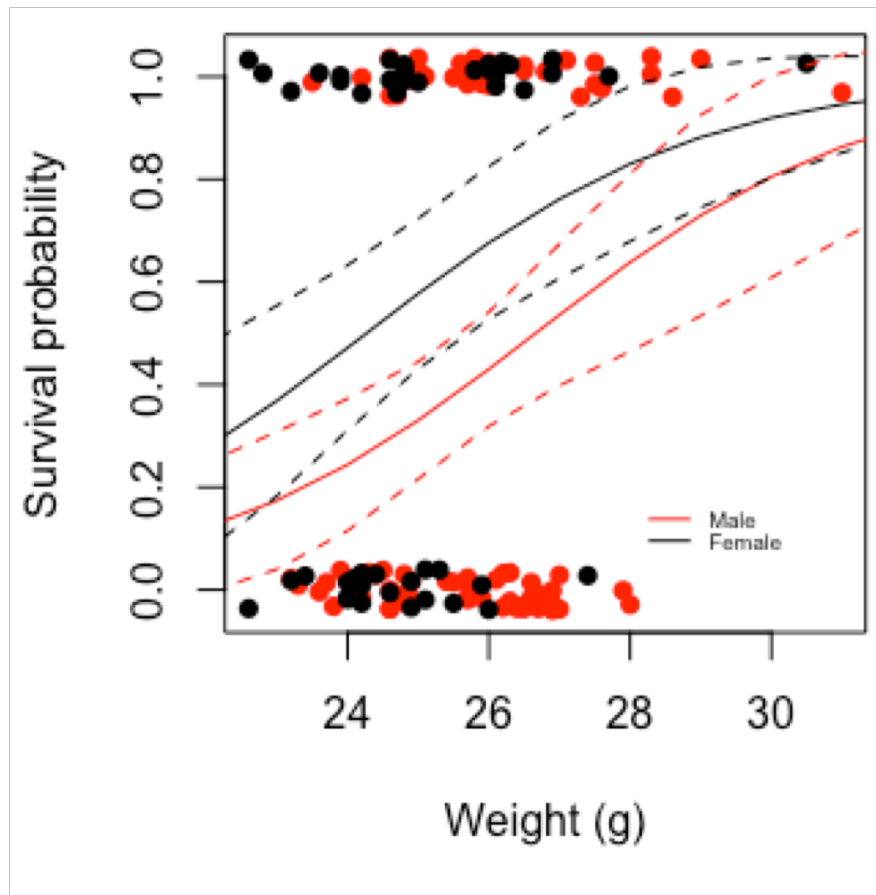Think about what kind of data went into the model (particularly the explanatory variables)

What was the biological question?

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

Remember the link function

- 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)

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
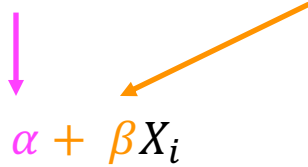
Exam style

Recap of last yesterday

Logit link

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

The inverse is

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

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

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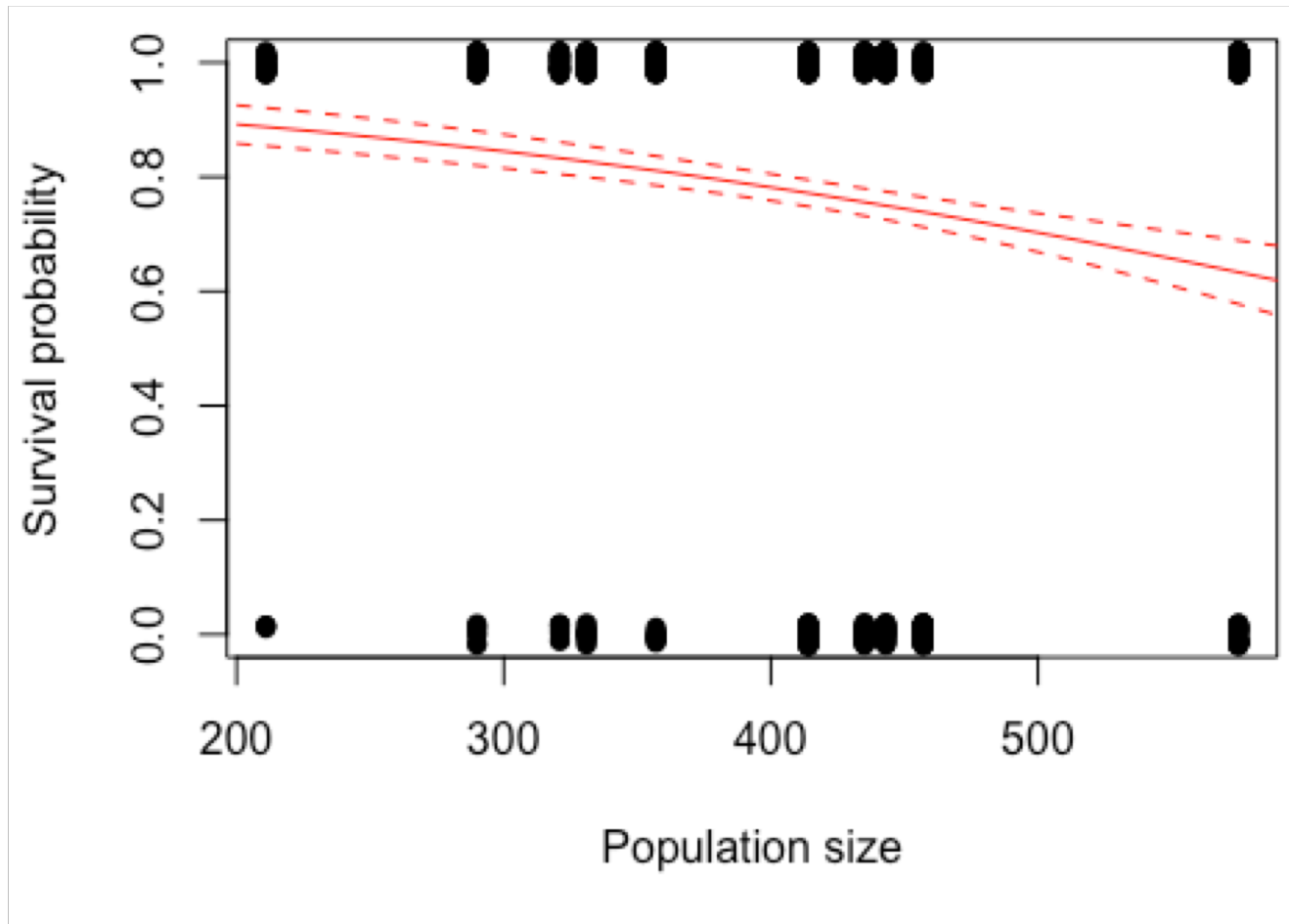
$$\frac{1}{1 + e^{-\mu}} = p$$

