

FEEG6017 lecture: Transformations, polynomial fitting, and interaction terms

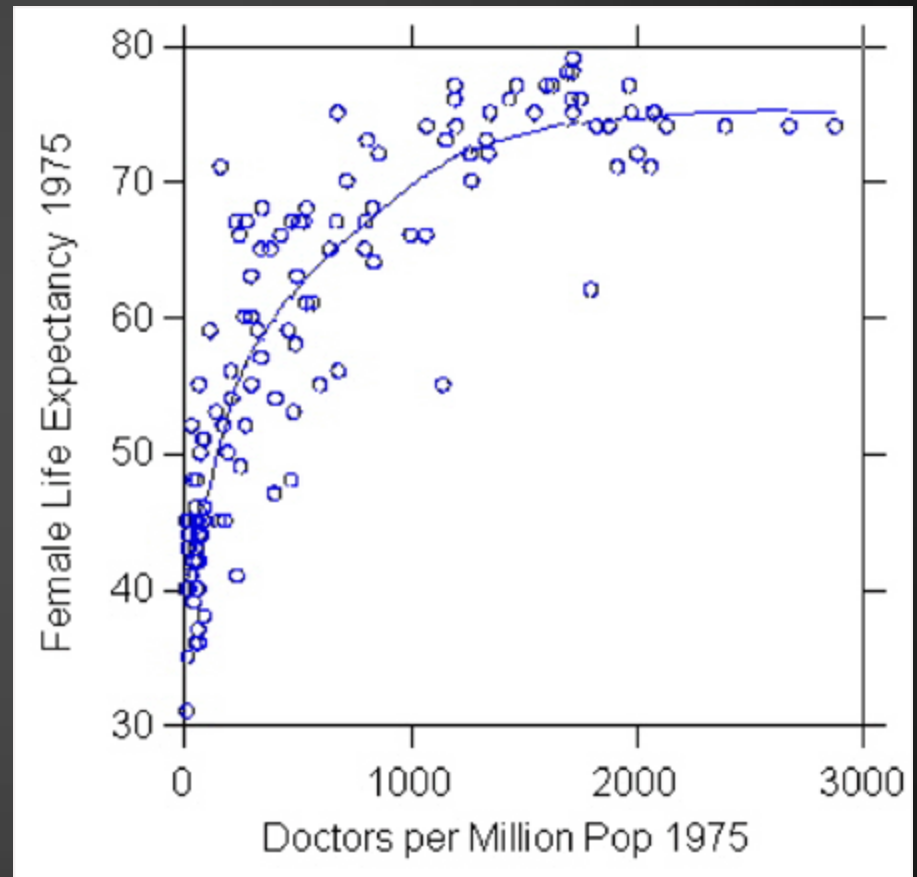
`bjn1c13@ecs.soton.ac.uk`

The linearity assumption

- Regression models are both powerful and useful.
- But they assume that a predictor variable and an outcome variable are related *linearly*.
- This assumption can be wrong in a variety of ways.

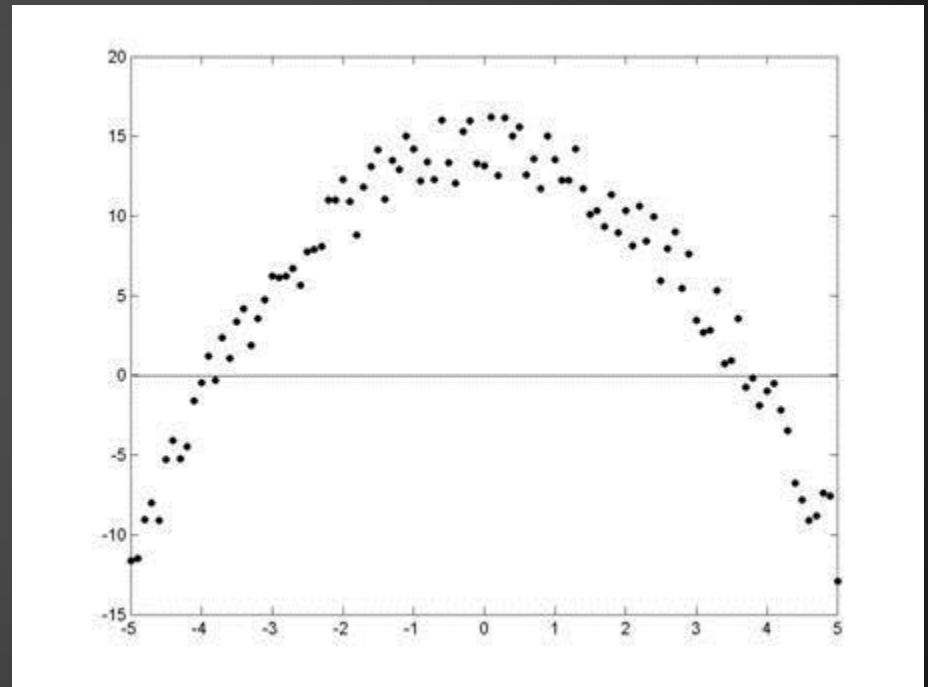
The linearity assumption

- Some real data showing a non-linear connection between life expectancy and doctors per million population.



The linearity assumption

- The Y and X variables here are clearly related, but the correlation coefficient is close to zero.
- Linear regression would miss the relationship.



The independence assumption

- Simple regression models also assume that if a predictor variable affects the outcome variable, it does so in a way that is *independent of all the other predictor variables*.
- The assumed linear relationship between Y and X_1 is supposed to hold no matter what the value of X_2 may be.

The independence assumption

- Suppose we're trying to predict happiness.
- For men, happiness increases with years of marriage. For women, happiness decreases with years of marriage.
- The relationship between happiness and time may be linear, but it would not be independent of sex.

Can regression models deal with these problems?

- Fortunately they can.
- We deal with non-linearity by transforming the predictor variables, or by fitting a polynomial relationship instead of a straight line.
- We deal with non-independence of predictors by including *interaction terms* in our models.

