Revision session 16.5.19

Outline

Exploratory model selection

Interpreting R outputs

Any other questions

Exploratory model selection

Exploratory model selection

Why do we need it?

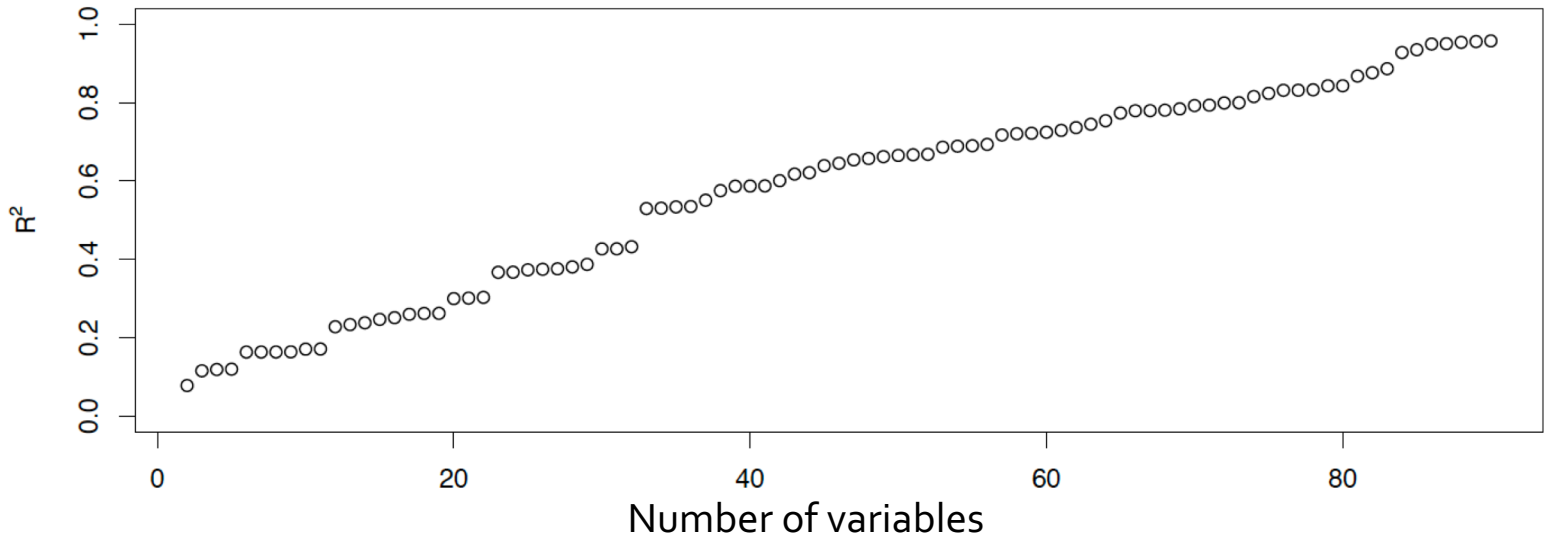
Exploratory model selection

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(sample size)*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

Interpreting R outputs

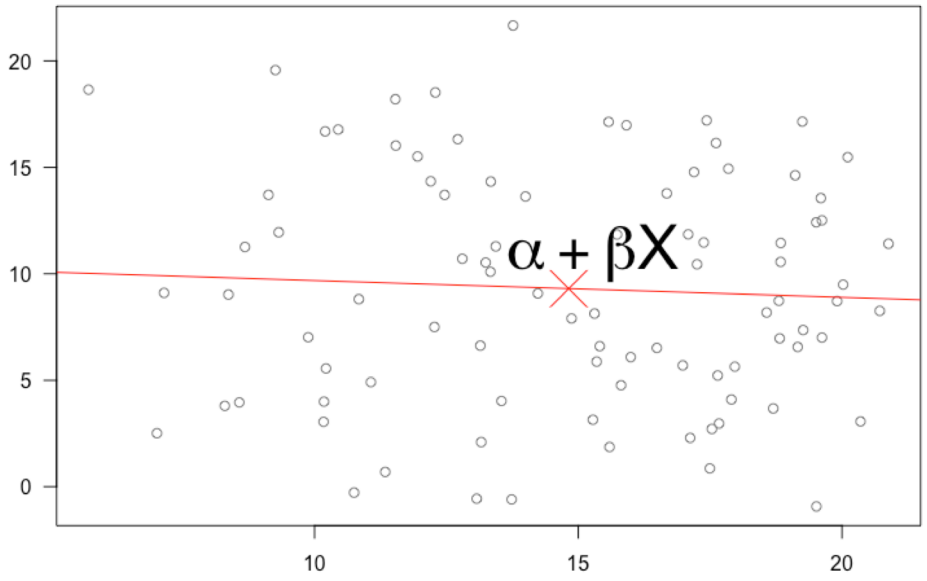
Linear models (glm) or (lm) are all based on this equation