E.g.

For X (PopSize) = 300

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

For X (PopSize) = 400

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

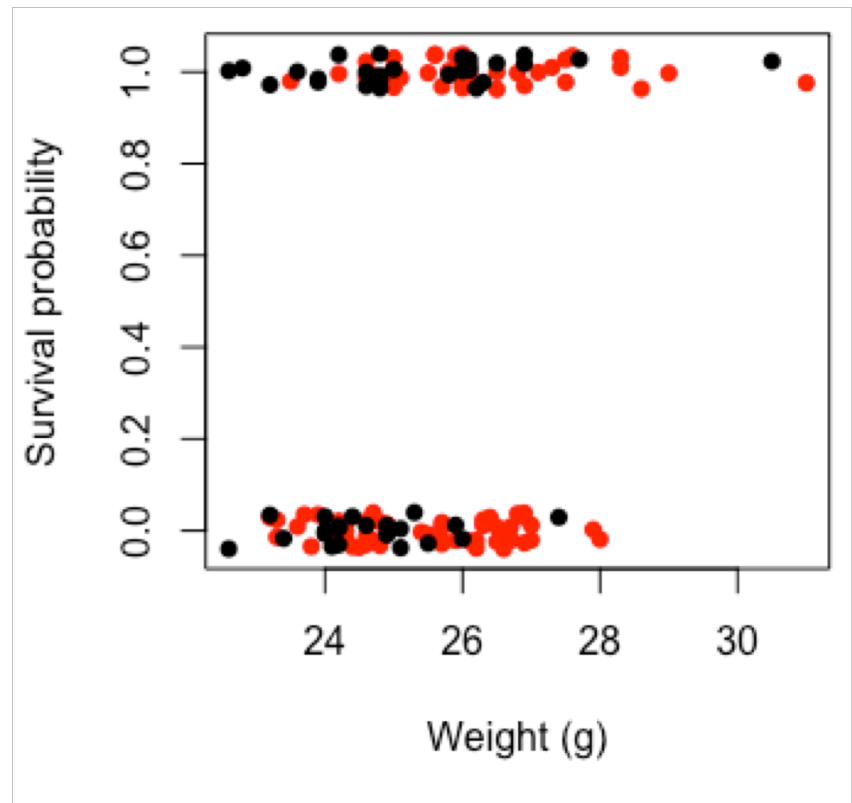**or plot it!**

Introduction to the data

Data

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

```
> 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
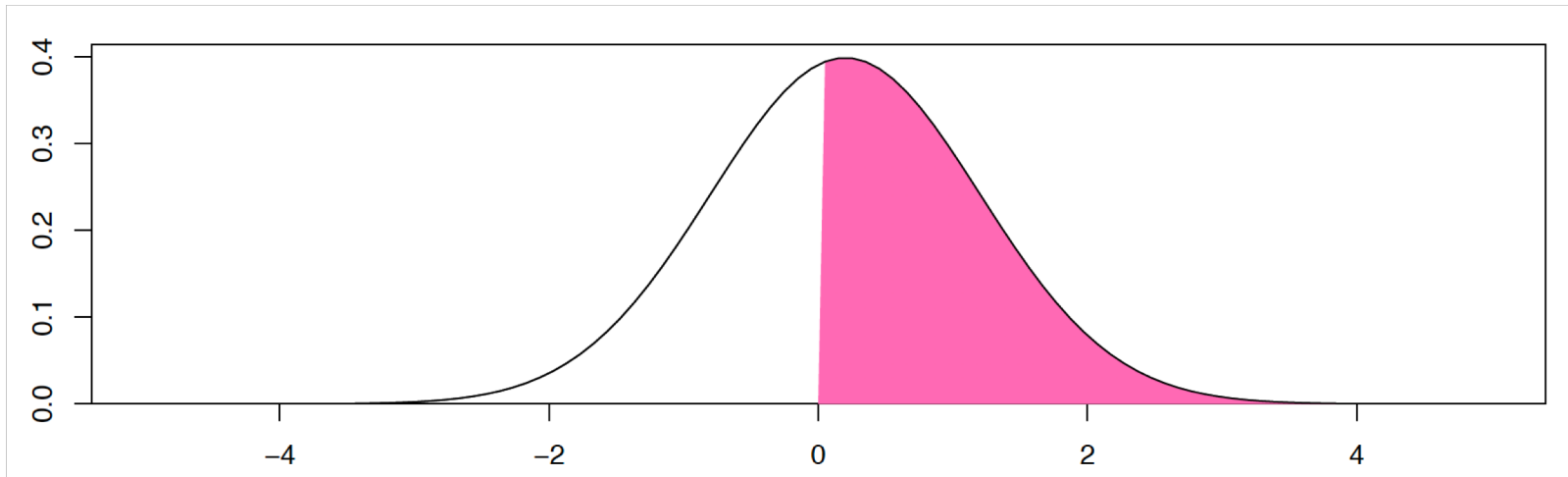