Dealing with non-linearity

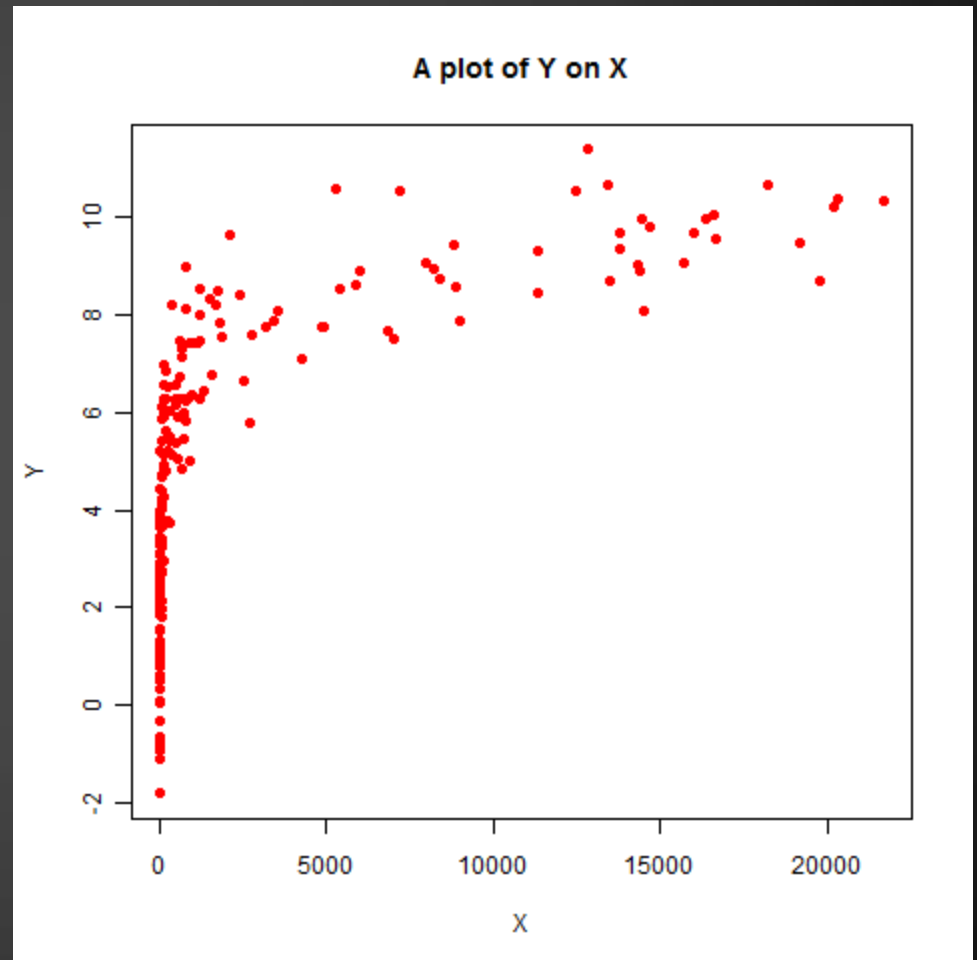
- Transformation is the simplest method for dealing with this problem.
- We work not with the raw values of X , but with some arbitrary function that transforms the X values such that the relationship between Y and $f(X)$ is now (closer to) linear.
- Like finding that you were measuring the wrong thing, e.g., surface area turns out to be a better predictor than height.

Dealing with non-linearity

- How do we know what transformation to use?
- Sometimes there will be a theoretical motivation.
- Certain relationships in a scatterplot start to look familiar and suggest a fix.
- We can experiment with different functions and assess them using R^2 or AIC.

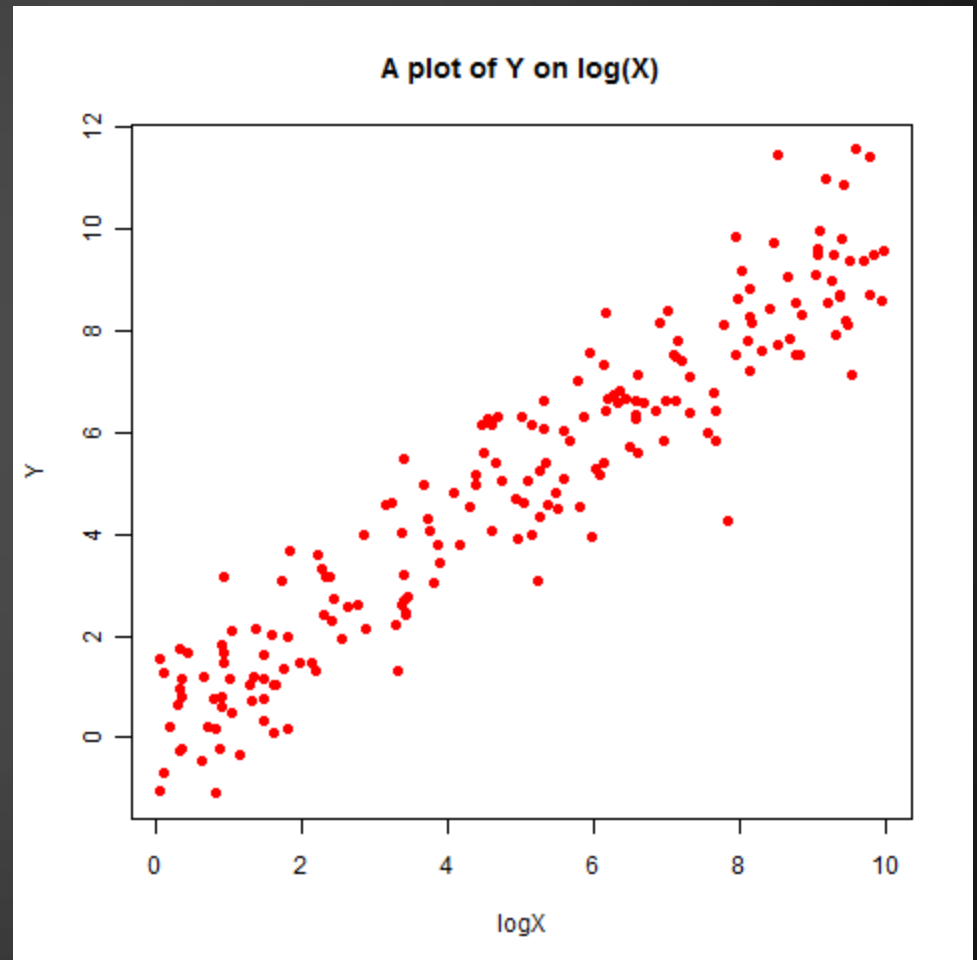
An example: Y on X

- Values of Y increase very rapidly as X increases, and then level off.
- Suggests a log transformation of X.