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

Interpreting R outputs

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

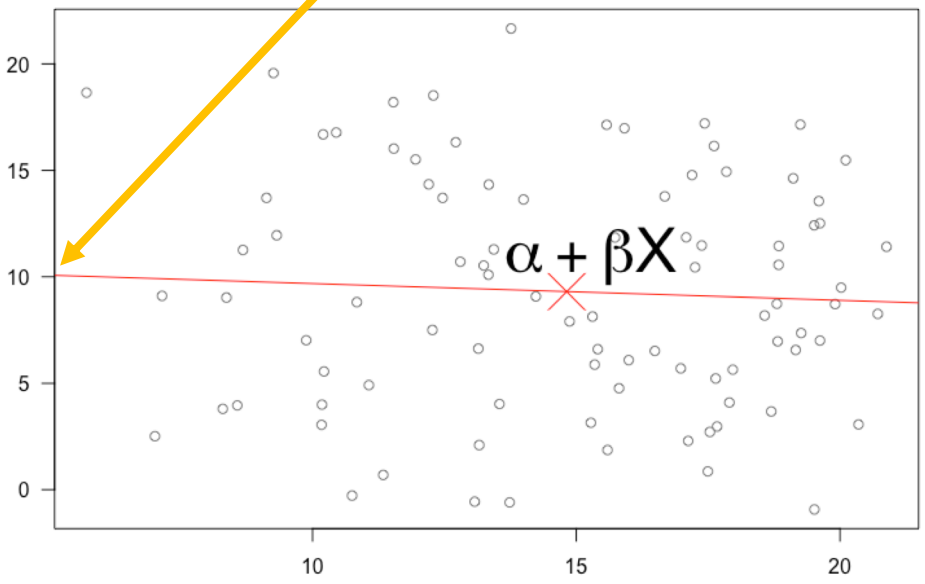
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Interpreting R outputs

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

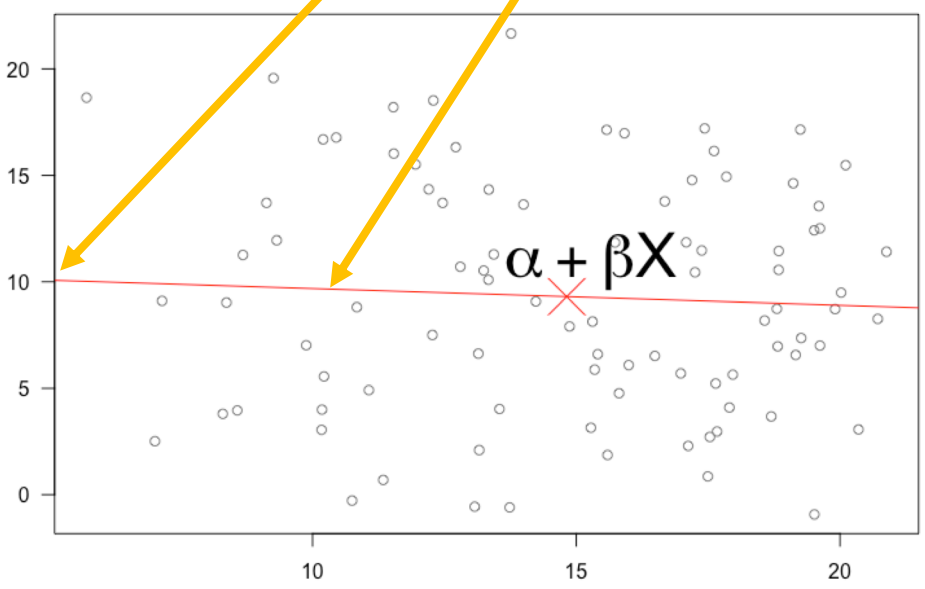
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Interpreting R outputs

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

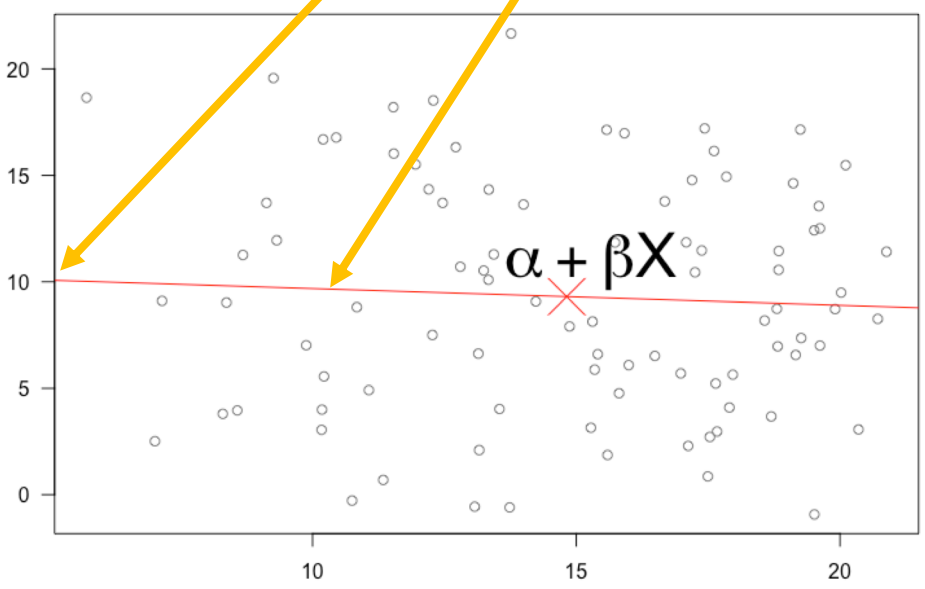
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Interpreting R outputs

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

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

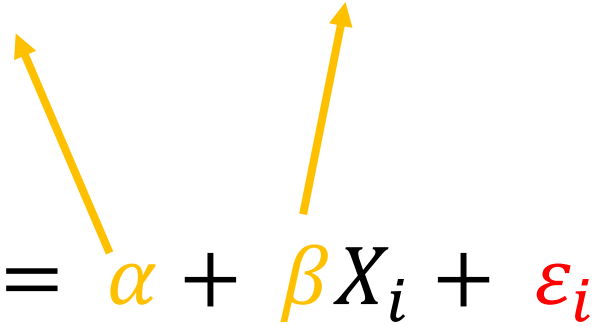
```
> coef(model)
(Intercept)           X
   2.358216    0.010118
```