This is the default (canonical) link in R

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

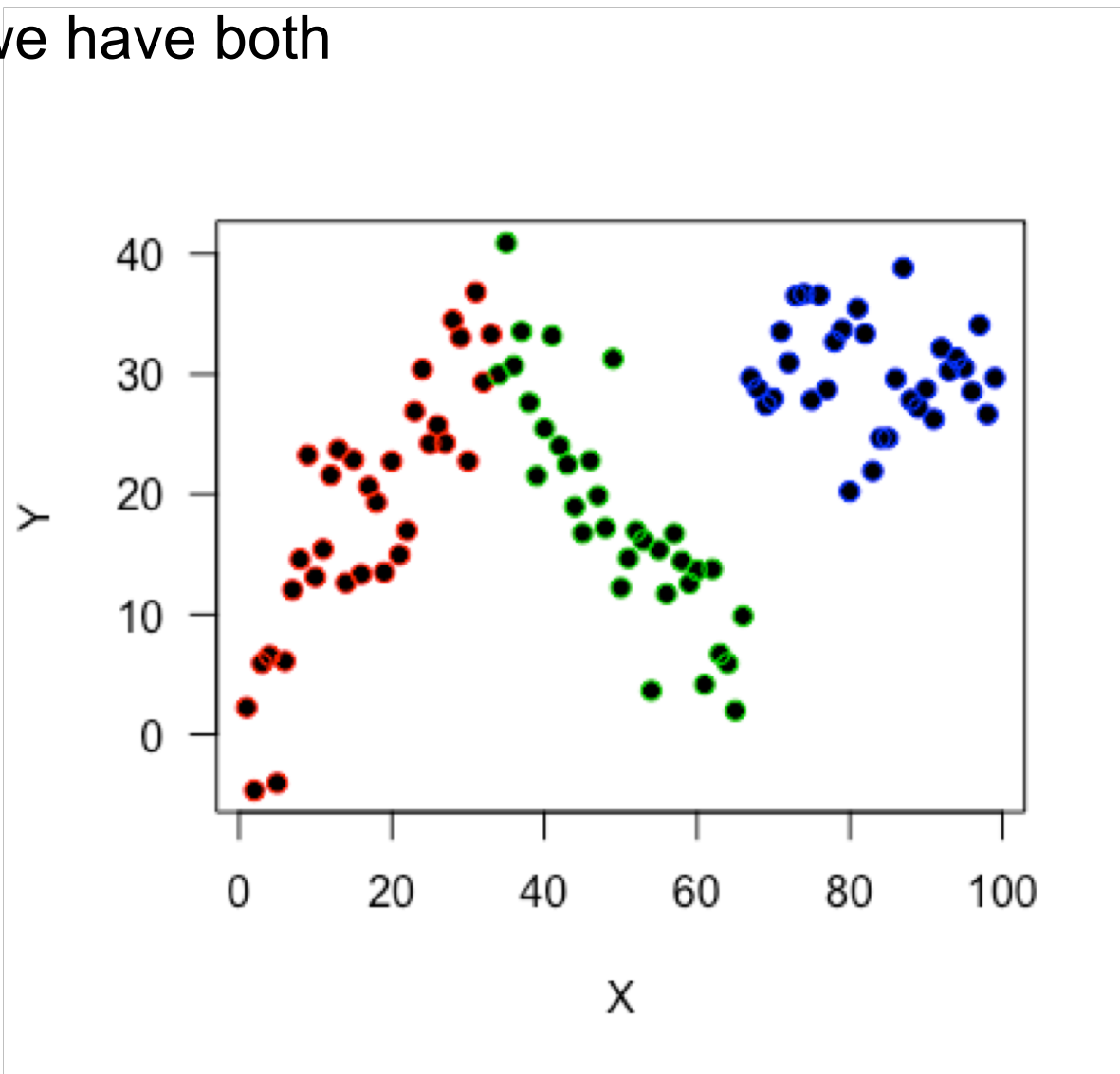**cloglog = if you have Poisson-like counts**

Categorical and continuous

Definitions

Categorical = in groups

Continuous = every value can exist

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
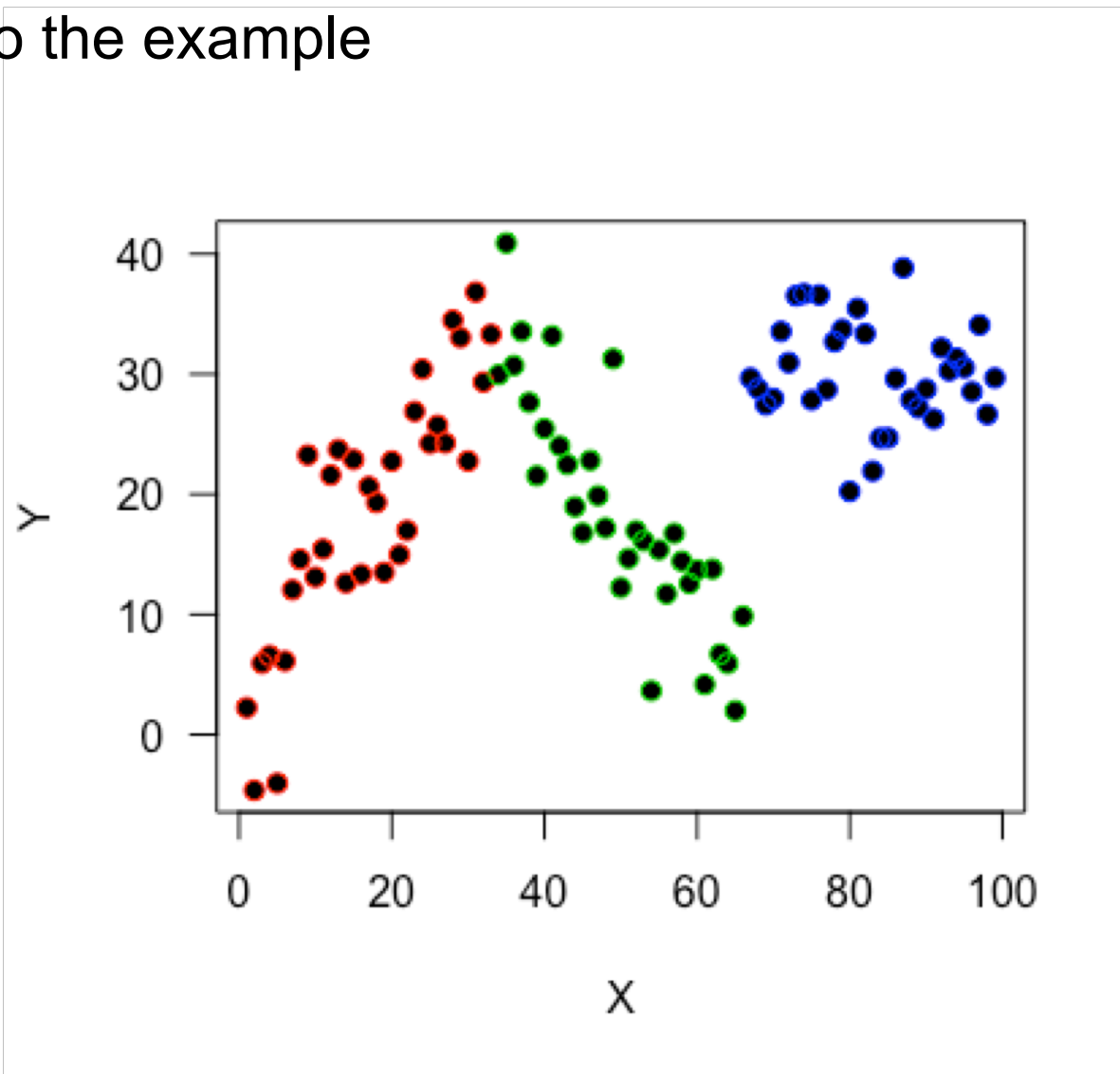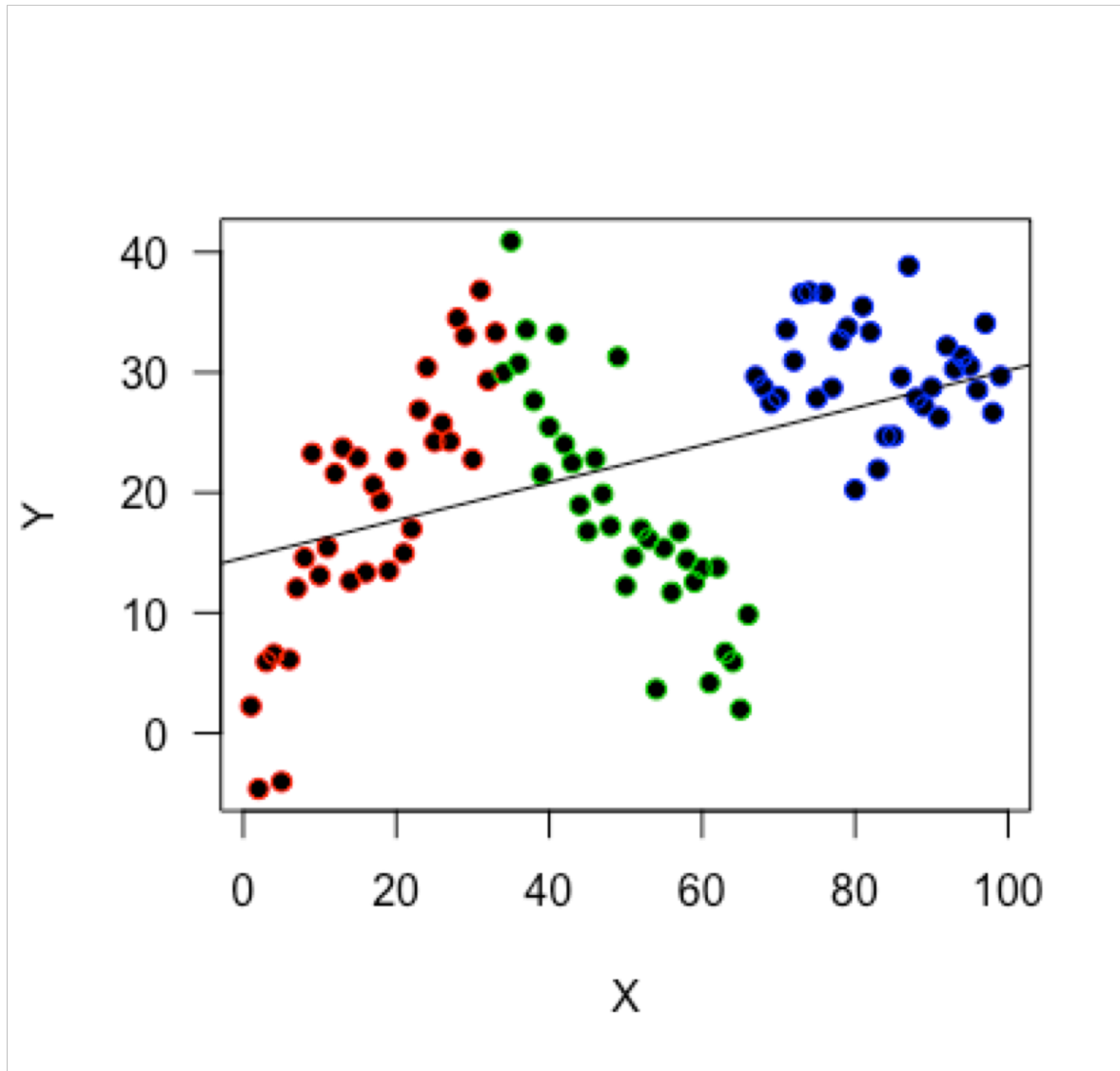