Plotting Y against $\log(X)$

- More linear relationship evident.
- Correlation coefficient has improved from 0.65 to 0.95.



Transformations

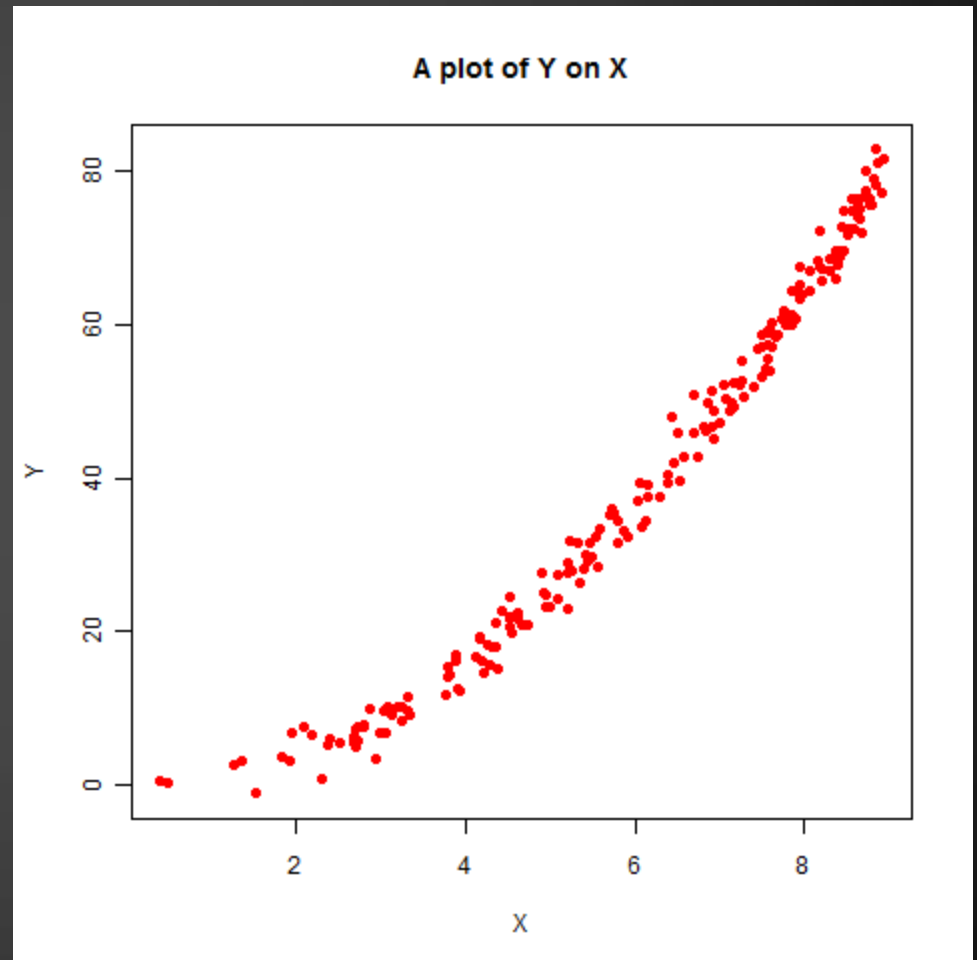
- If the fit to the transformed version of the variable *looks* more linear, we will get a better fit using regression.
- We can experiment with transforming the dependent variable as well.
- When making predictions with our model, we need to remember to "de-transform" the variable.

Assessing a transformation

- If we are doing linear regression on the transformed variables, we can ask simply whether the R^2 value indicates that we are explaining more of the variance now.
- We can also use the incredibly convenient AIC measure to ask which of two competing models of the variation in the outcome measure is preferable.

