Revision lectures 2020v

Outline

Summary of the course

Interpreting R outputs

Model selection

Model checking

Overdispersion

Course summary

Steps of modelling (week 2 recap)

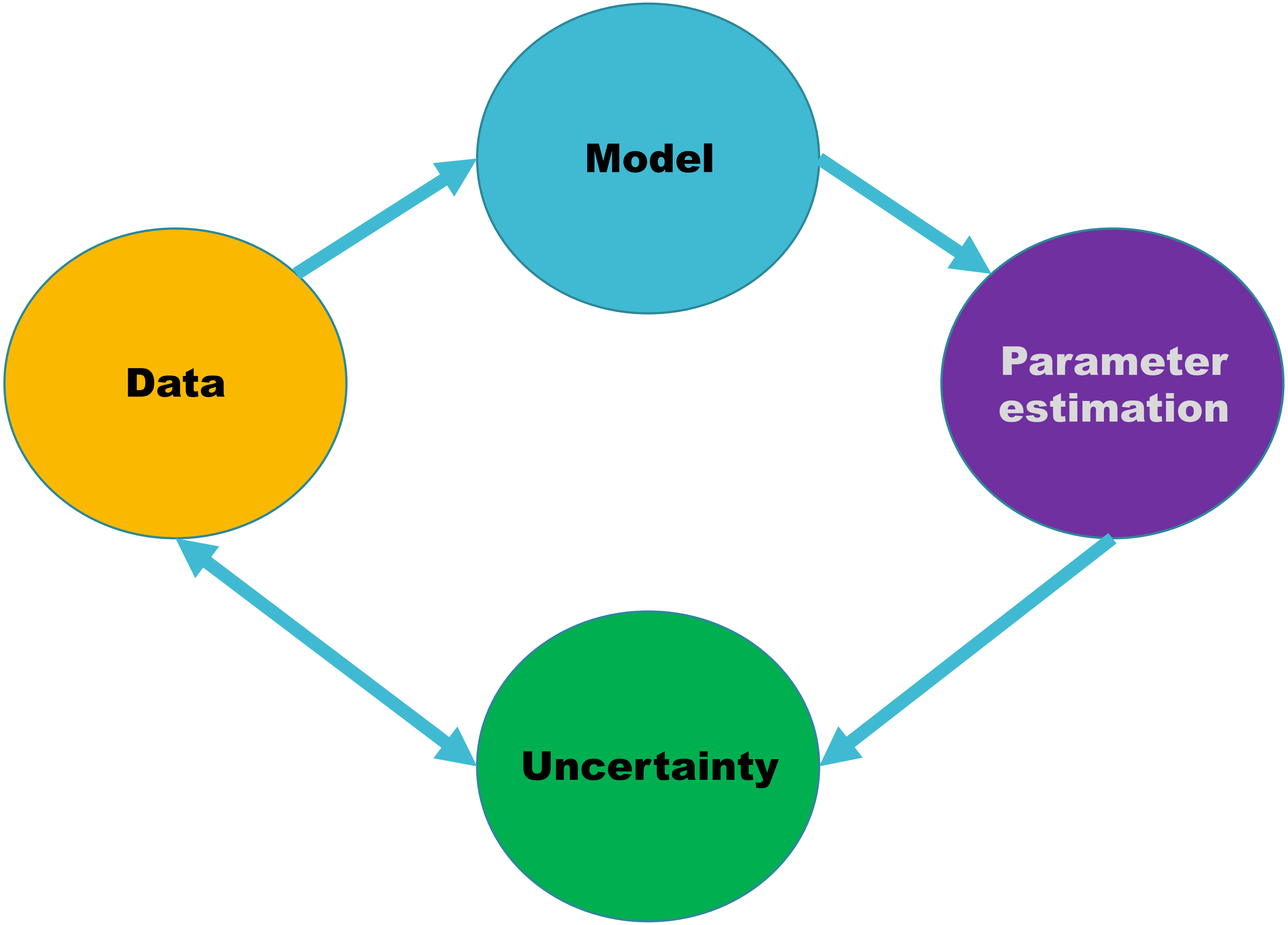
1. Choose a model for your data

2. Get estimates of the parameters

3. Quantify uncertainty in the estimates

4. Interpret the results

Schematic



Types of models

Distributions (Binomial, Poisson, and Normal)	Weeks 1-4
Linear models	Weeks 5-9 + 11
Generalised linear models	Weeks 10 + 12-13
- Binomial	Week 12
- Poisson	Week 13

GOOD LUCK!!!

Model selection

Two aims of model selection

Confirmatory: to test a specific hypothesis

Exploratory: to find which variables of many influence the response

Confirmatory model selection

Why do we need it?

Confirmatory model selection

Why do we need it?

To test a specific hypothesis

e.g. **Does variable X influence variable Y?**

Confirmatory model selection

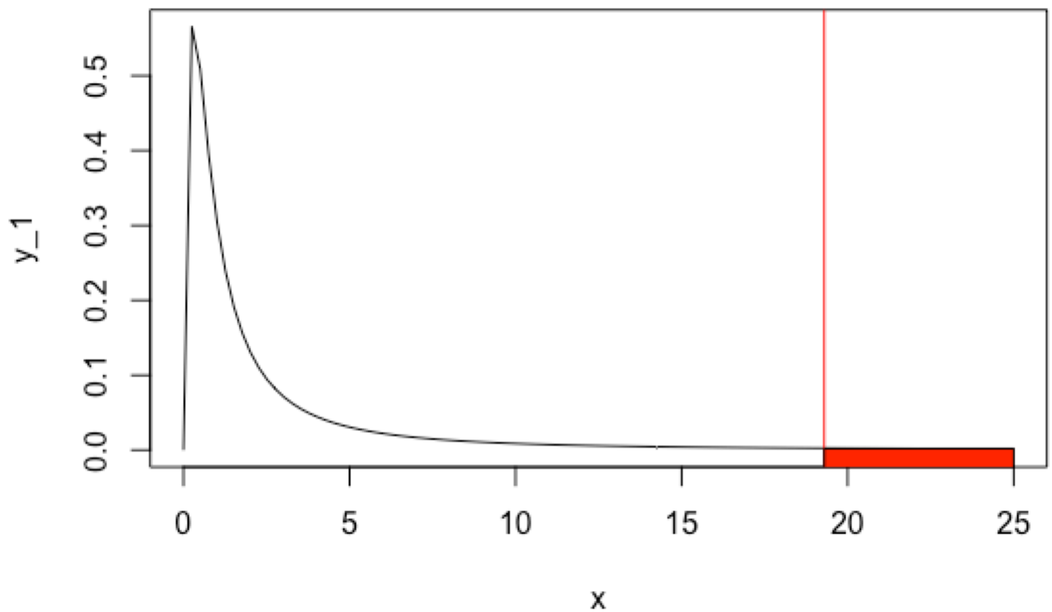
Even with nothing going on….

sometimes you would expect to collect data that shows an effect.

Need to distinguish whether the effect you found is likely to occur if a null assumption is true.

Confirmatory model selection

Use a hypothesis test of H0 and H1 (null and alternative) to decide if the estimated statistic fits with the null assumption.



Confirmatory model selection

To use them:

Construct lm() or glm() for H0 model and H1 model.

Then use ANOVA or Analysis of deviance to test the hypothesis.

Want to see the probability of getting the F-ratio or deviance value estimated if the null were true.

Confirmatory model selection

```
## Analysis of Variance Table
##
## Model 1: PopularityScore ~ 1
## Model 2: PopularityScore ~ Weight
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     99 510.43
## 2     98 484.88  1    25.545 5.1628 0.02526 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear model