Back to the example

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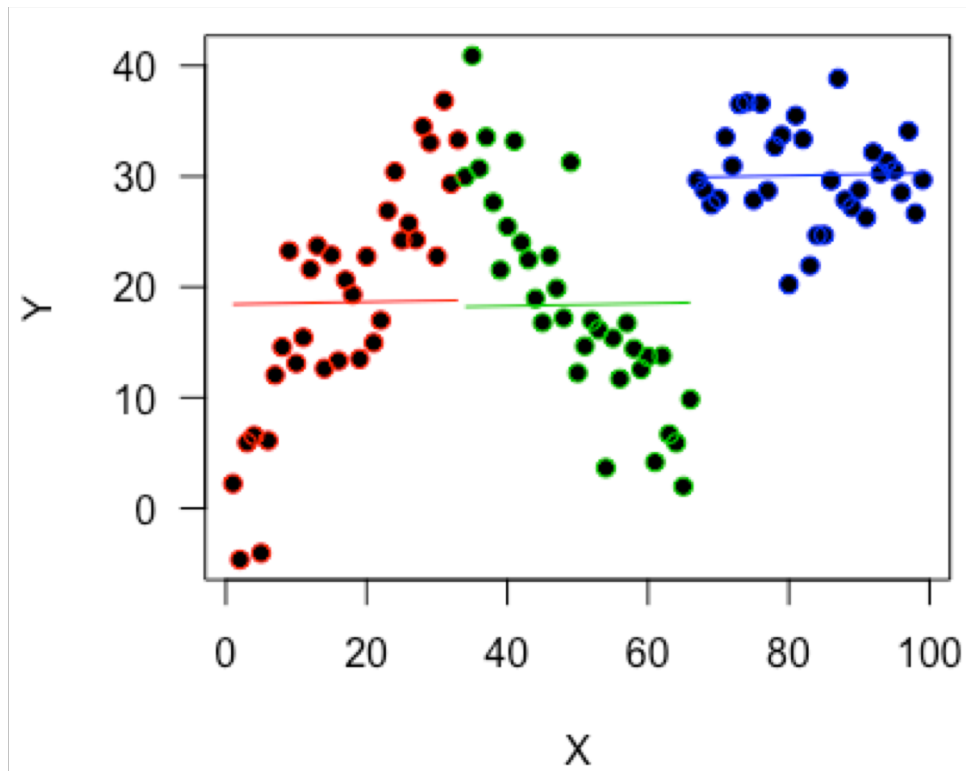
```
> coef(model2)
(Intercept)             X            GB             GC          X:GB          X:GC
  2.7816210     0.9314119   57.9696096   31.4551418   -1.7785780   -0.9812481
```

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