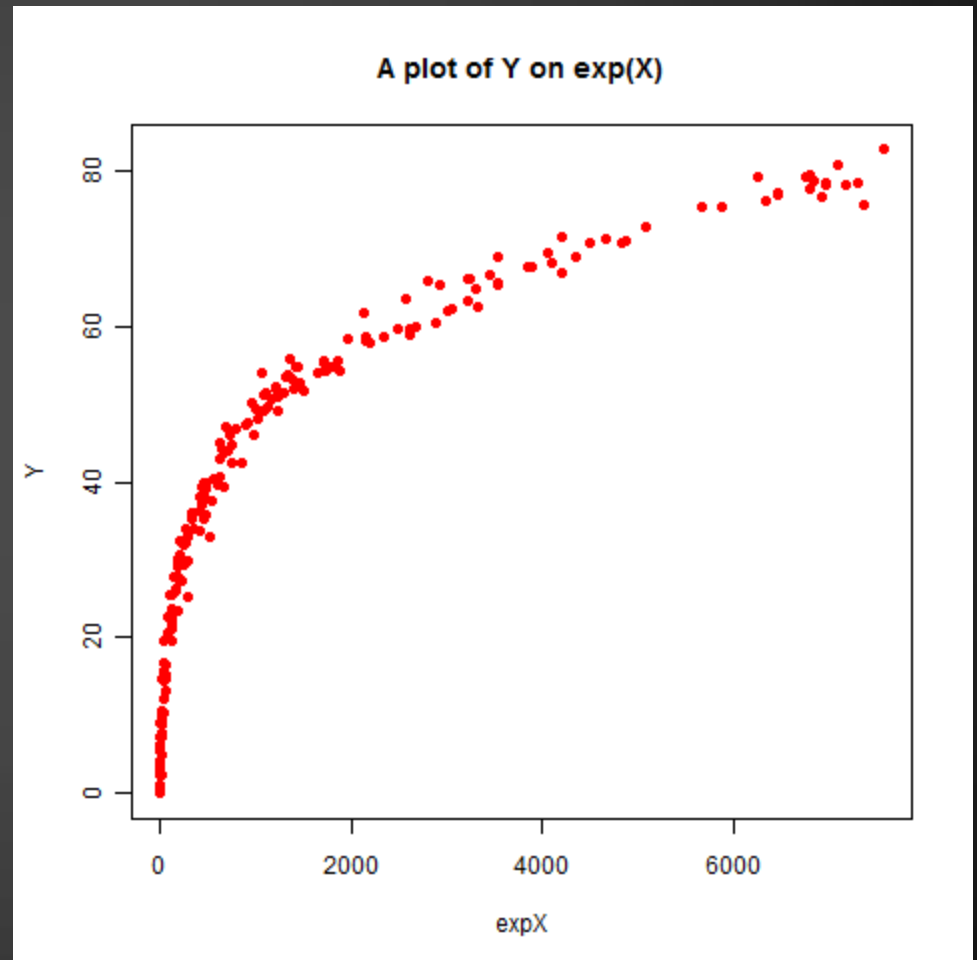
Another example: Y on X

- In this case, Y appears to be increasing at a greater rate as X increases.
- Could try $\exp(X)$?



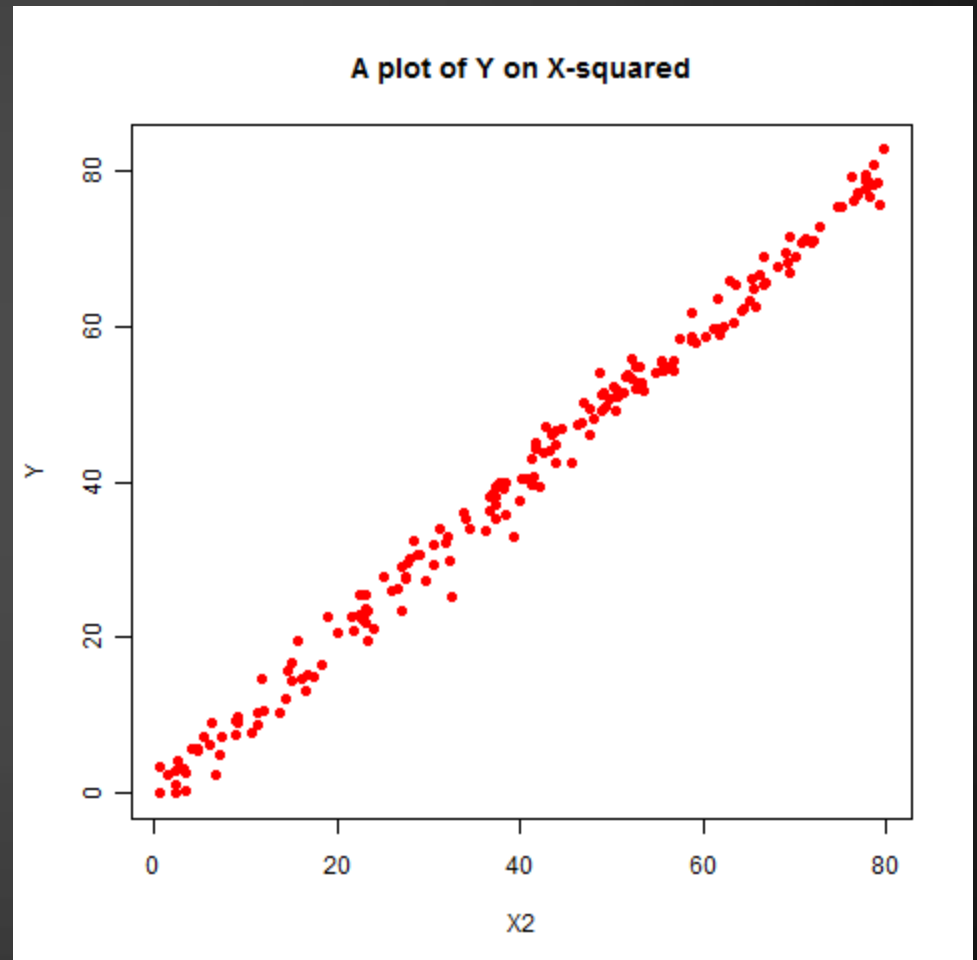
Transforming with $\exp(X)$

- The $\exp(X)$ transform seems too extreme; this doesn't look linear either.



Transforming with X^2

- The X^2 transformation looks much more linear.