```
anova(mod, mod1, test="LRT")
```

LRT = likelihood ratio test

```
Analysis of Deviance Table

Model 1: SimR ~ 1
Model 2: SimR ~ X
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        99     95.487
2        98     94.961  1  0.52572   0.4684
```

Generalised linear model

Exploratory model selection

Why do we need it?

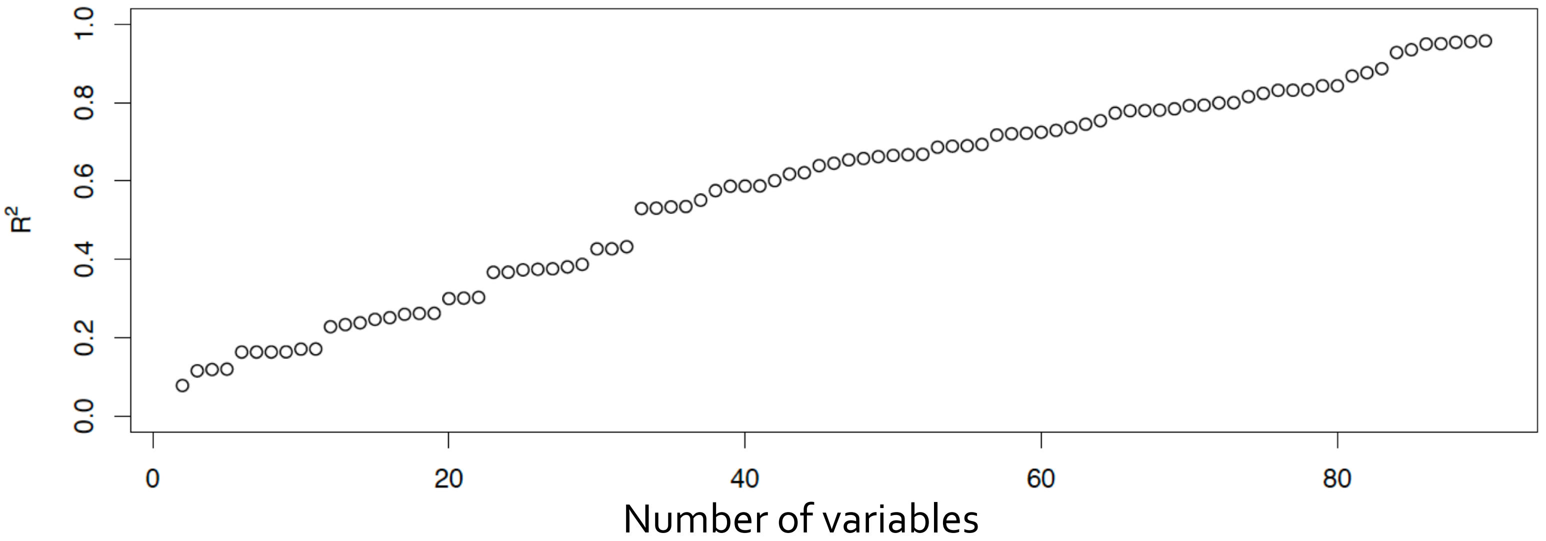
Exploratory model selection

Why do we need it?

To find a ‘best’ model from several candidates

e.g. **Which of these 20 variables I collected data on explain my response variable?**

Exploratory model selection



Every time we add an explanatory variable to a model, the R^2 increases

R^2 = our measure of how much of the variation in the data is explained by our model

Even if variables are random

Exploratory model selection

Every time we add an explanatory variable to a model, the degrees of freedom decrease.

This is because the number of parameters estimated increases.

Exploratory model selection

We need a way to work out what a good or ‘best’ model is.

We need to balance **fit** with **simplicity**.

We have the AIC and BIC to do this.

Exploratory model selection

AIC: tries to find the model that best predicts the data.

BIC: tries to find the model most likely to be true.

You can often choose which works best for you. Just remember to justify the choice!

You cannot do both.

Exploratory model selection

Both AIC and BIC add penalties for model complexity.

AIC = (-2*Likelihood) + (2*Number of parameters)

BIC = (-2*Likelihood) + log(n)*Number of parameters

BIC has the higher penalty for complexity.

Both use Likelihood as a measure of fit.

For both **lower = better.**

Exploratory model selection

To use them:

Construct lm() or glm() for all combinations of the variables you want to test.

Then calculate AIC or BIC for each.

Pick the lowest. (within 2 of lowest is considered pretty similar)

Can be quicker in bestglm()

Interpreting R output:

interactions

Interpreting R outputs

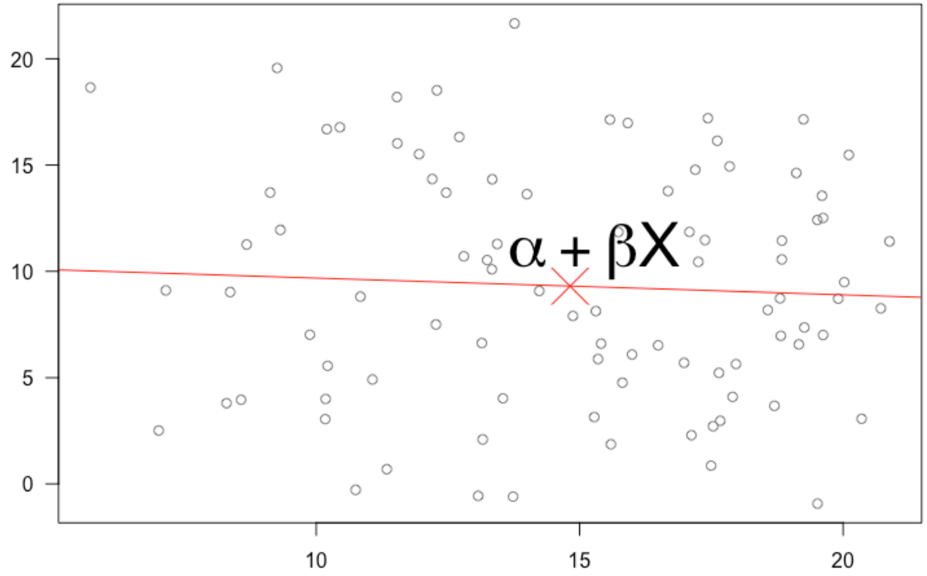
Linear models (glm) or (lm) are all based on this equation

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Interpreting R outputs

Linear models (glm) or (lm) are all based on this equation

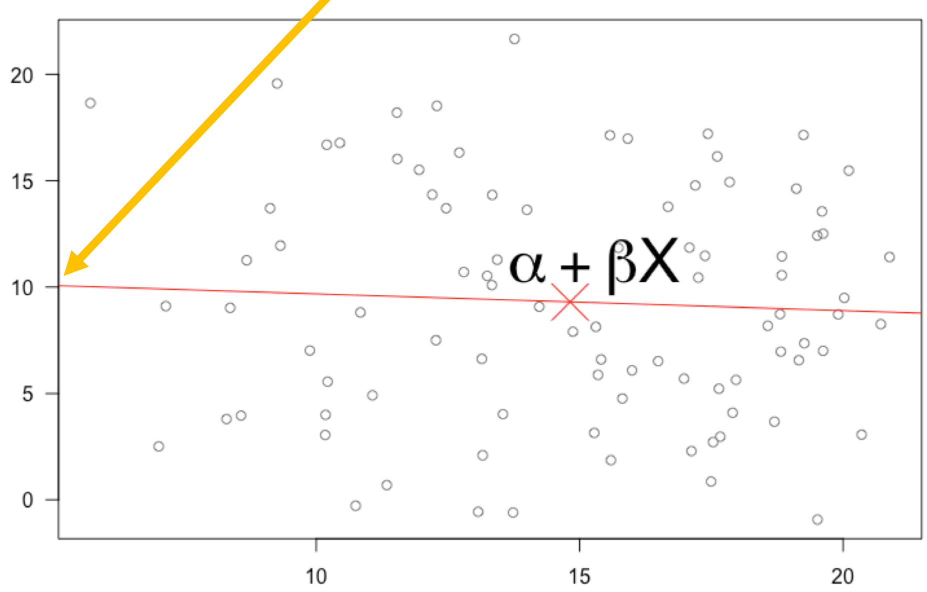
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Interpreting R outputs

Linear models (glm) or (lm) are all based on this equation

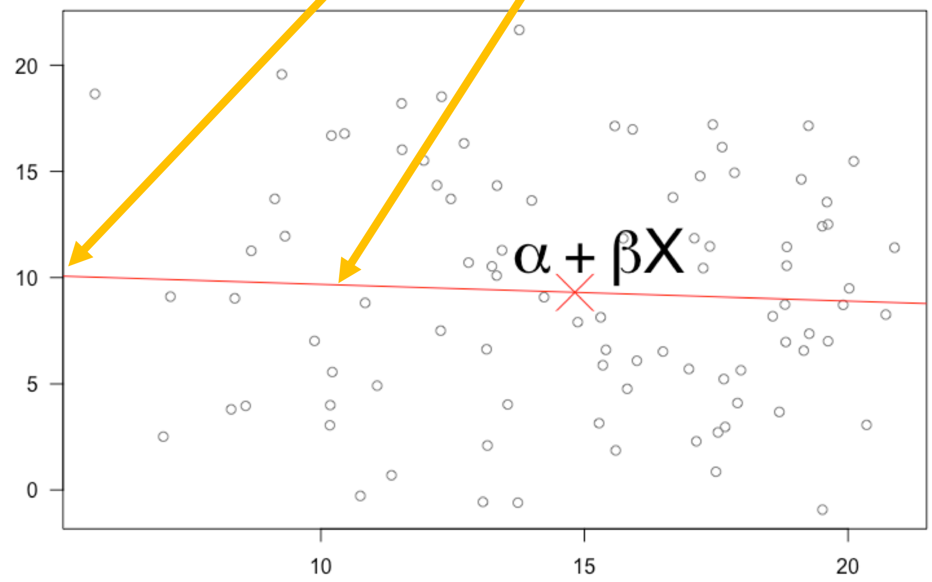
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Interpreting R outputs

Linear models (glm) or (lm) are all based on this equation

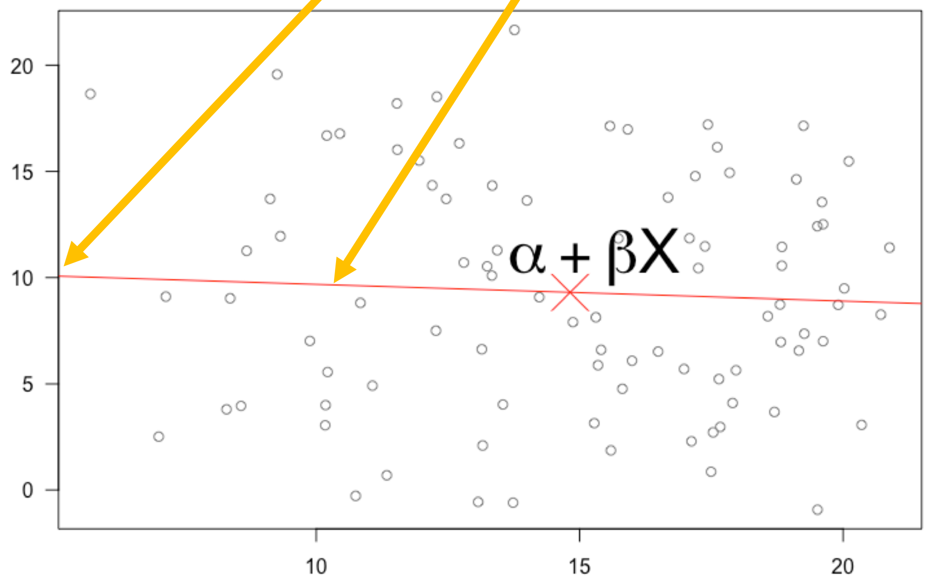
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Interpreting R outputs

Linear models (glm) or (lm) are all based on this equation

$$Y_i = \alpha + \beta X_i + \varepsilon_i \text{ error}$$



Interpreting R outputs

The outputs of the models also map onto the linear equation.

Several different functions to look at the output: coef(), confint(), summary()

Interpreting R outputs

coef

```
> coef(model)
(Intercept)           X
   2.358216    0.010118
```

Interpreting R outputs

coef

```
> coef(model)
(Intercept)           X
   2.358216    0.010118
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Interpreting R outputs

coef

```
> coef(model)
(Intercept)           X
   2.358216    0.010118
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Interpreting R outputs

confint

```
> confint(model)
                  2.5 %     97.5 %
(Intercept)  0.74006514 3.97636758
X           -0.01770064 0.03793665
```

Interpreting R outputs

confint

```
> confint(model)
                  2.5 %     97.5 %
(Intercept)  0.74006514 3.97636758
X           -0.01770064 0.03793665
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Interpreting R outputs

confint

```
> confint(model)
                   2.5 %     97.5 %
(Intercept)  0.74006514 3.97636758
X           -0.01770064 0.03793665
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Interpreting R outputs

summary

```
> summary(model)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5535 -2.9695  0.3335  3.1508  9.0043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.35822    0.81541   2.892  0.00471 **
X            0.01012    0.01402   0.722  0.47215   
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.047 on 98 degrees of freedom
Multiple R-squared:  0.005288,	Adjusted R-squared:  -0.004862 
F-statistic: 0.521 on 1 and 98 DF,  p-value: 0.4722
```

Interpreting R outputs

summary

```
> summary(model)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5535 -2.9695  0.3335  3.1508  9.0043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.35822    0.81541   2.892  0.00471 **
X            0.01012    0.01402   0.722  0.47215   
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.047 on 98 degrees of freedom
Multiple R-squared:  0.005288,	Adjusted R-squared:  -0.004862 
F-statistic: 0.521 on 1 and 98 DF,  p-value: 0.4722
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Interpreting R outputs

summary

```
> summary(model)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5535 -2.9695  0.3335  3.1508  9.0043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.35822    0.81541   2.892  0.00471 **
X            0.01012    0.01402   0.722  0.47215   
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.047 on 98 degrees of freedom
Multiple R-squared:  0.005288,	Adjusted R-squared:  -0.004862 
F-statistic: 0.521 on 1 and 98 DF,  p-value: 0.4722
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Important considerations:

What α and β represent can be slightly different depending on your explanatory variables.

Interpreting R outputs

Important considerations:

What α and β represent can be slightly different depending on your explanatory variables.

You could have several βs

You could have multiple values relating to α **but still only one intercept!**

You could have differences as well as absolute values

Interpreting R outputs

Important considerations:

It all depends on the explanatory variables

What type of data are they? (categorical or continuous)

How many are there?

Interpreting R outputs

Categorical vs continuous

Interpreting R outputs

Categorical vs continuous

How to identify

Interpreting R outputs

Categorical vs continuous

How to identify

explanatory variables

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

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

Interpreting R outputs

Categorical vs continuous

How to identify

explanatory
variables

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

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

Continuous = single value with same name as variable

Interpreting R outputs

Categorical vs continuous

How to identify

explanatory variables

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

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

Categorical = can be multiple values. Name is variable name + one of the levels/categories/groups of the variable.

Interpreting R outputs

Categorical vs continuous

How it changes interpretation

Interpreting R outputs

Categorical vs continuous

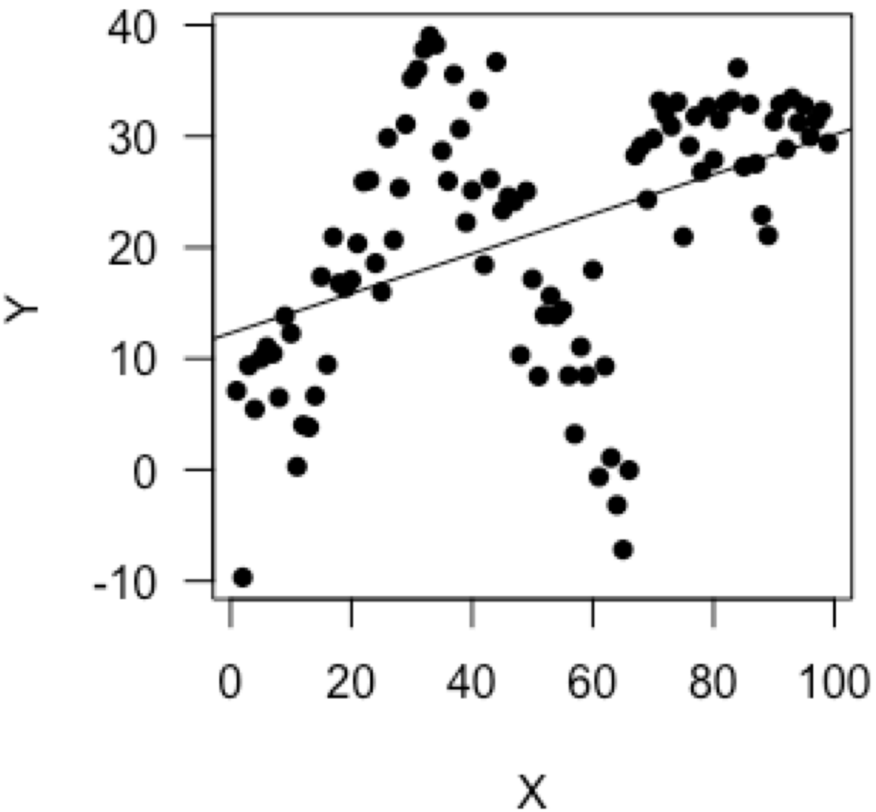
ONLY continuous explanatory variables

Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