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

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
```

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

Interpreting R outputs

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

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

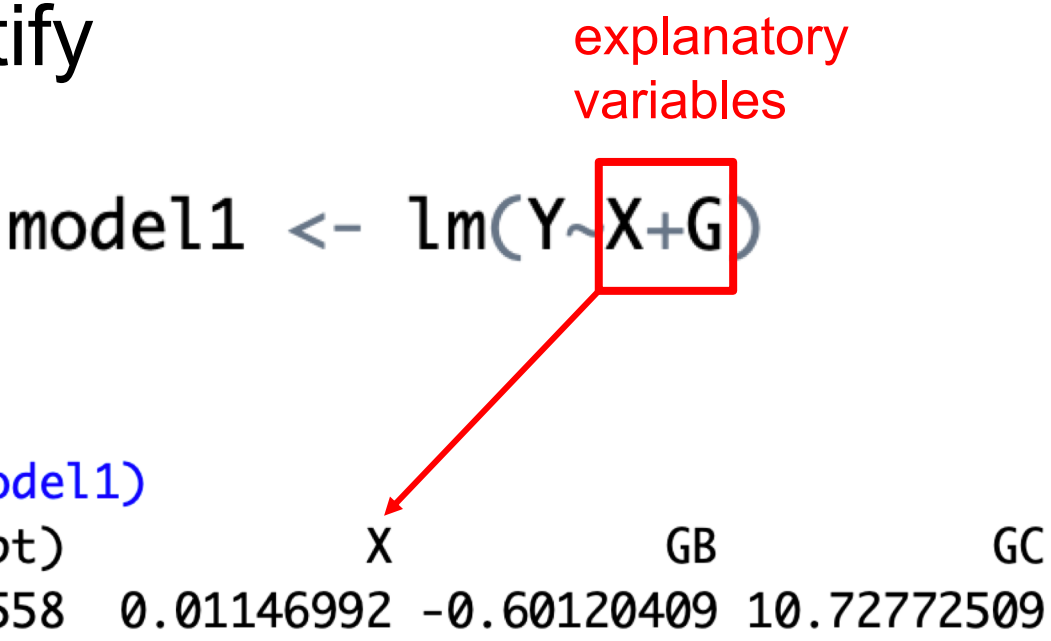
```
model1 <- lm(Y~X+G)
```

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

Interpreting R outputs

Categorical vs continuous

How to identify