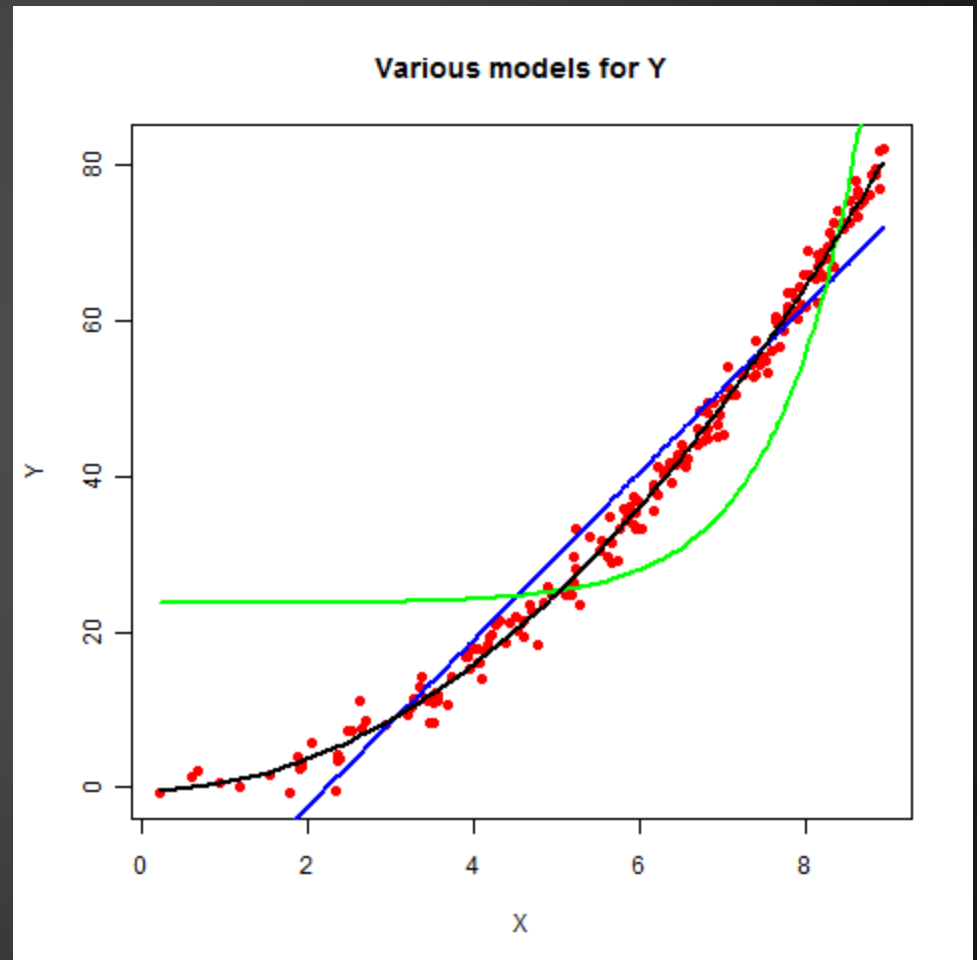
Assessing a transformation

- We can fit regression lines to all three cases.
- AIC analysis confirms that the X^2 transformation is the best.

	df	AIC
unchanged	3	1227.2688
expTrans	3	1555.0627
squareTrans	3	846.1275

Visual assessment

- The three regression models plotted together.
- Blue is no transform, green is $\exp(X)$, and black is X^2 .



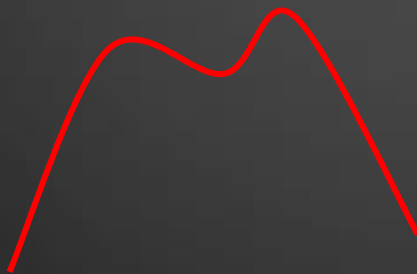
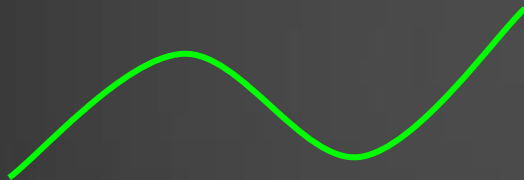
Polynomial fitting

- A different but related way of dealing with non-linear relationships is to fit a polynomial model rather than a simpler linear one.
- The idea of regression as fitting a straight line still applies. We are just turning one predictor variable into two or more predictors.
- Instead of predicting Y on X , we predict Y on X , and X^2 (and possibly X^3 , X^4 , etc.)

Polynomial fitting

- This allows us to deal with obvious non-linearity but without having to specify in advance precisely what the appropriate transformation would be.
- The degree of the polynomial can be estimated based on the shape of the relationship in the scatterplot.

Degrees of polynomials



- A parabola...
- a cubic...
- and a quartic relationship between Y and X.

How to do it in R

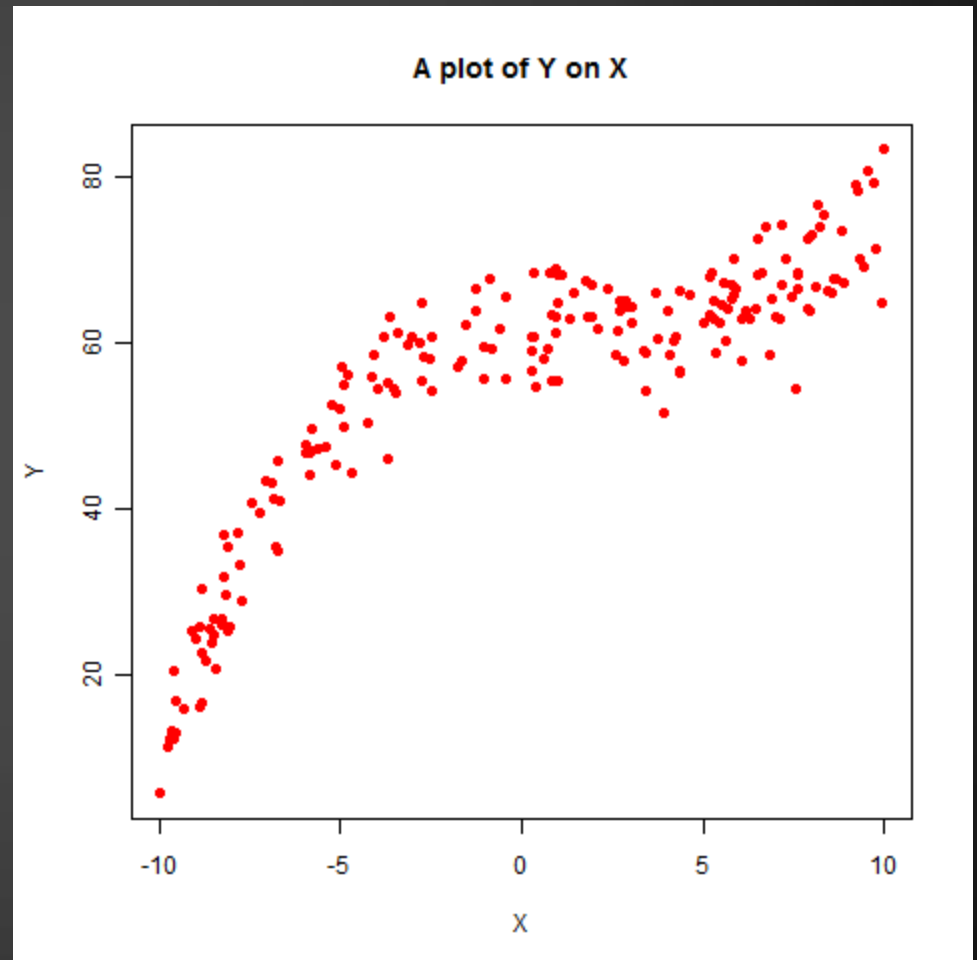
- There are two ways to add a polynomial term to the model.
- You can use
$$\text{model} = \text{lm}(y \sim x + I(x^2))$$
- and this will give you the expected coefficients.
- Recommended.