```
> coef(lm(Y~X))
(Intercept)           X
 12.2918037   0.1783776
```



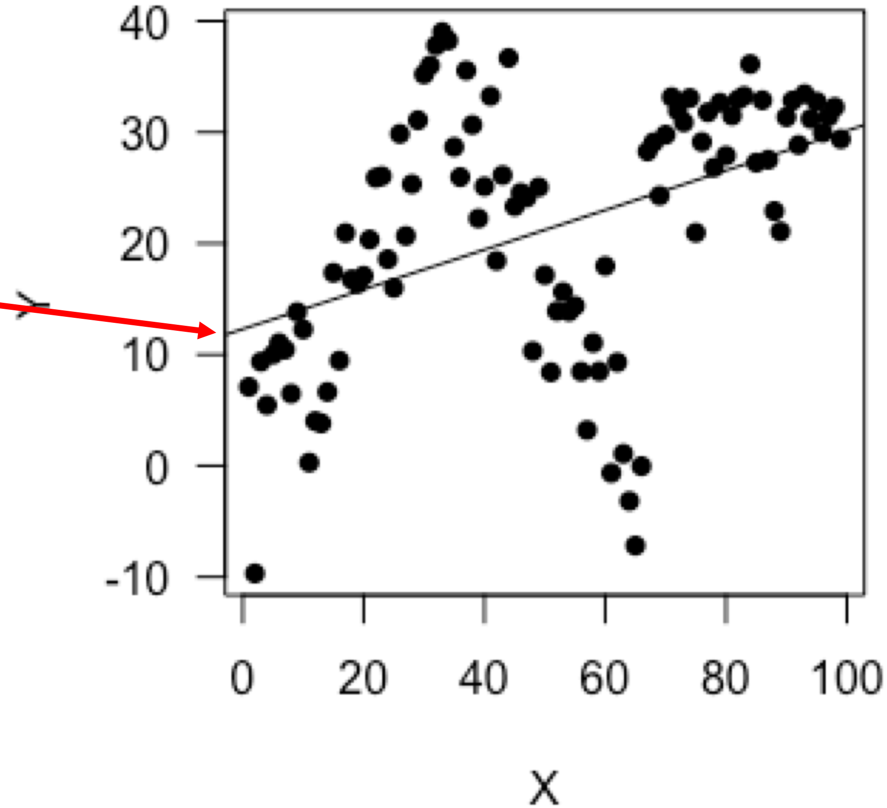
Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

```
> coef(lm(Y~X))
(Intercept)           X
 12.2918037   0.1783776
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



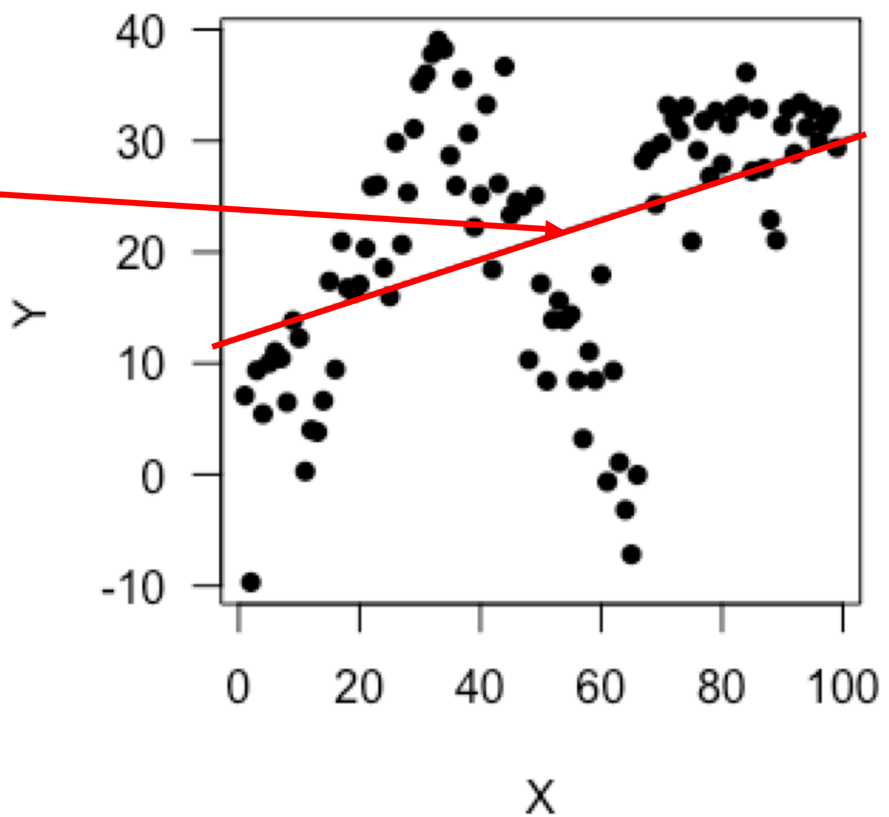
Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

```
> coef(lm(Y~X))
(Intercept)           X
 12.2918037   0.1783776
```

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



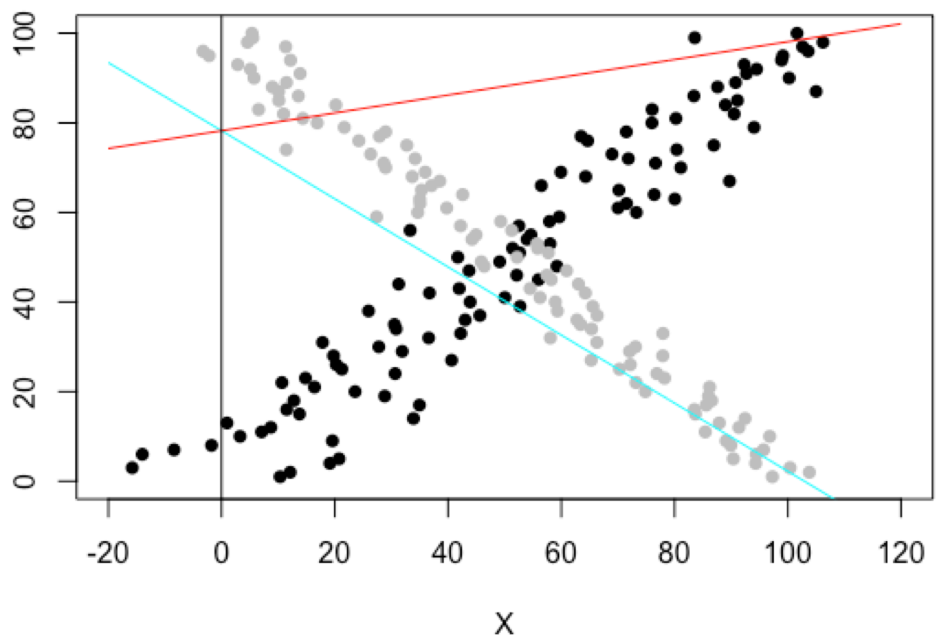
Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

More than one

```
> model <- lm(Y~ X + X2)
> coef(model)
(Intercept)           X          X2
 78.2467216   0.1987219  -0.7602549
```



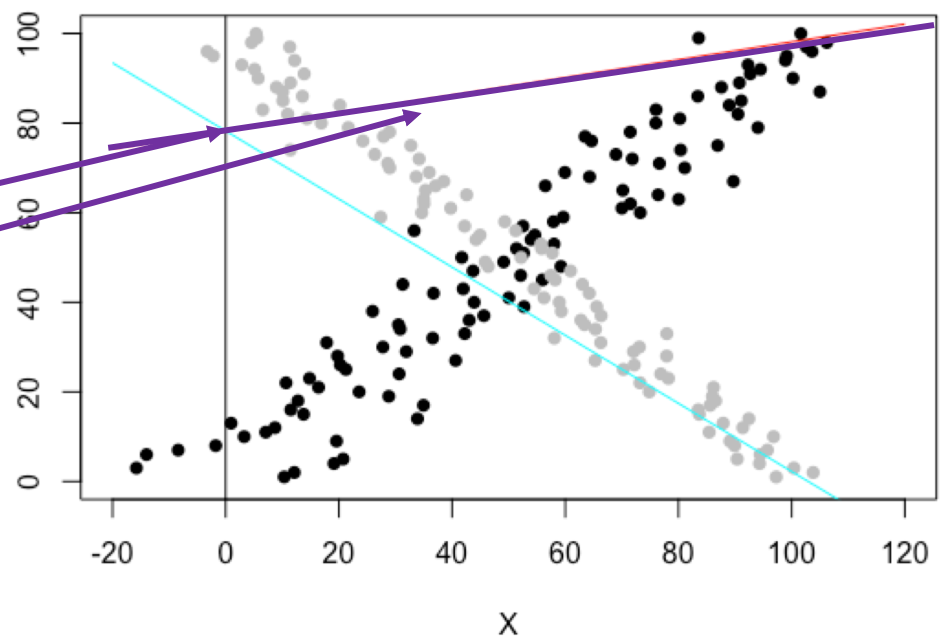
Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

More than one

```
> model <- lm(Y~ X + X2)
> coef(model)
(Intercept)           X          X2
 78.2467216   0.1987219  -0.7602549
```



$$Y_i = \alpha + \beta_1 X_i + \beta_2 X2_i + \varepsilon_i$$

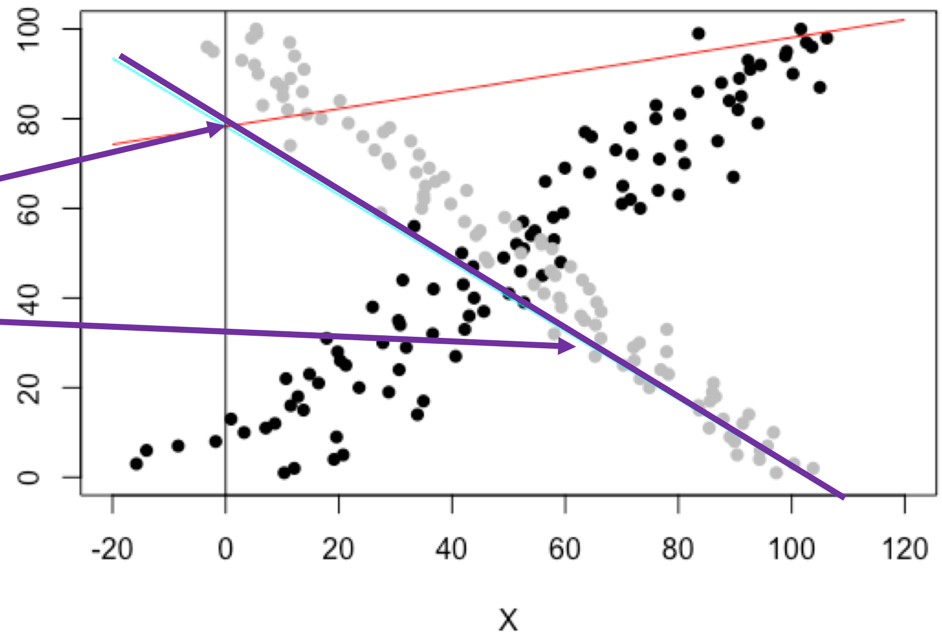
Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