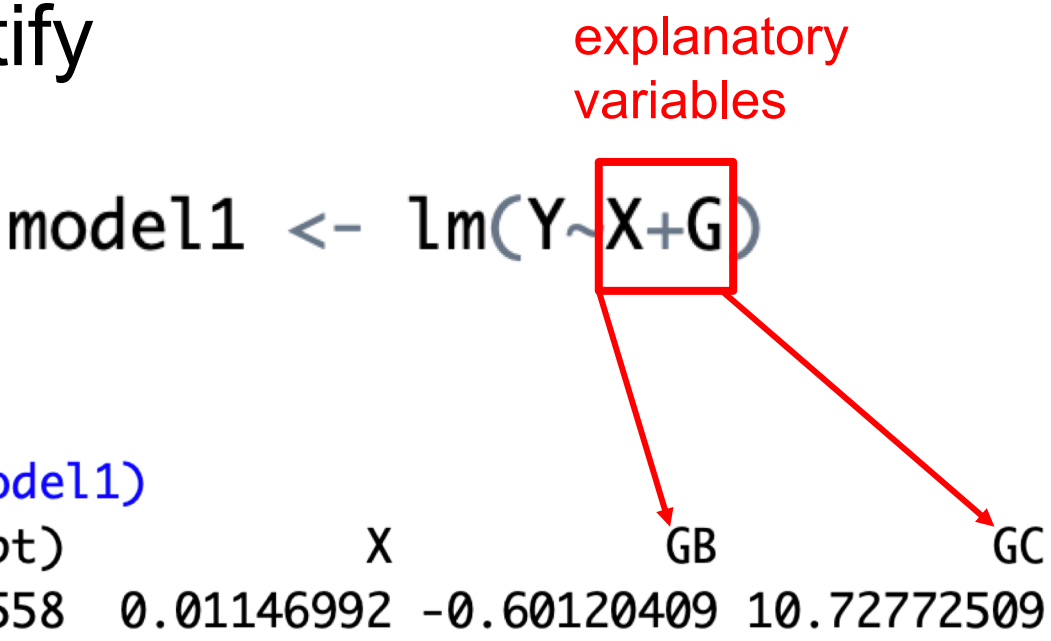


Continuous = single value with same name as variable

Interpreting R outputs

Categorical vs continuous

How to identify



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

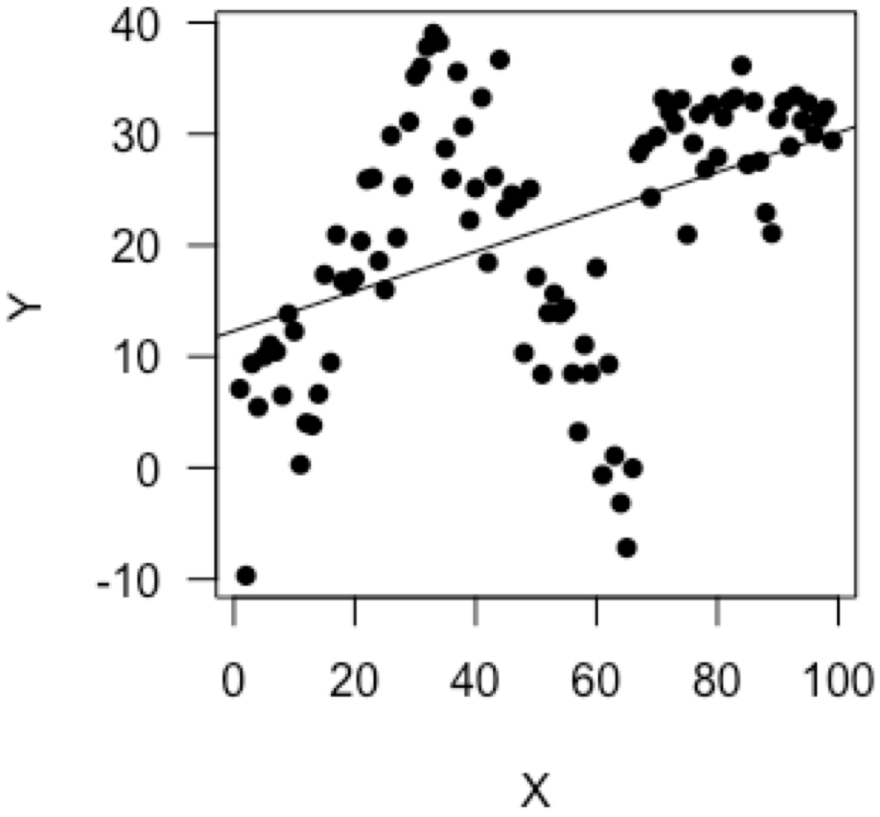
ONLY continuous explanatory variables

Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

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

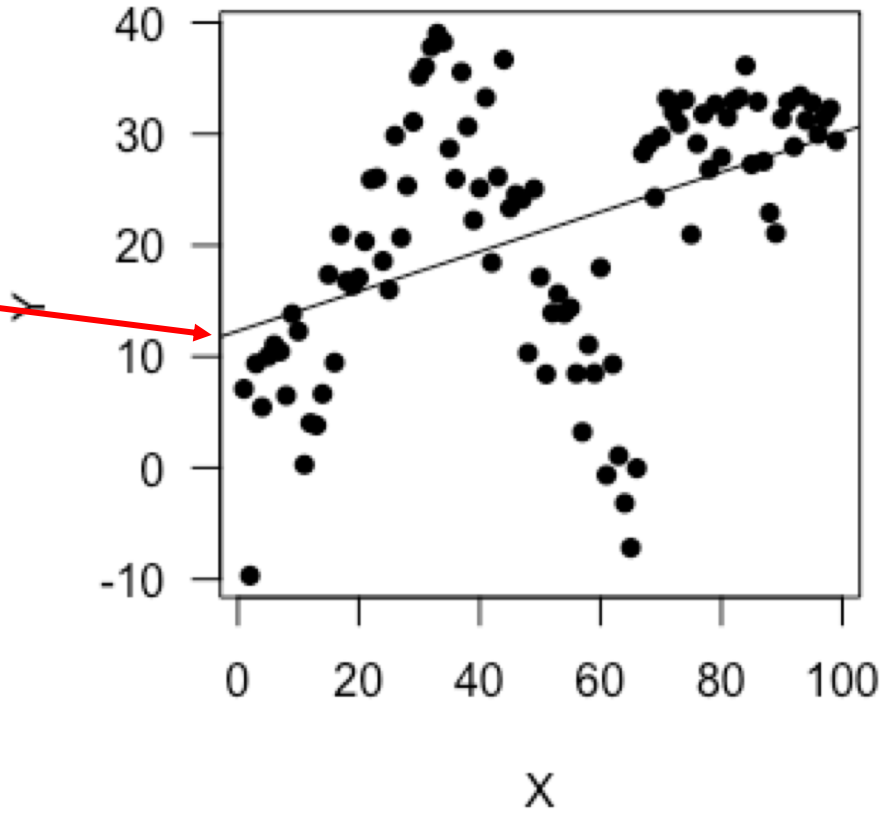


Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