```
> coef(model1)
(Intercept)              X              GB              GC
18.42063558   0.01146992  -0.60120409  10.72772509
```
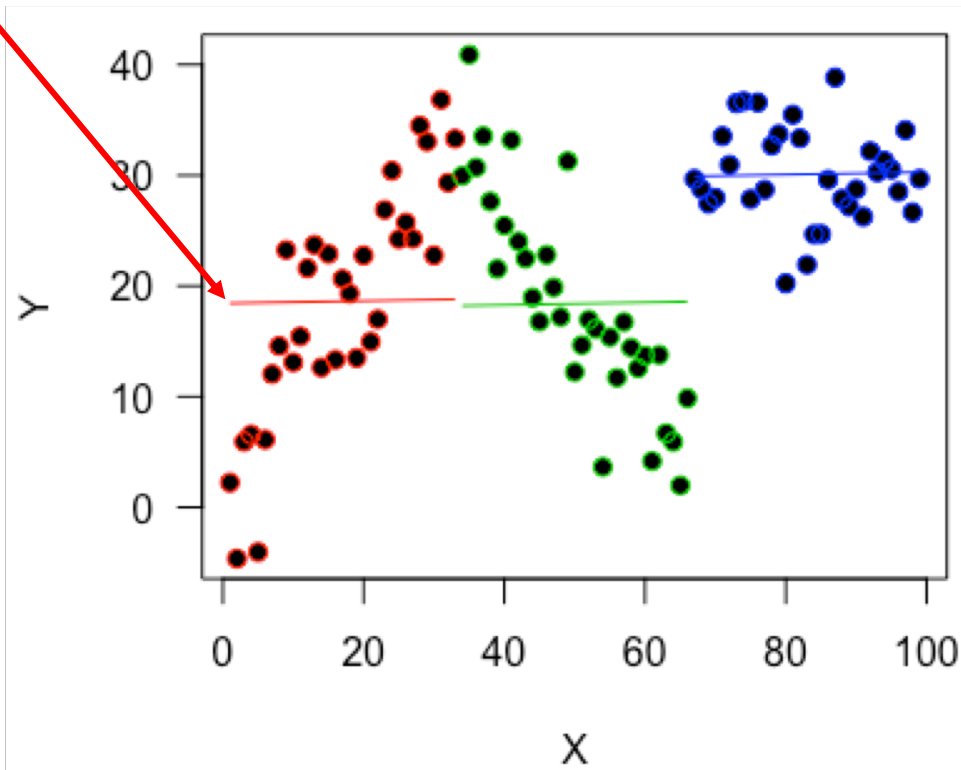
```
model1 <- lm(Y~X+G)
```

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

Intercept
of line of
Group A

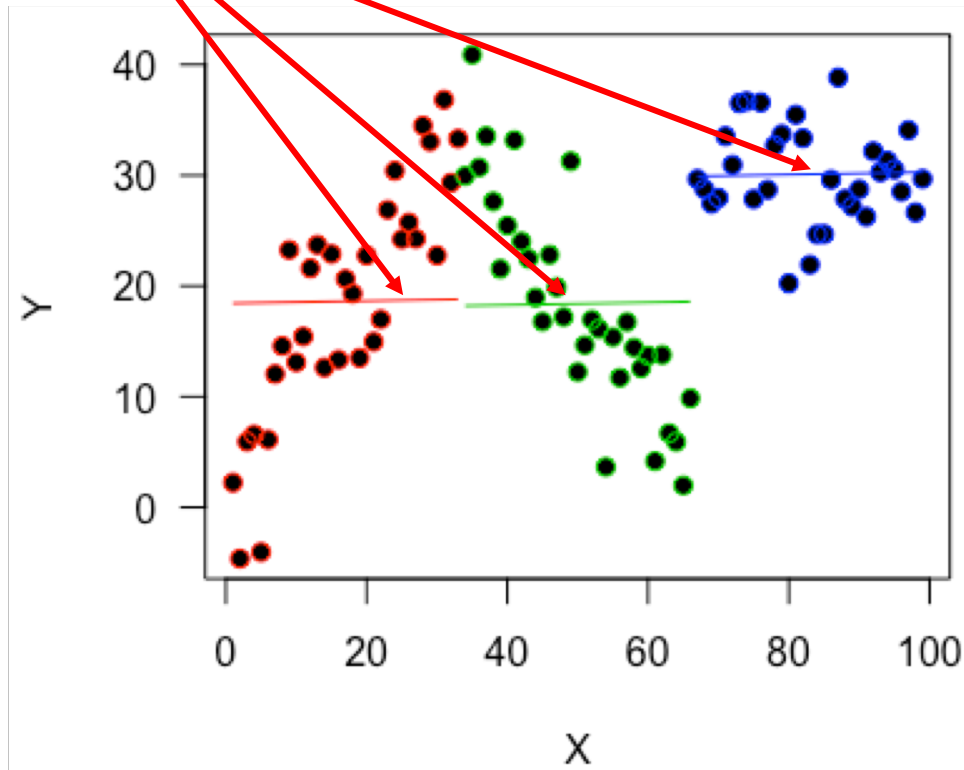# Interpreting