More than one

```
> model <- lm(Y~ X + X2)
> coef(model)
(Intercept)           X          X2
 78.2467216   0.1987219  -0.7602549
```



$$Y_i = \alpha + \beta_1 X_i + \beta_2 X2_i + \varepsilon_i$$

Interpreting R outputs

Categorical vs continuous

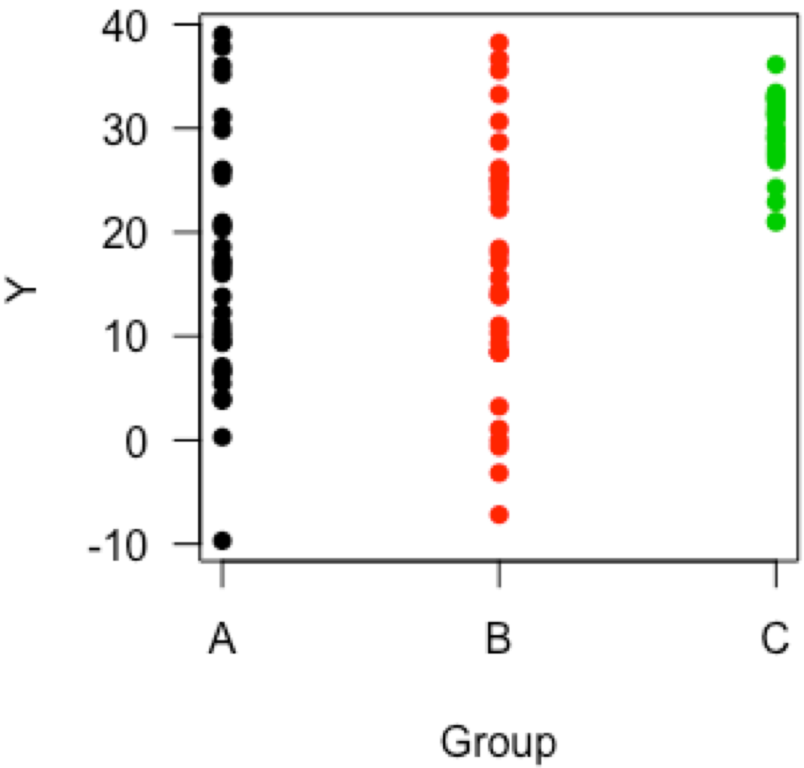
ONLY categorical explanatory variables

Interpreting R outputs

Categorical vs continuous

ONLY categorical explanatory variables

```
> coef(lm(Y~G))
(Intercept)          GB          GC 
 16.6924682   0.2848863  13.2697609 
```



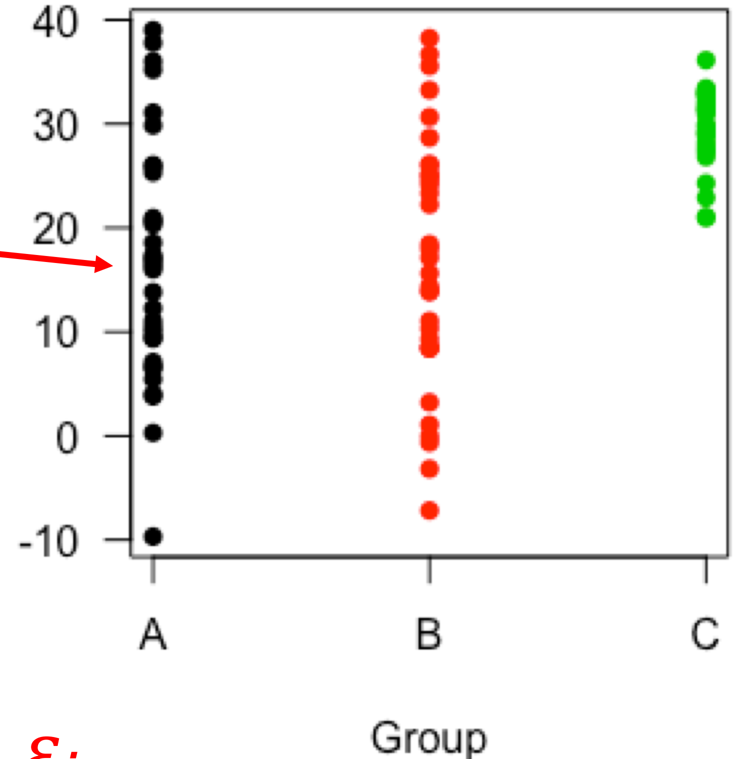
Interpreting R outputs

Categorical vs continuous

ONLY categorical explanatory variables

```
> coef(lm(Y~G))
(Intercept)          GB          GC
 16.6924682   0.2848863  13.2697609
```

When X=0 is the mean of group A



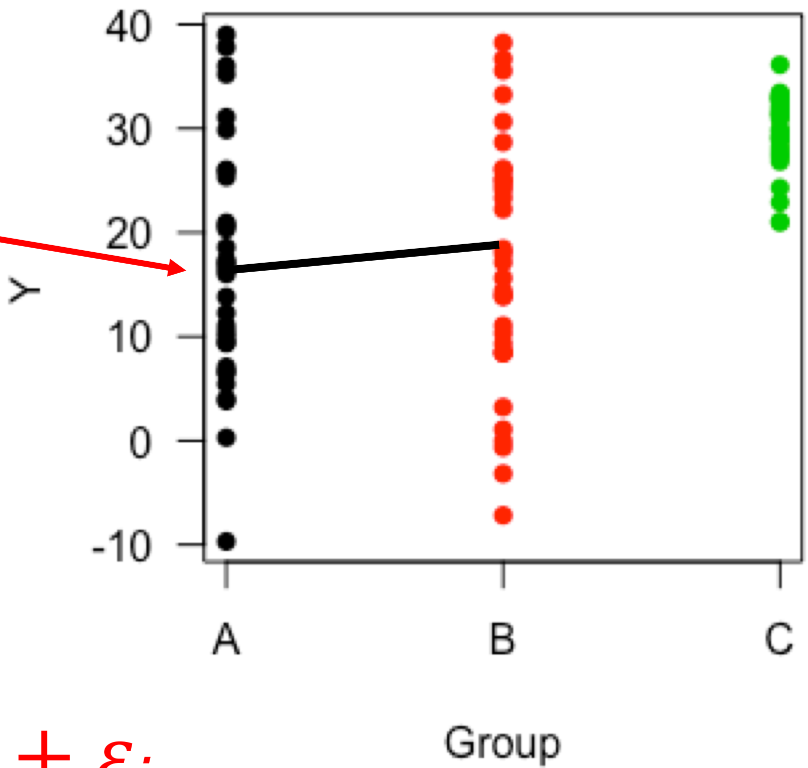
$$Y_i = \boxed{\alpha} + \beta_1 X_i + \beta_2 X2_i + \varepsilon_i$$

Interpreting R outputs

Categorical vs continuous

ONLY categorical explanatory variables

```
> coef(lm(Y~G))
(Intercept)          GB          GC 
 16.6924682   0.2848863  13.2697609 
```



The first beta value shows the slope one unit change in X, which is the same as the difference in mean between group A and group B

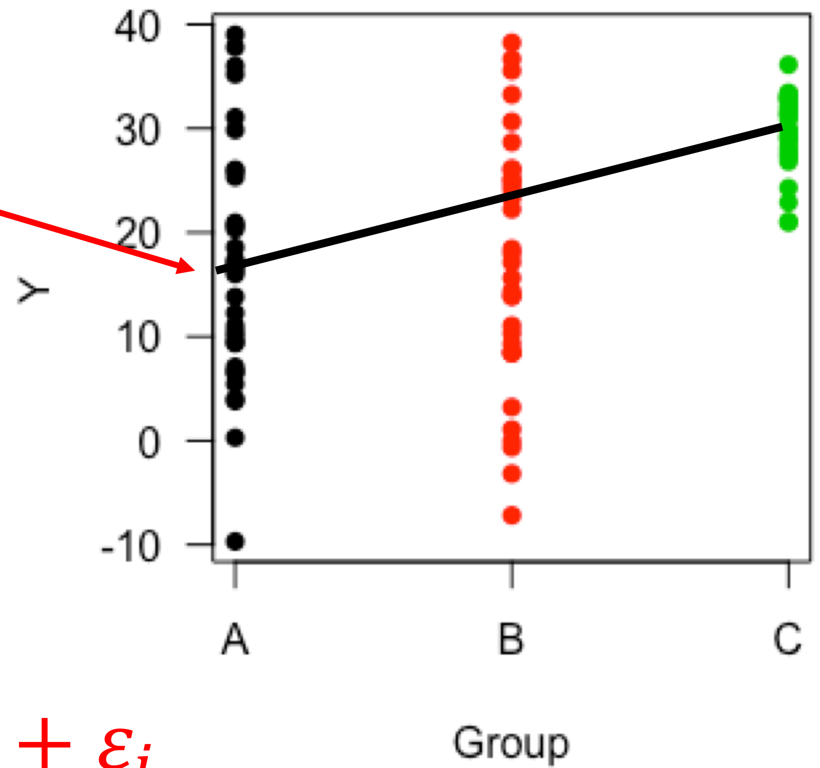
$$Y_i = \alpha + \boxed{\beta_1 X_i} + \beta_2 X2_i + \varepsilon_i$$

Interpreting R outputs

Categorical vs continuous

ONLY categorical explanatory variables

```
> coef(lm(Y~G))
(Intercept)          GB          GC
 16.6924682   0.2848863  13.2697609
```



The second beta value shows the slope for one unit change in X, which is the same as the difference in mean between group A and group C