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

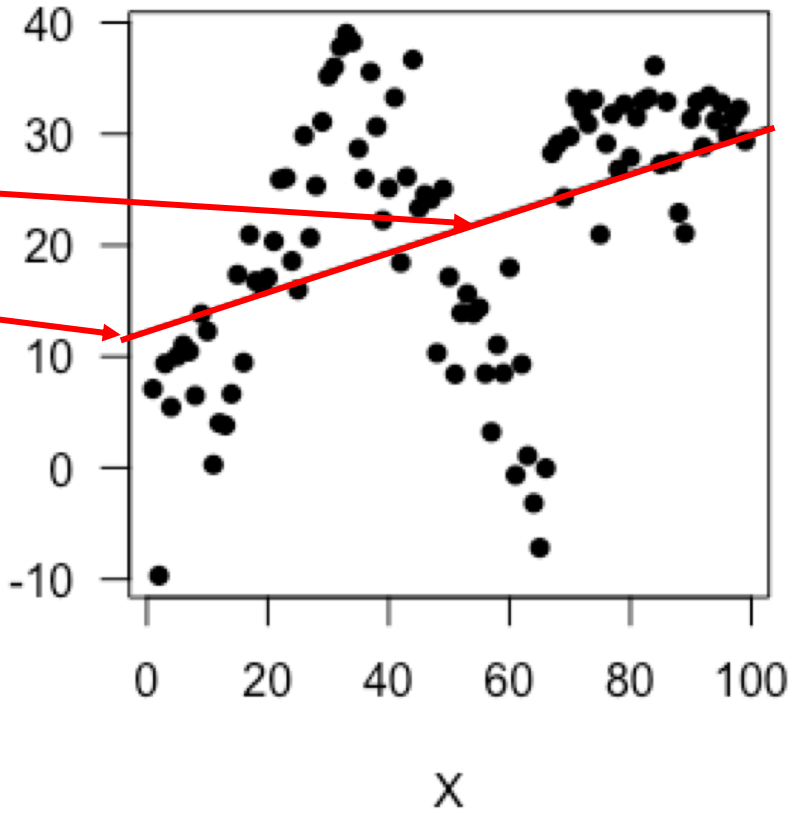


Interpreting R outputs

Categorical vs continuous

ONLY continuous explanatory variables

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



Interpreting R outputs

Categorical vs continuous

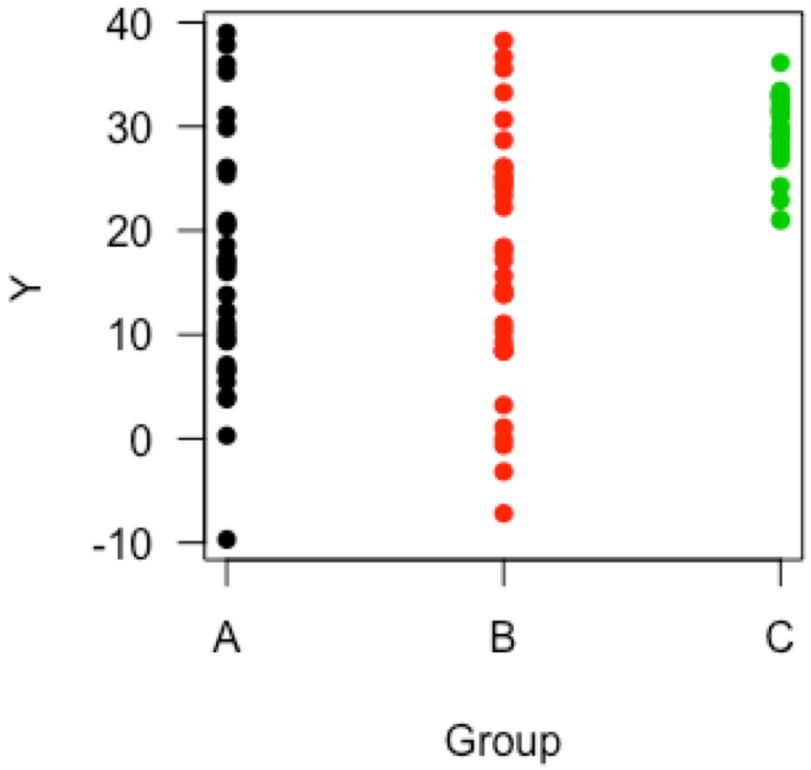
ONLY categorical explanatory variables

Interpreting R outputs

Categorical vs continuous

ONLY categorical explanatory variables

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