```
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)

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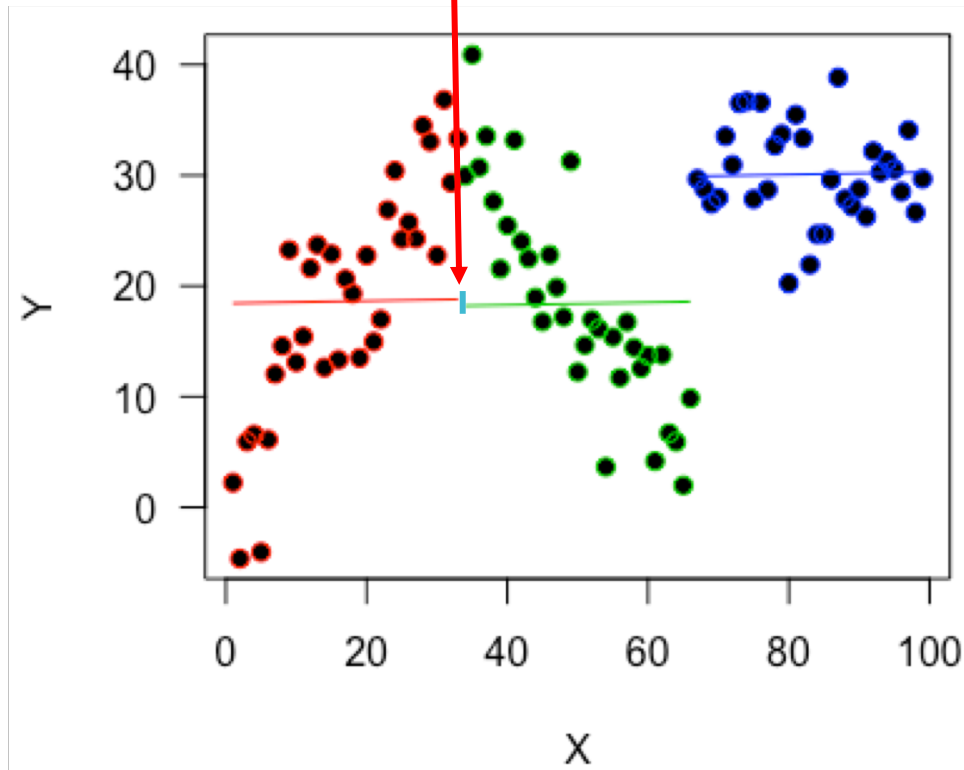
```
model1 <- lm(Y~X+G)
```