$$Y_i = \alpha + \beta_1 X_i + \boxed{\beta_2 X2_i} + \varepsilon_i$$

Interpreting R outputs

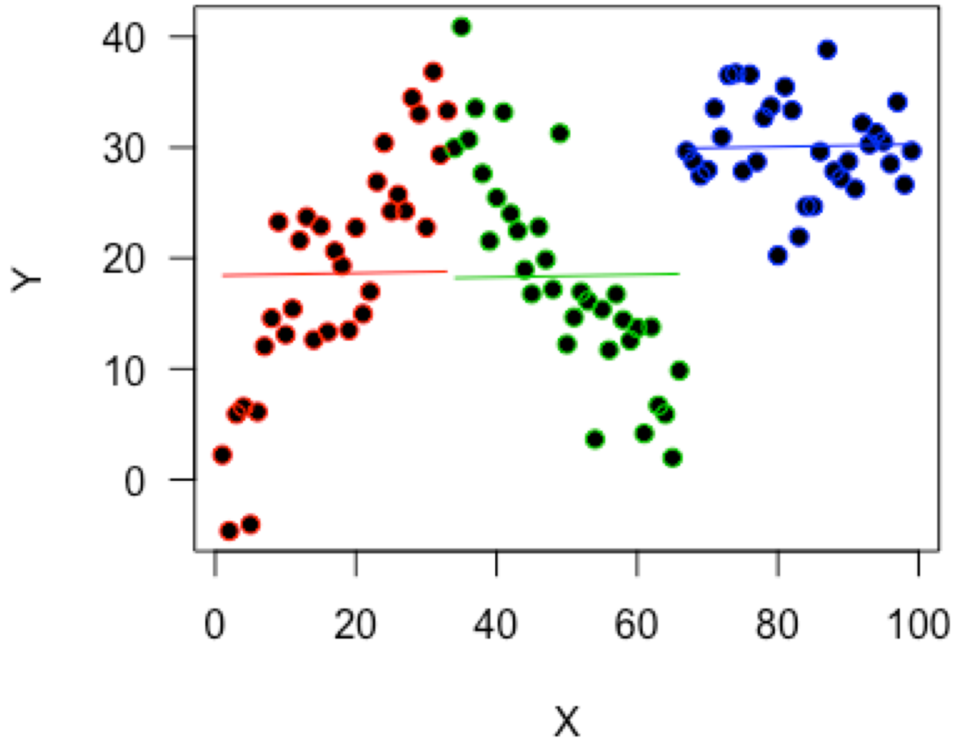
Categorical vs continuous

BOTH categorical and continuous explanatory variables

Interpreting

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

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



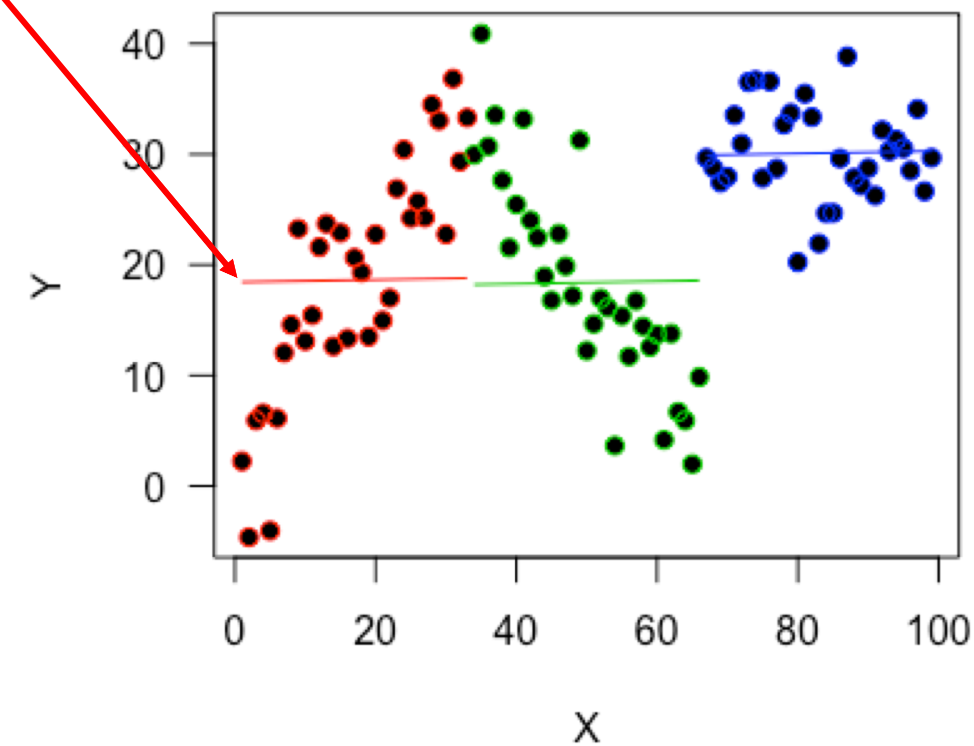
Interpreting

$$Y_i = \boxed{\alpha +} \beta_x X_i + \beta_B XcatB_i + \beta_C XcatC_i + \varepsilon_i$$

```
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



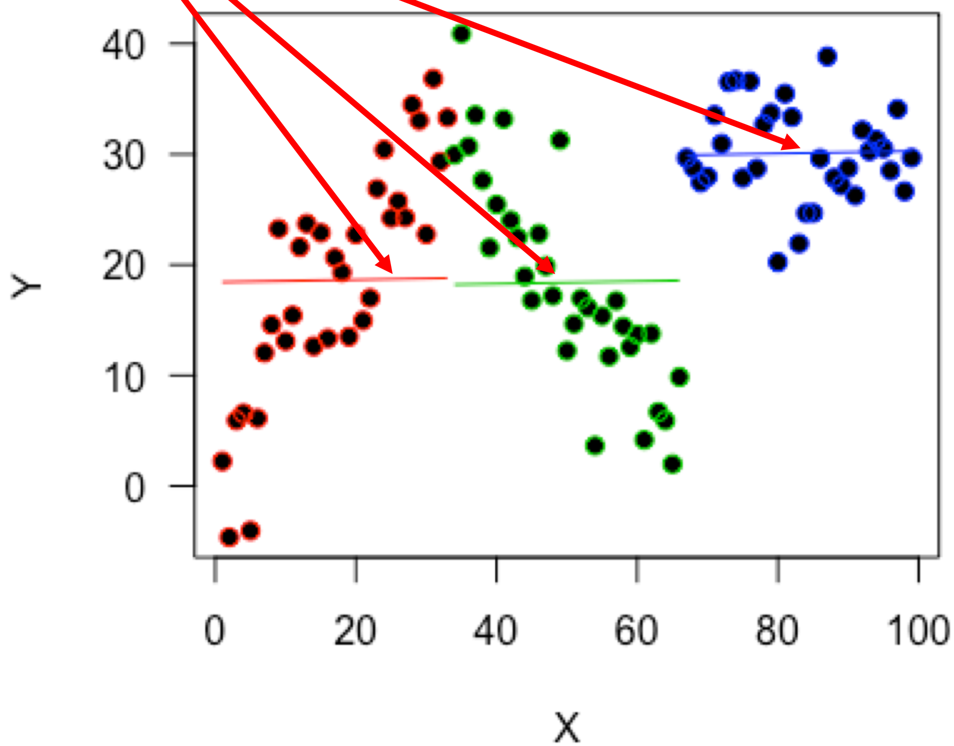
Interpreting

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

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

$$Y_i = \alpha + \boxed{\beta_x X_i} + \beta_B XcatB_i + \beta_C XcatC_i + \varepsilon_i$$

Slope value for all groups (same)



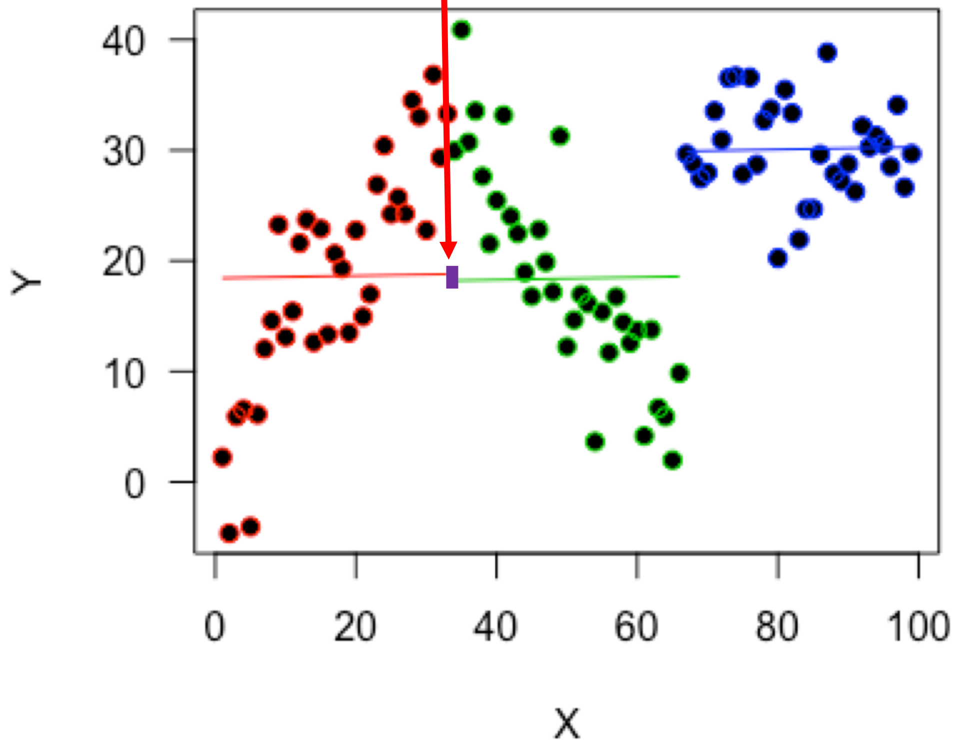
Interpreting

$$Y_i = \alpha + \beta_x X_i + \boxed{\beta_B XcatB_i} + \beta_C XcatC_i + \varepsilon_i$$

```
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



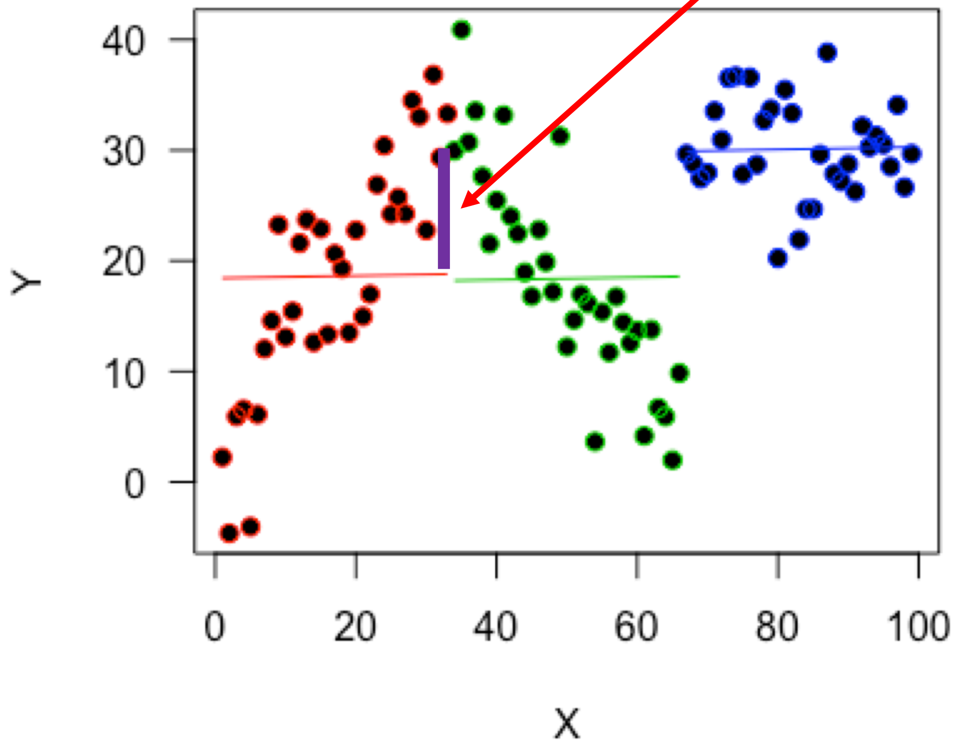
Interpreting

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

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

$$Y_i = \alpha + \beta_x X_i + \beta_B XcatB_i + \boxed{\beta_C XcatC_i} + \varepsilon_i$$

Difference in intercept from Group A to Group C

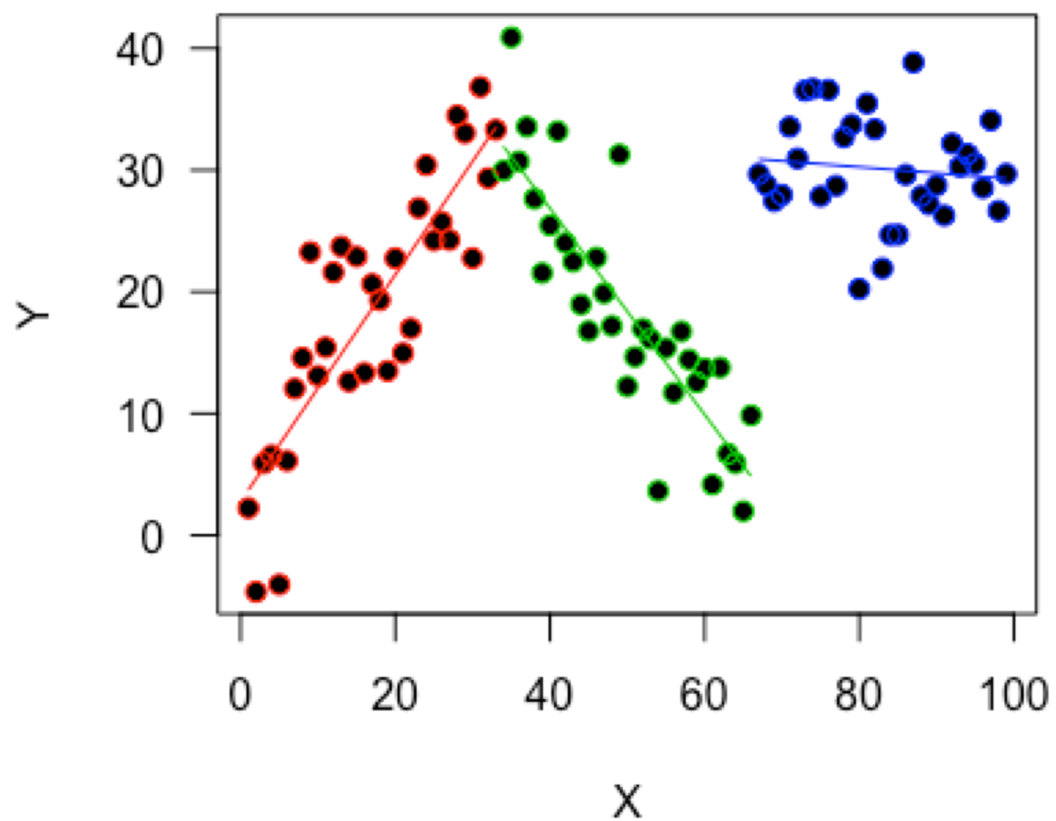


Interpreting

```
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 
```

Interaction includes differences in slope



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 
```

An example: predict Y for X = 40, G = B

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
```

An example: predict Y for X = 40, G = B

$$Y_i = \alpha + \beta_B XcatB_i + ((\beta_x + \beta_{intB})X_i) + \varepsilon_i$$

26.74 = 2.78 + (57.96*1) + ((0.93+-1.78)*40)

Interpreting R outputs

Things to remember:

Everything is based on this $Y_i = \alpha + \beta X_i + \varepsilon_i$

Check whether your explanatory variables are categorical or continuous before interpreting

Sometimes there will be differences as well as slopes and intercepts

Interpreting R outputs

Other bits:

Look out for interactions indicated by * in the model and : in the output e.g. X:GB

In glms need to consider the link too, especially for the intercept and interpretation of predictions

Model checking

Model checking

For all models, we make some assumptions.