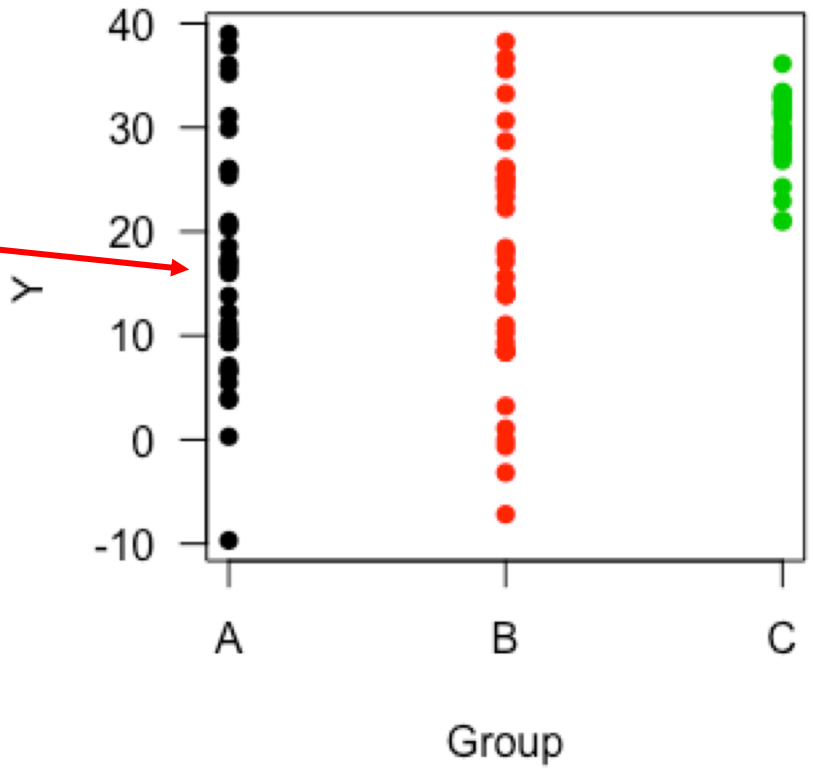
Interpreting R outputs

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



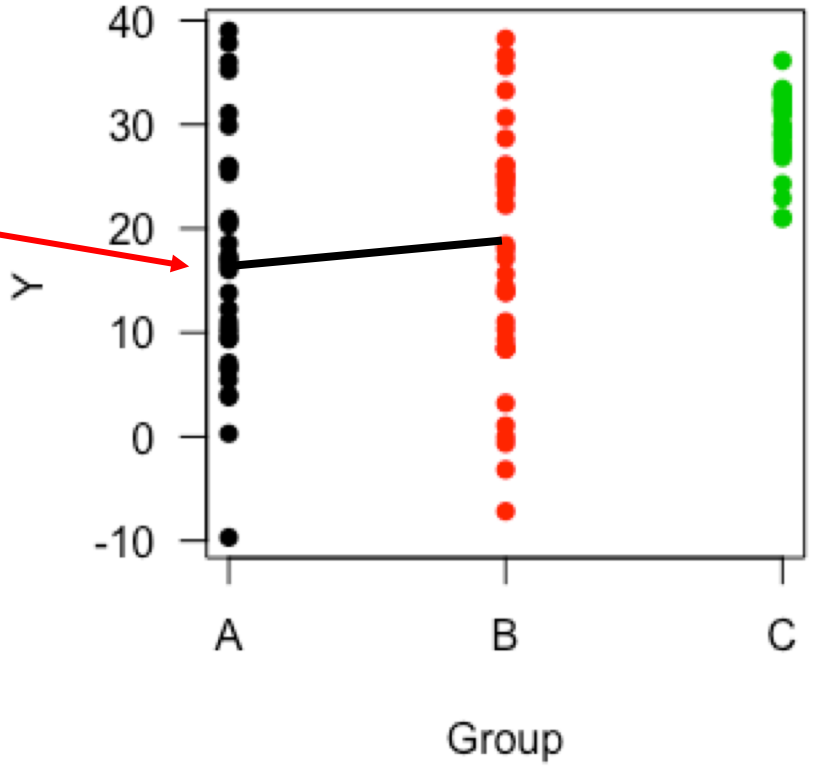
Interpreting R outputs

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



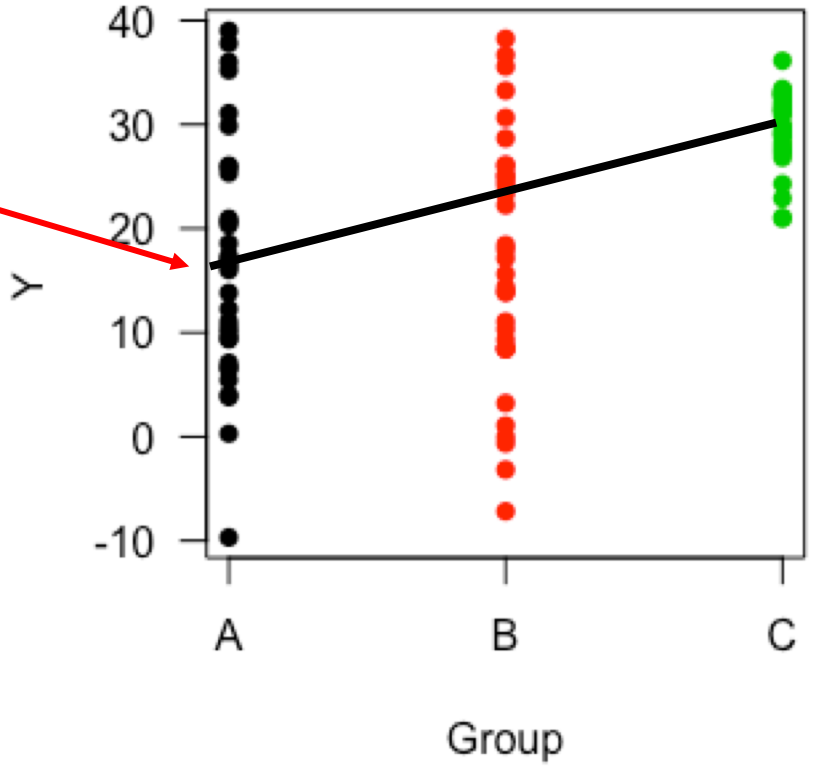
Interpreting R outputs

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



Interpreting R outputs

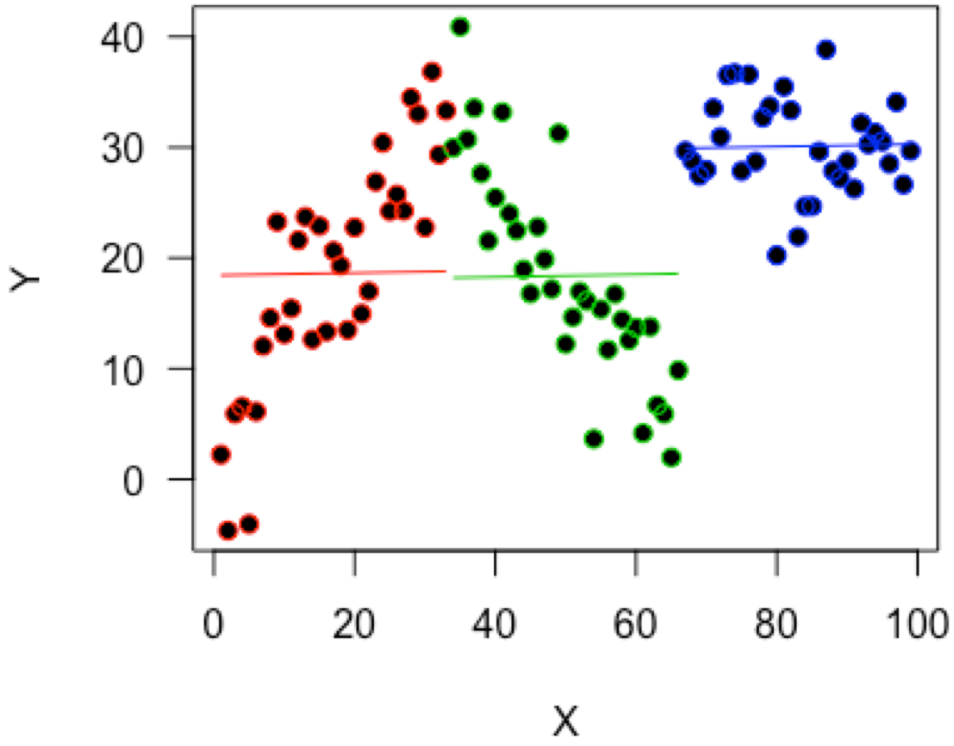
Categorical vs continuous

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
```