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

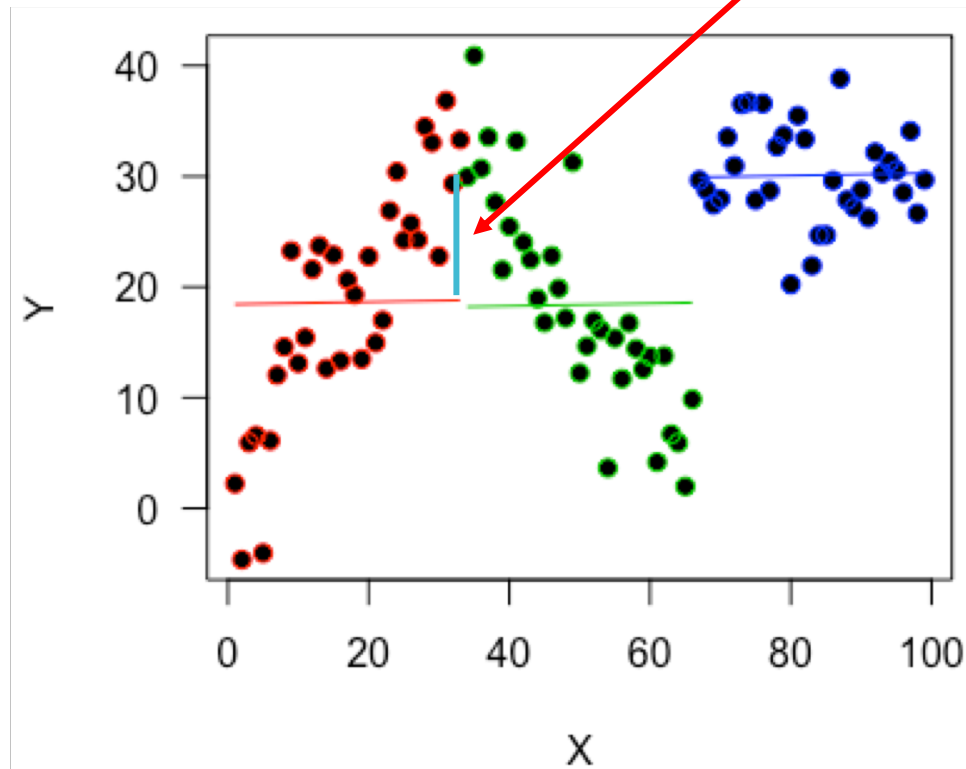Difference in intercept from Group A to Group C

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
```
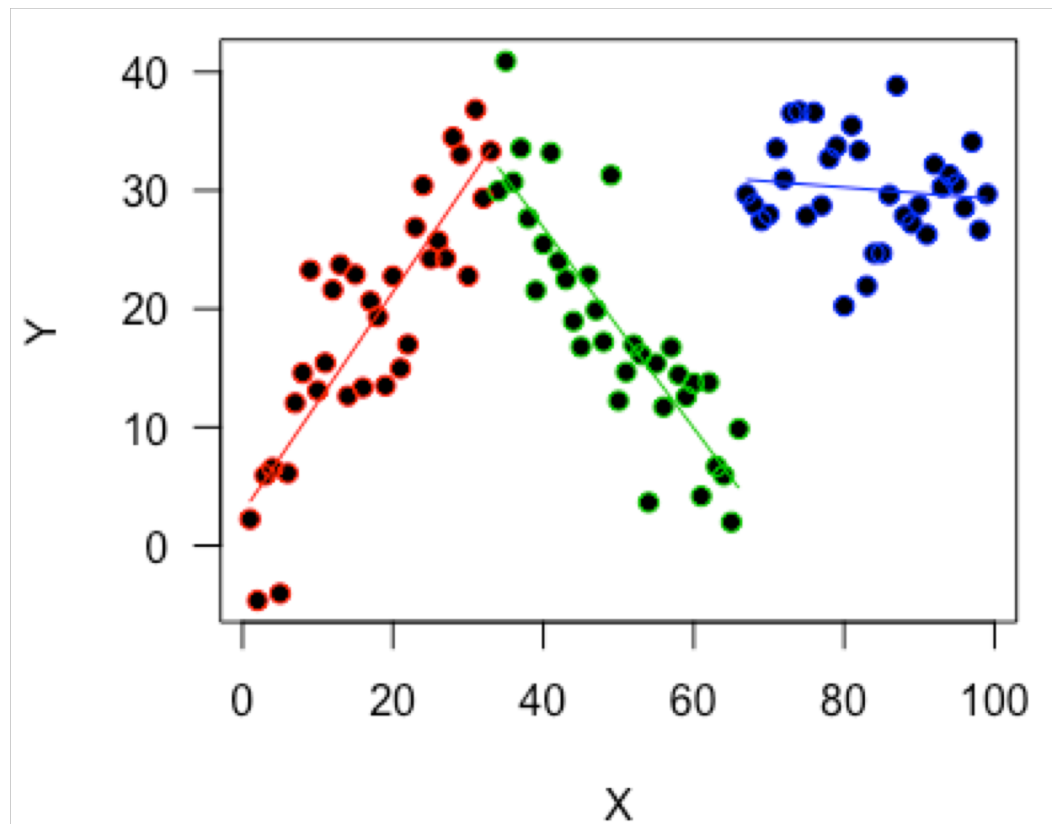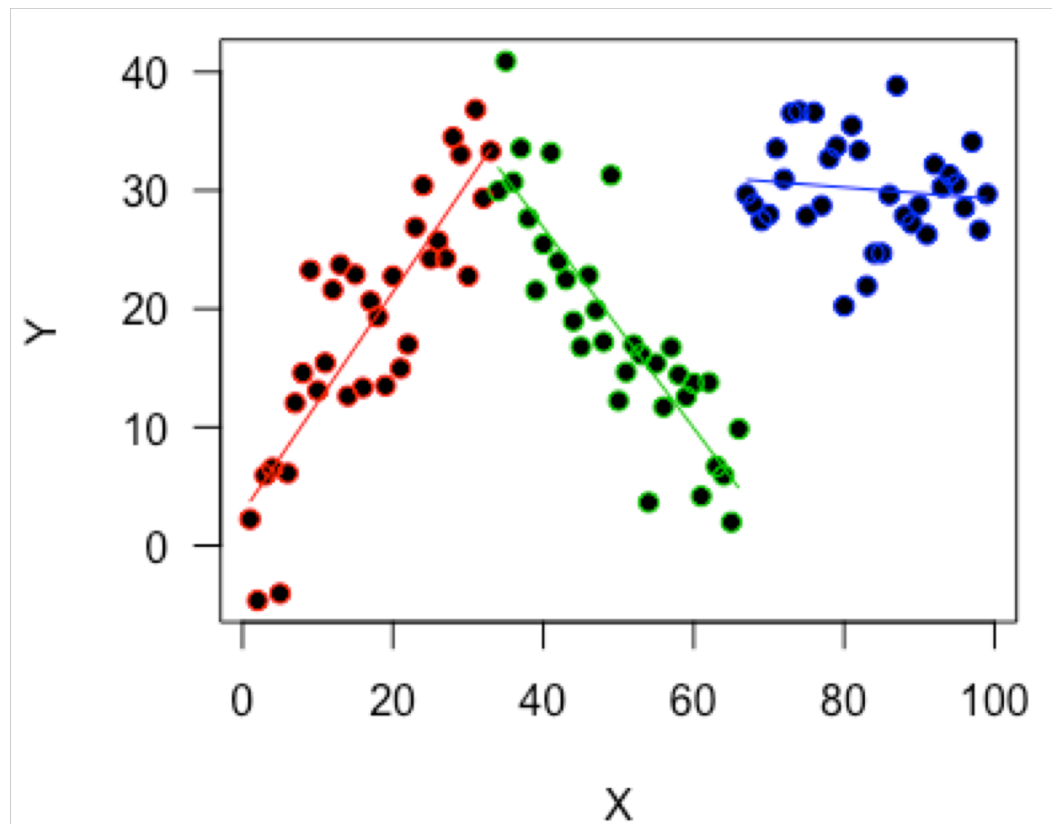
Interpreting