During checking, we determine if these assumptions are met.

Model checking

Reminder of assumptions (LM):

Residuals (error) is normally distributed

Error has a mean of 0

The relationship is linear

The variance is equal for all fitted values

No outliers

Independence of Y

Model checking

Reminder of assumptions (GLM):

~~Residuals (error) is normally distributed~~ Correct distribution used

~~Error has a mean of 0~~ Correct link function is used

The relationship is linear on the link scale

~~The variance is equal for all fitted values~~ Dispersion parameter is constant

No outliers

Independence of Y

Model checking

Reminder of assumptions (GLM):

~~Residuals (error) is normally distributed~~ Correct distribution used

~~Error has a mean of 0~~ Correct link function is used

The relationship is linear on the link scale

~~The variance is equal for all fitted values~~ Dispersion parameter is constant

No outliers

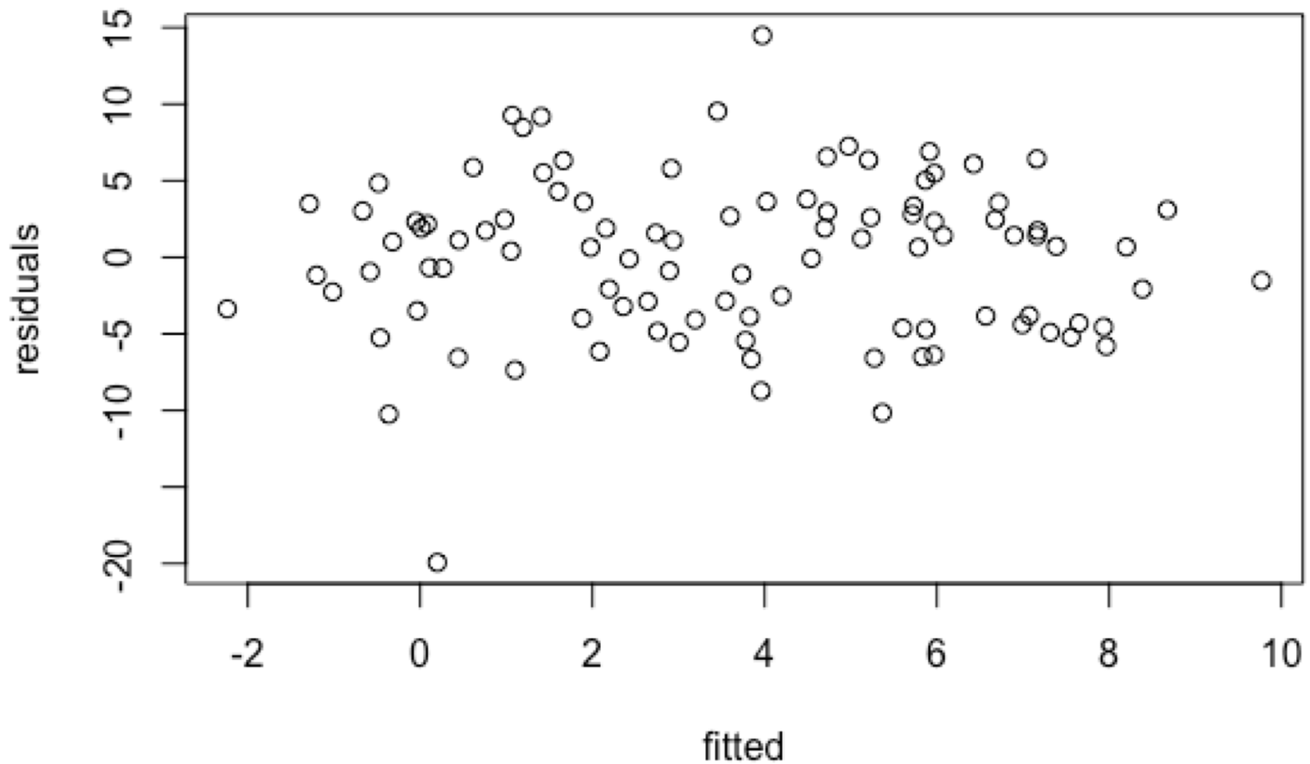
Independence of Y

Model checking

Linearity and equal variance: residual vs fitted plot

Model checking

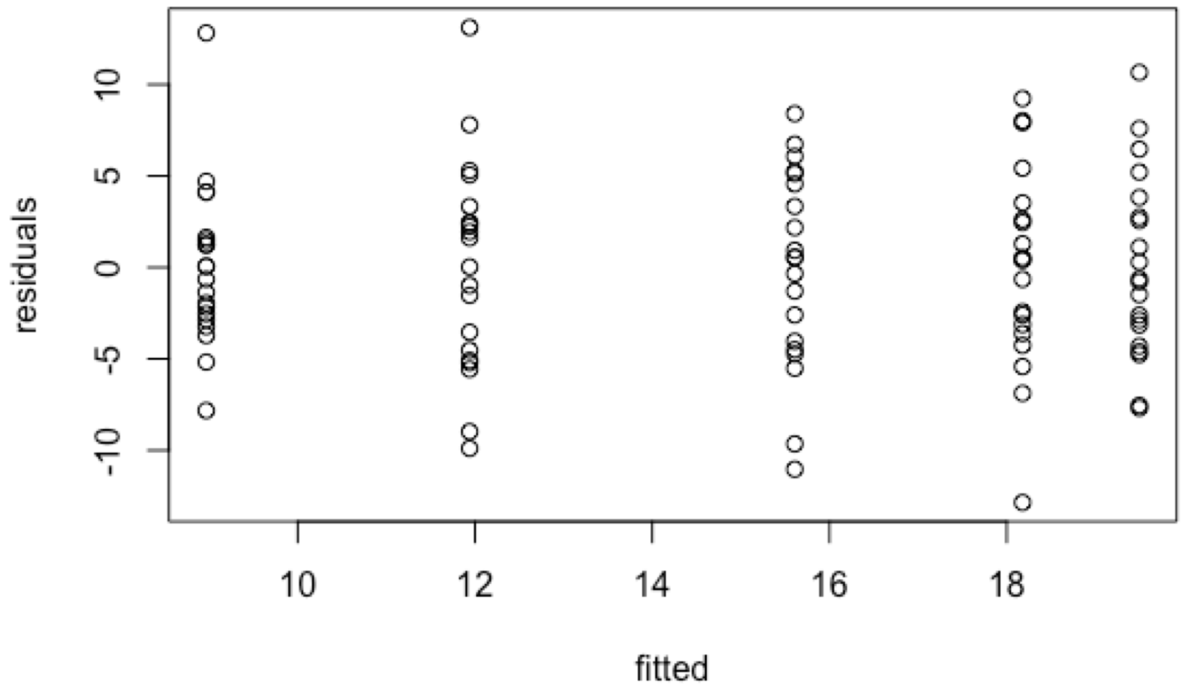
Linearity and equal variance: residual vs fitted plot



OK

Model checking

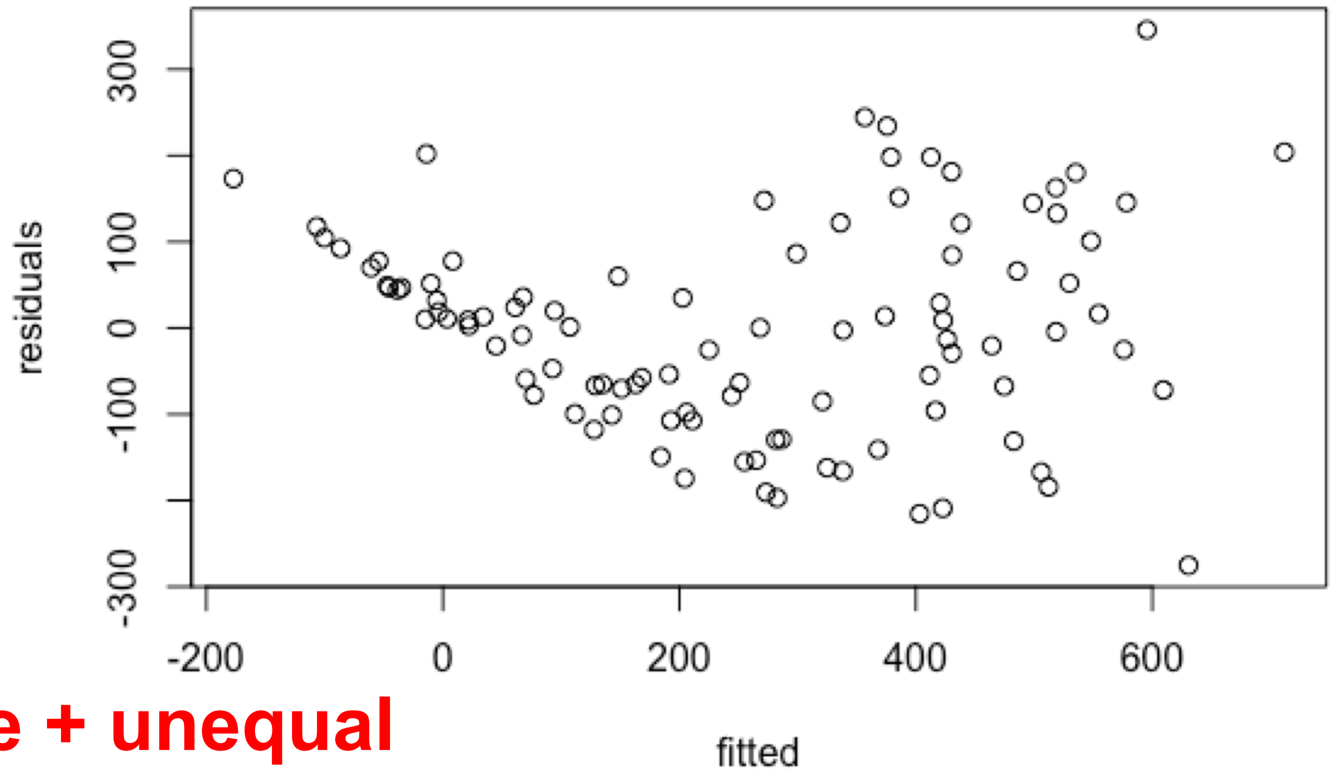
Linearity and equal variance: residual vs fitted plot