How to do it in R

- You can also use orthogonal polynomials, which are more efficient for fitting, with
`model = lm (y ~ poly(x,2) )`
- But the coefficients aren't as transparent.
- Use this method if you want to return predictions but don't need to directly interpret the polynomial coefficients.

A polynomial fitting example

- There's clearly a strong relationship, but it's not linear.
- No single transform looks like it will obviously fix things.



Polynomial fitting example

- We can fit a series of models to compare.

```
m0 = lm ( Y ~ 1 )
```

```
m1 = lm ( Y ~ X )
```

```
m2 = lm ( Y ~ X + I (X^2) )
```

```
m3 = lm ( Y ~ X + I (X^2) + I (X^3) )
```

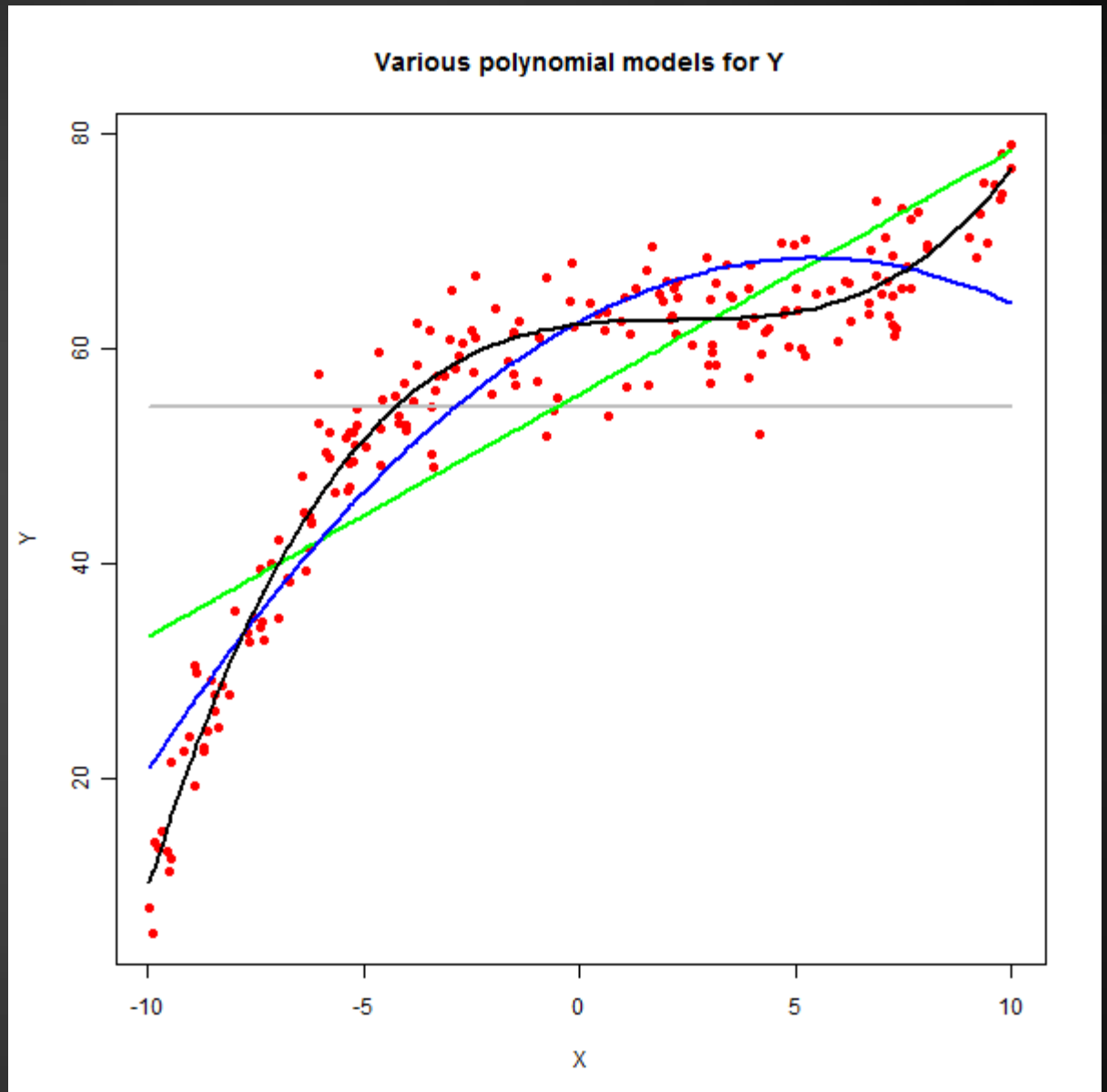
```
m4 = lm ( Y ~ X + I (X^2) + I (X^3) + I (X^4) )
```

```
AIC (m0, m1, m2, m3, m4)
```

	df	AIC
m0	2	1691.918
m1	3	1435.907
m2	4	1278.603
m3	5	1110.864
m4	6	1111.103

← **m3, the cubic model, has the lowest AIC score**

- Grey is the null model (the mean).
- Green is the linear model.
- Blue is the order-2 polynomial.
- Black is the order-3 polynomial, with the lowest AIC value.



Interaction terms

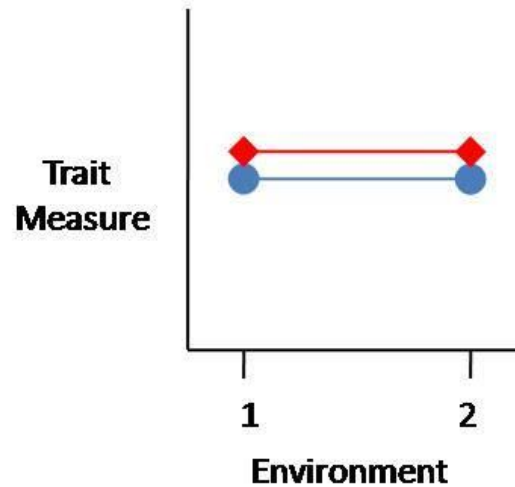
- Independence Assumption: Violated
- If two predictor variables affect the outcome variable in a way that is non-additive, we need to include an *interaction term* in the model to capture this effect.
- This is the same as epistasis in NK landscapes: e.g., I can't say whether anchovies will improve the pizza without knowing whether it has prawns on it.