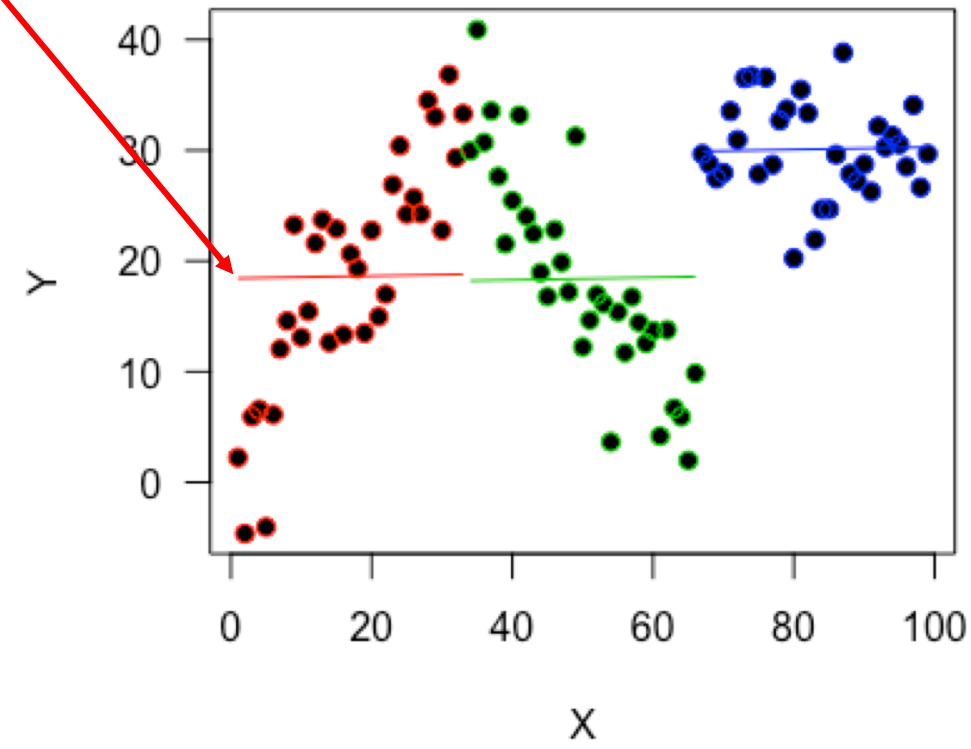


Interpreting

```
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

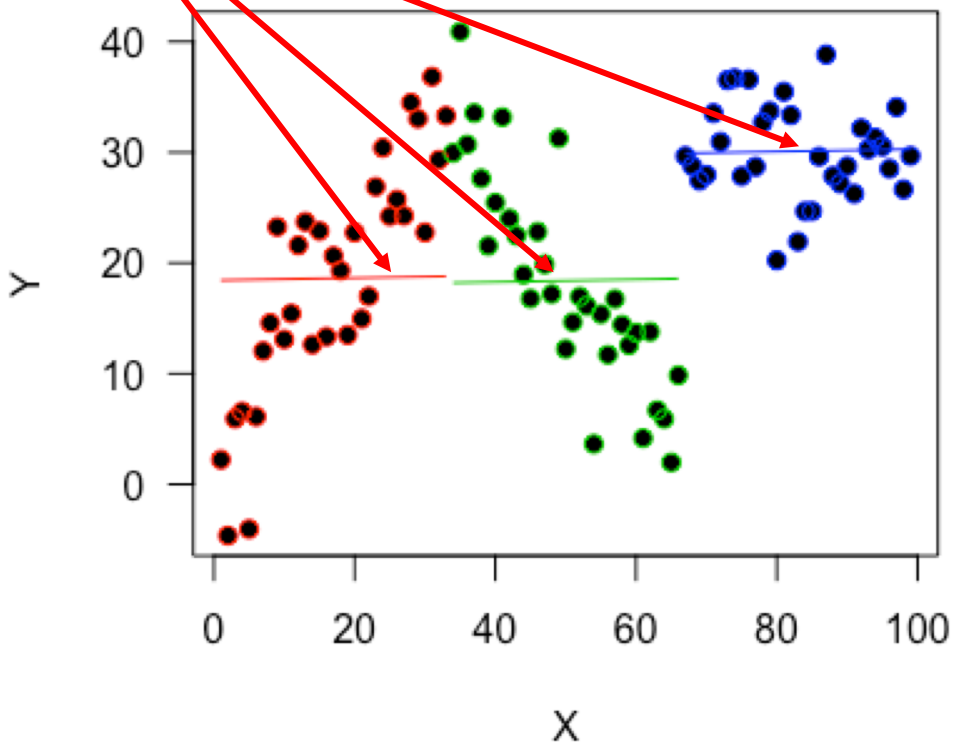


Interpreting