OK

Model checking

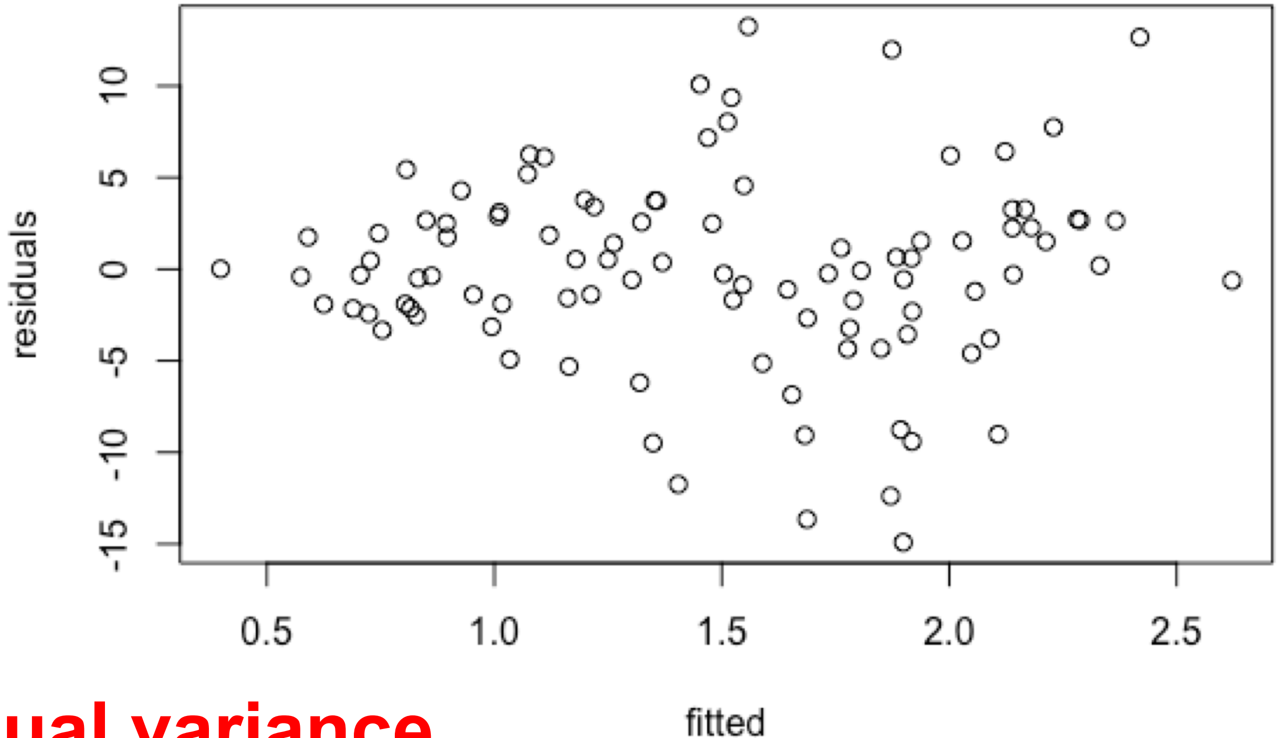
Linearity and equal variance: residual vs fitted plot



Curve + unequal variance

Model checking

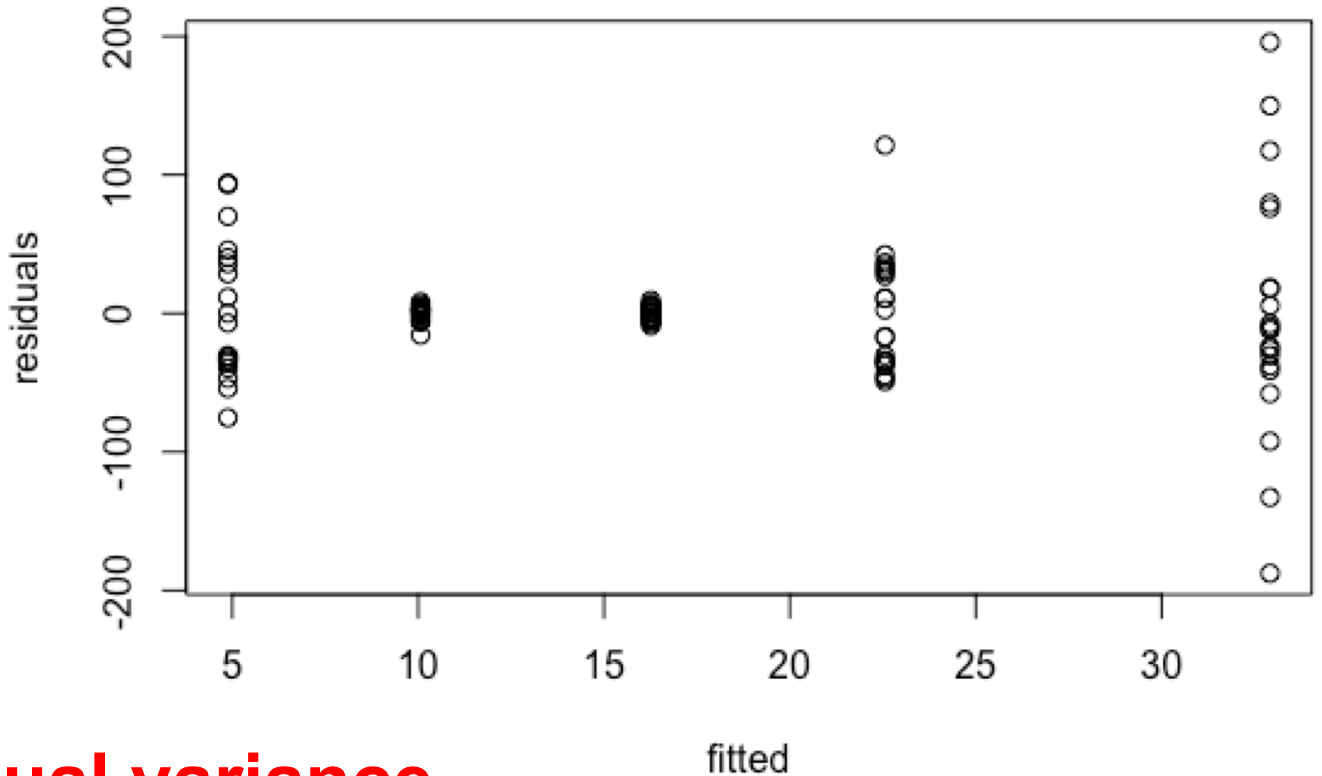
Linearity and equal variance: residual vs fitted plot