```
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
```
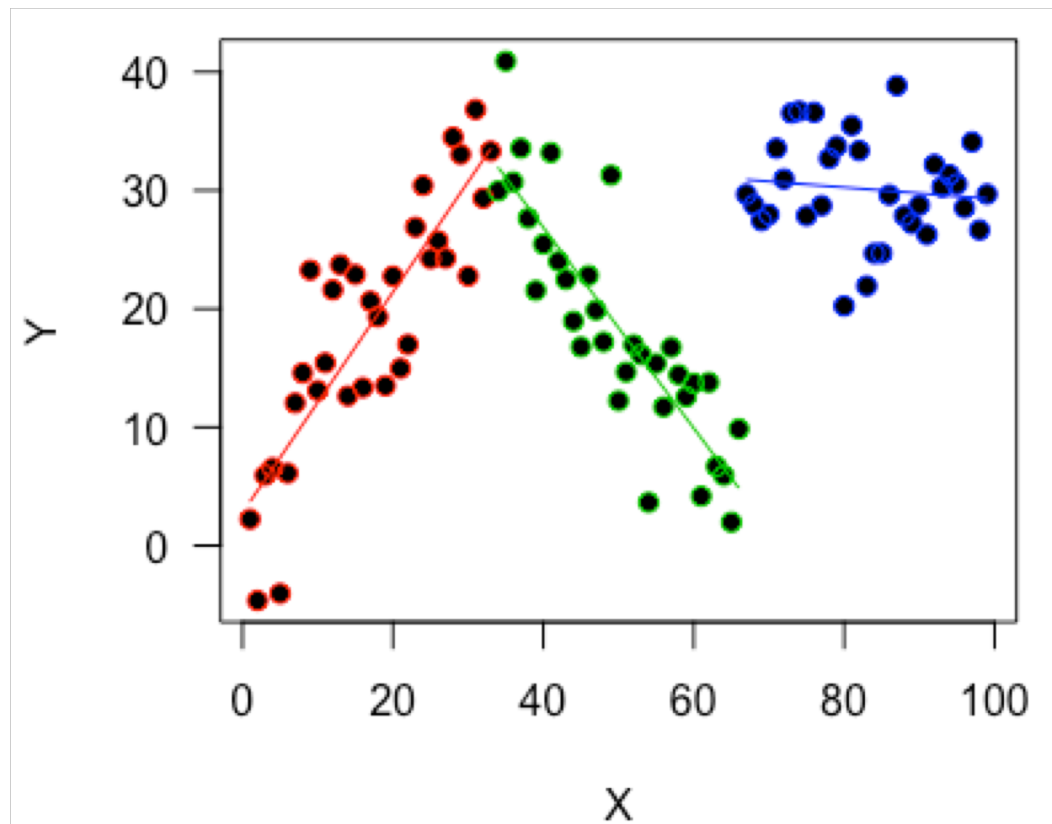
Differences in slopes

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