```
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)

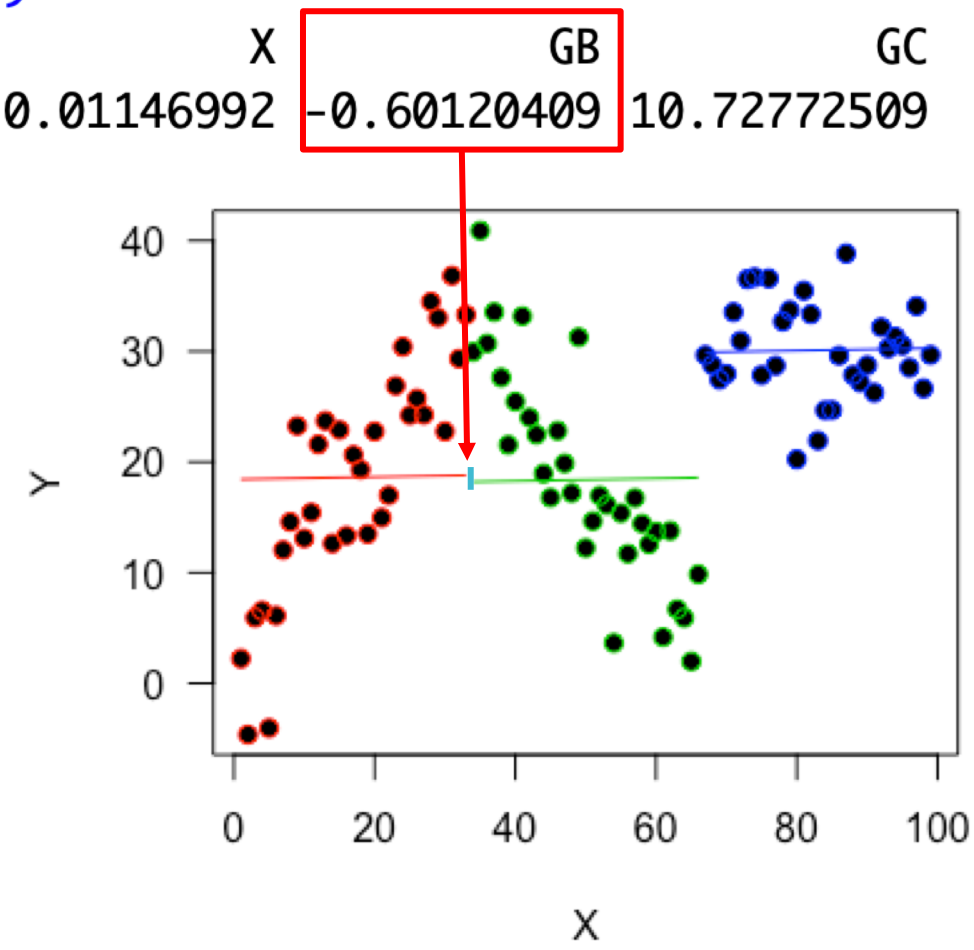


Interpreting

```
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

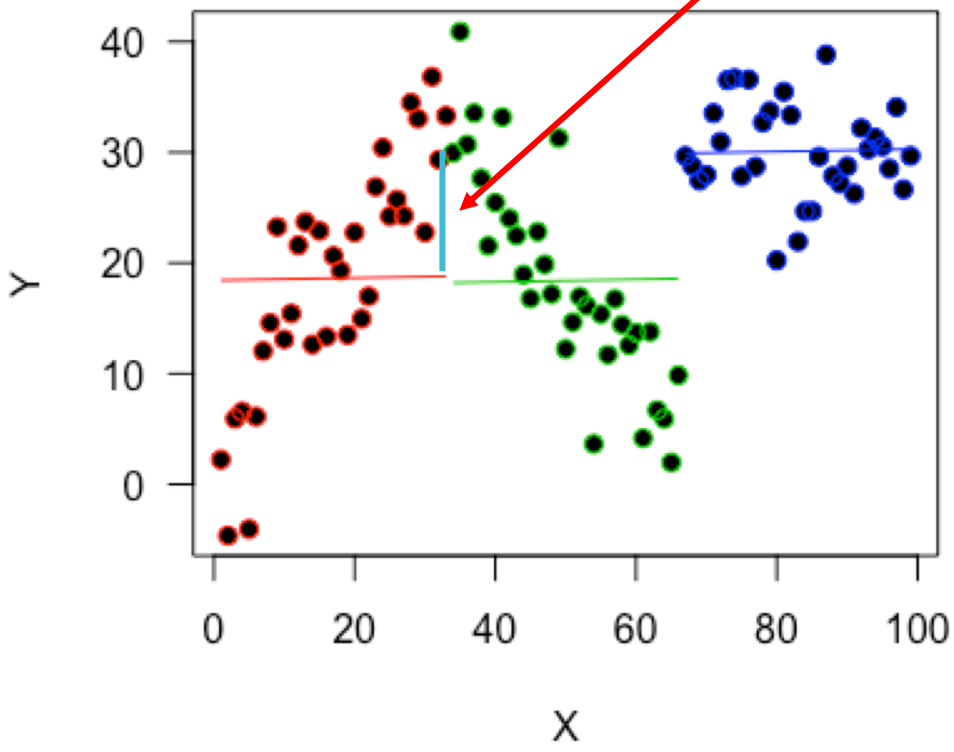


Interpreting

```
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 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

Any other questions?