Interaction terms

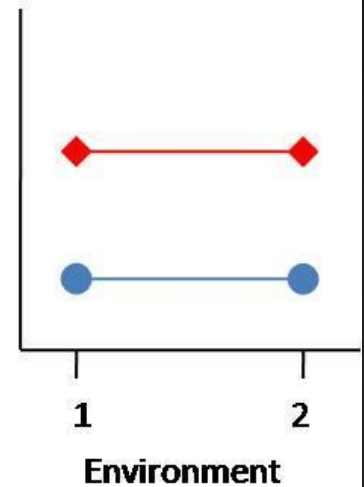
- Consider an example where we have two species, red and blue, and we measure their fitness in two environments, 1 and 2.
- A species effect means that the red species does better or worse than the blue one.
- An environment effect means that average fitness in environment 1 is different from average fitness in environment 2.

- What about interaction effects?
- Red species vs Blue Species
- Environment 1 vs Environment 2

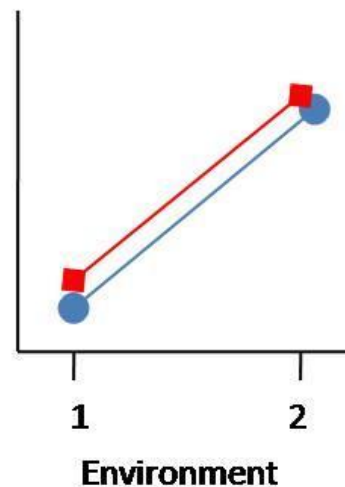
(A) No significant effects



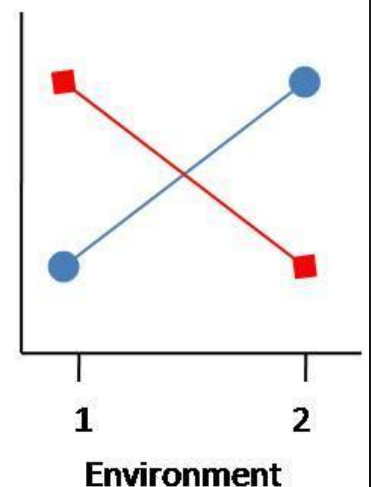
(B) Species effect



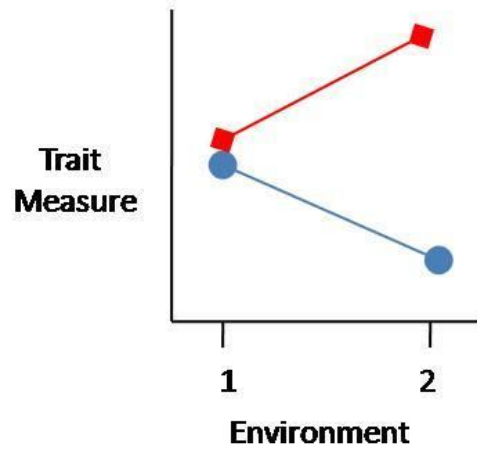
(C) Environment effect



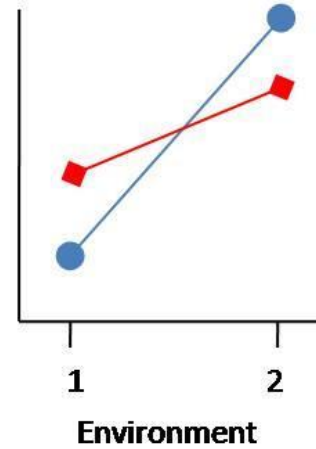
(D) Interaction effect



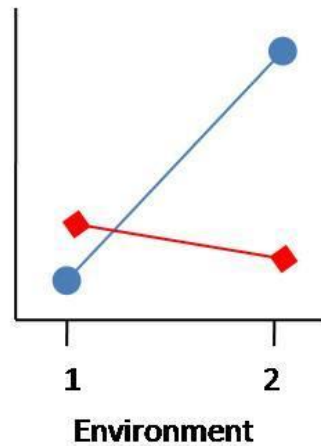
(E) Interaction and Species effect



(F) Interaction and Environment effect



(G) Interaction, Species, and Environment effect



How do I do it in R?

- `model = lm( Y ~ X1 * X2 )`

will regress Y on X1, X2, and the X1-by-X2 interaction term.

- Equivalent to:

```
model = lm(Y ~ X1 + X2 + X1:X2)
```

- To specify more complex interactions see the [R documentation for "formula"](#).

How do I know if an interaction term is worth including?

- If you're dealing with large numbers of predictors, it's not practical to try to investigate all possible interaction terms.
- Due to the combinatorics of interaction, your model will rapidly have more predictor variables than your data has cases.

How do I know if an interaction term is worth including?