Unequal variance

Model checking

Linearity and equal variance: residual vs fitted plot



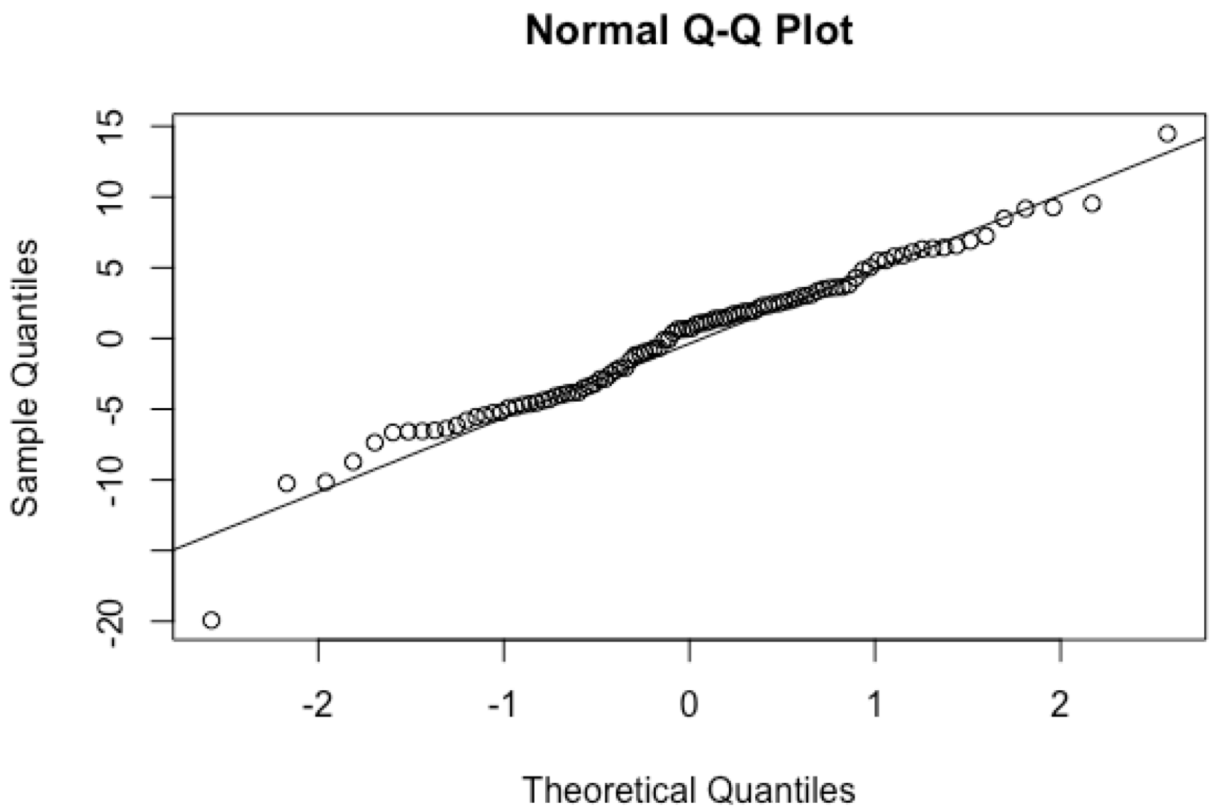
Unequal variance

Model checking

Normality of residuals: normal QQ

Model checking

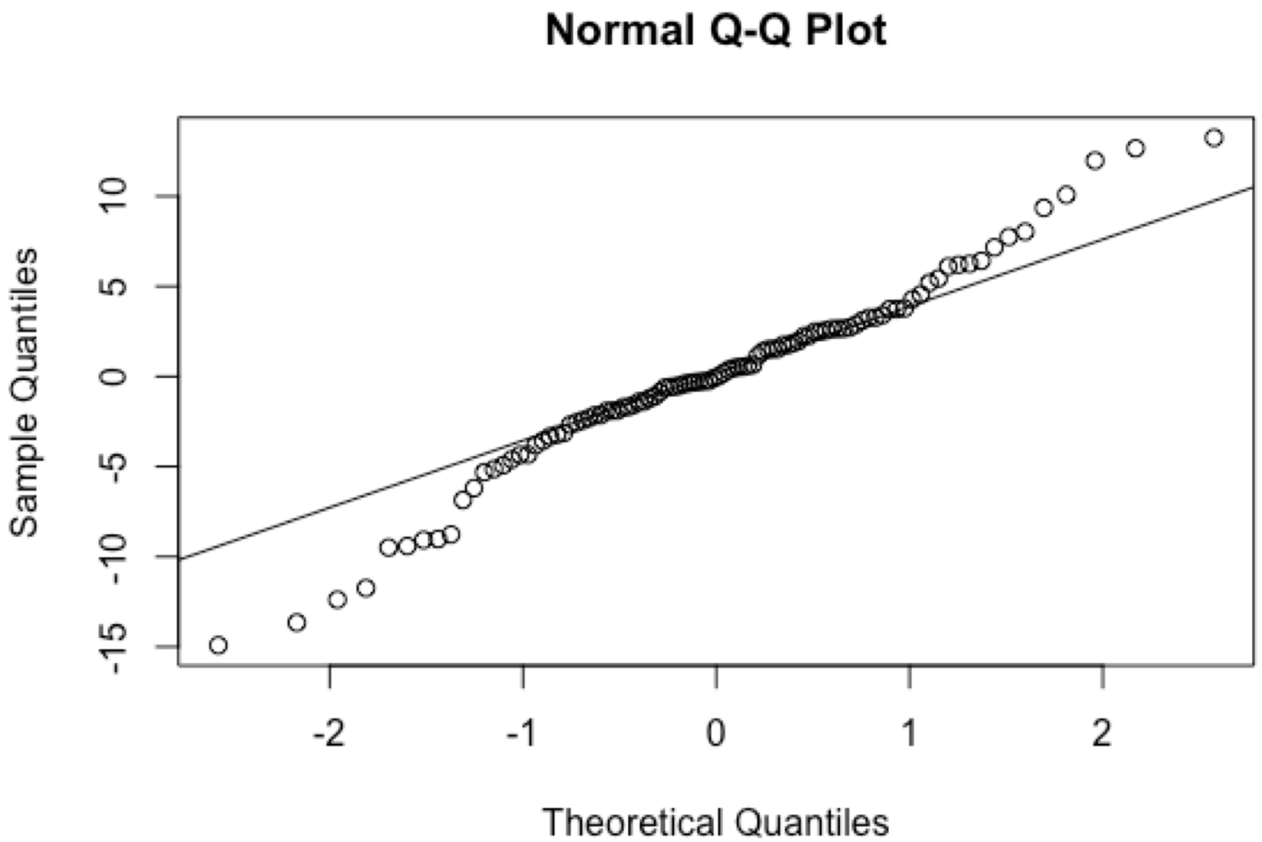
Normality of residuals: normal QQ



OK

Model checking

Normality of residuals: normal QQ



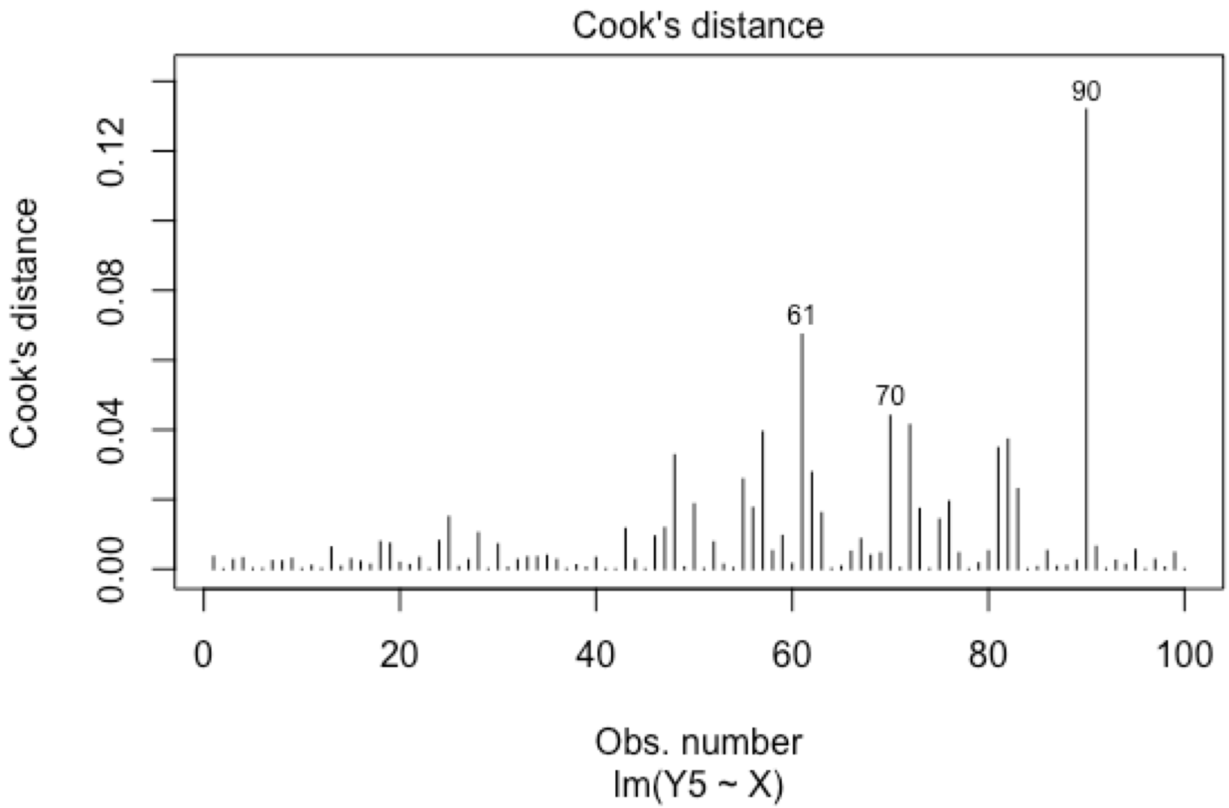
not OK

Model checking

Outliers: Cook's Distance

Model checking

Outliers: Cook's Distance



Model checking GLMS: Overdispersion

Overdispersion

When there is more variation than the model assumes

Overdispersion

An example: using Poisson GLM

Overdispersion

What we know about GLMs:

Systematic part = $\alpha + \beta X_i$

Random part = the error around this (the Poisson bit)

A link function = here log()

Overdispersion

What we know about GLMs:

Systematic part = $\alpha + \beta X_i$

$\boxed{\text{Random part = the error around this (the Poisson bit)}}$

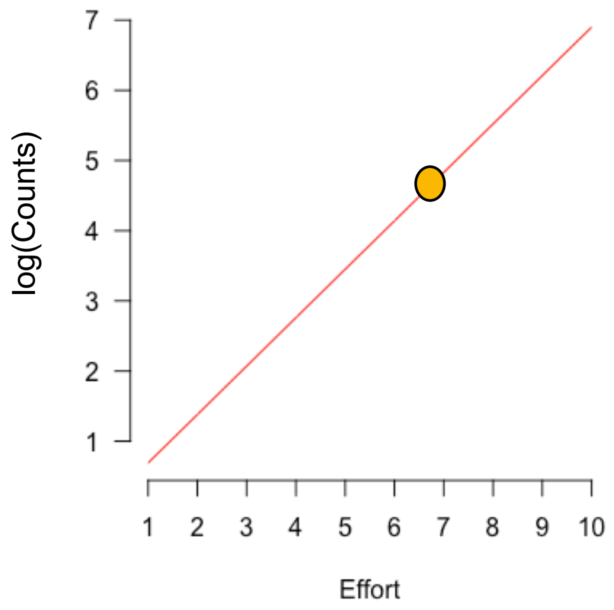
A link function = here log()

Overdispersion

Systematic part = $\alpha + \beta X_i$

Gives us the fitted values on the link scale:

$$E(\log(Y)) = \alpha + \beta X_i$$



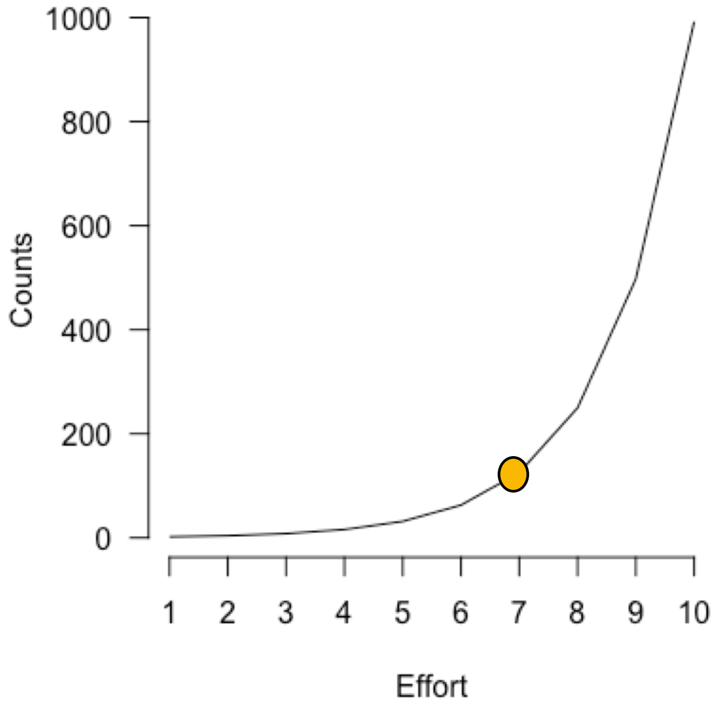
Predicting the expected value of Y on the link scale

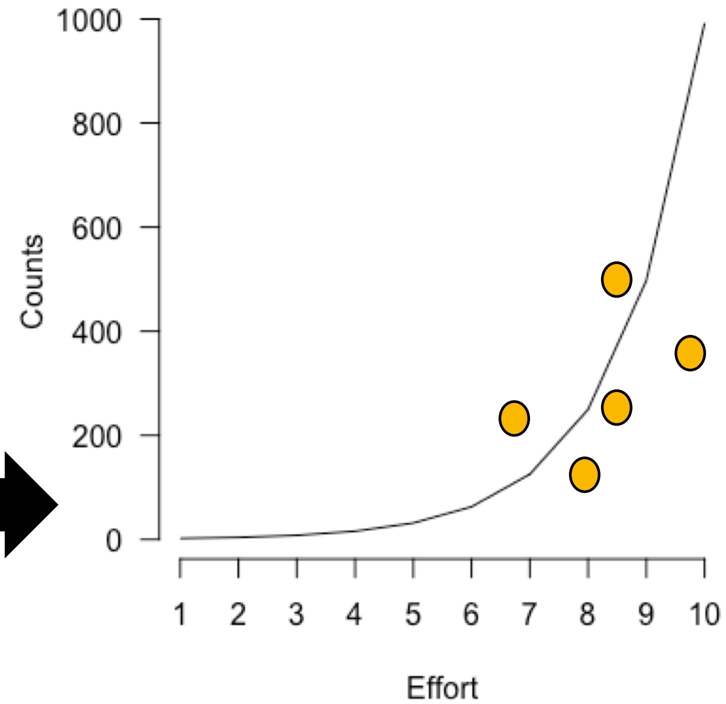
Overdispersion

The Random part (the Poisson bit)

Data will not always lie exactly on our estimated line

Need to capture difference between data and model (residuals)





Overdispersion

For the Poisson, we assume that the variance = the mean

So, dispersion = 1 and is constant

Overdispersion

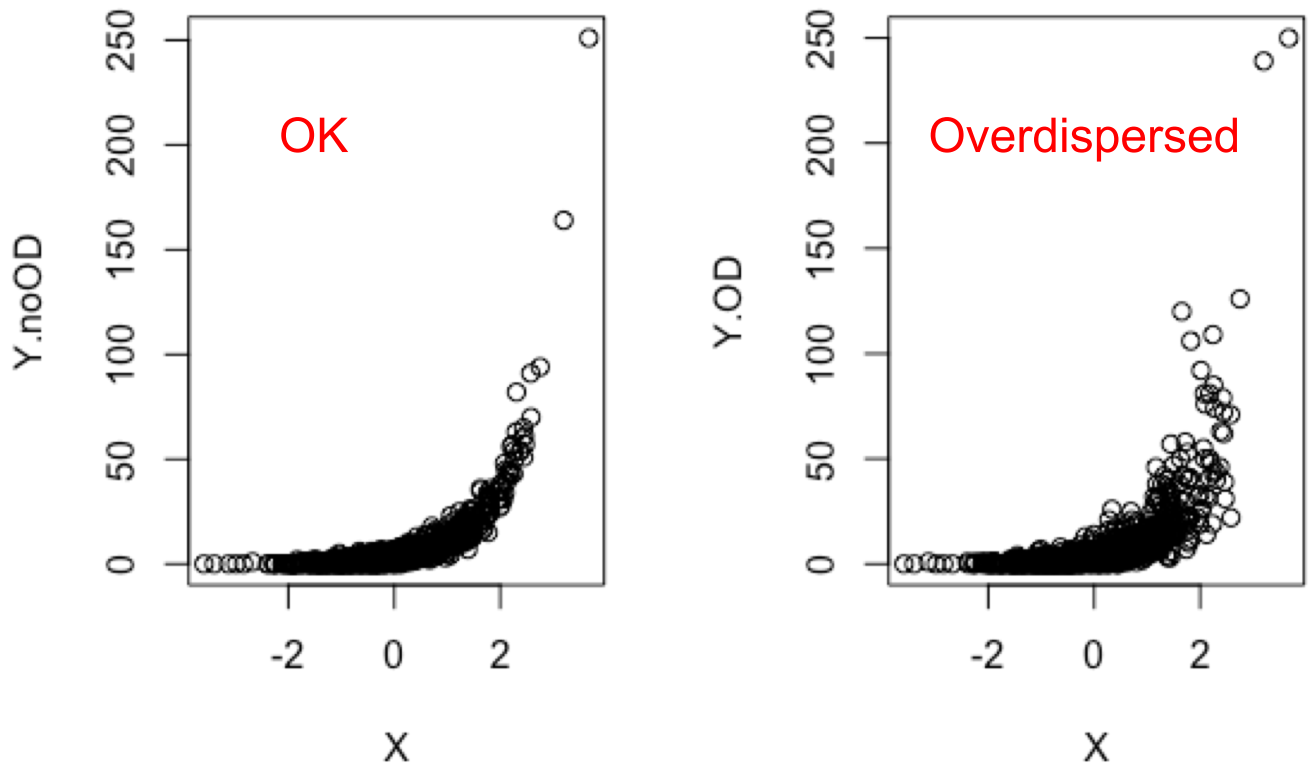
Can be caused by another variable we didn't measure

If you don't account for it – uncertainty is too narrow!

Overdispersion

For the Poisson, we assume that the variance = the mean

This means dispersion = 1 and is constant



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
>1.2 is a problem

Not good here

Overdisperion

Fix by correcting the likelihood OR

Use a negative binomial GLM

Code in Exercise 11 and Poisson GLM module