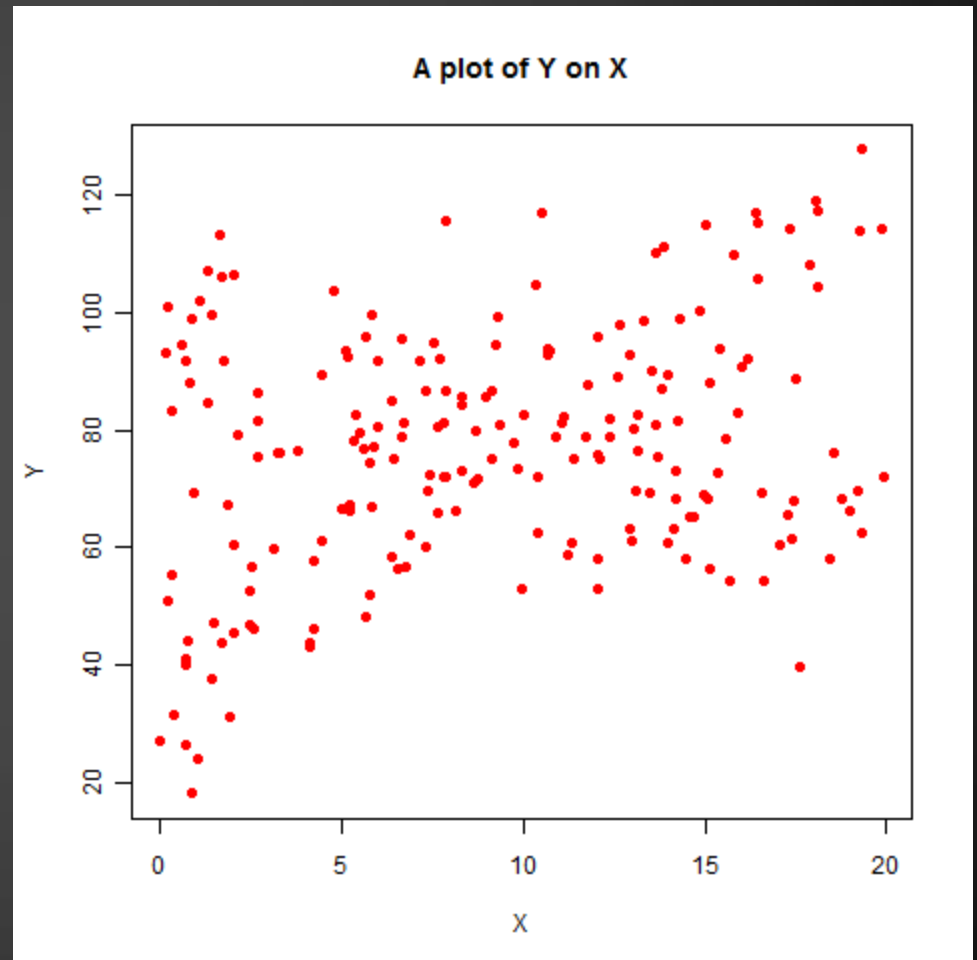
- Therefore you want to keep to interaction terms that you either have theoretical reasons to expect, or which are strongly suggested by your use of descriptive statistics to look at the data.
- Once an interaction term is in the model, it's up for potential elimination in the usual way as you try to reduce the model.

How do I know if an interaction term is worth including?

- AIC analysis may indicate that an interaction term does not add significantly to our ability to predict the dependent variable.
- A complication: if an interaction term does stay in the model, the component terms have to stay too in order to make the model interpretable.

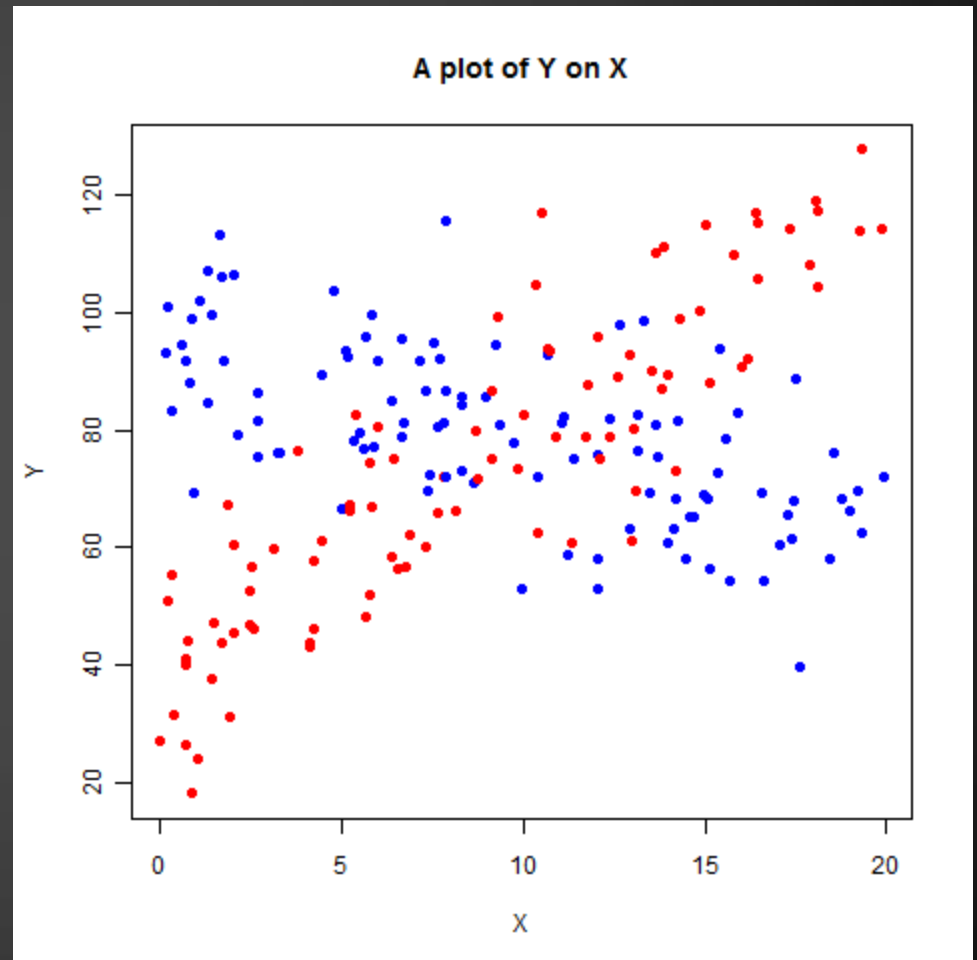
An example

- The outcome measure Y is related to the predictor X , but is also dependent on the categorical variable "Group".



An example

- If we display group A in red and group B in blue the relationship becomes clearer.



Fitting a regression model

```
lm(formula = Y ~ X * Group)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	33.4776	2.5465	13.15	<2e-16	***
X	4.3830	0.2144	20.44	<2e-16	***
GroupB	67.2934	3.2639	20.62	<2e-16	***
X:GroupB	-6.3115	0.2756	-22.91	<2e-16	***

Sig. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.99 on 196 degrees of freedom

Multiple R-squared: 0.7348, Adjusted R-squared: 0.7308

F-statistic: 181 on 3 and 196 DF, p-value: < 2.2e-16

Can the model be reduced?

```
> drop1(fullModel)
```

Single term deletions

Model:

Y ~ X * Group

	Df	Sum of Sq	RSS	AIC
<none>			23674	962.76
X:Group	1	63366	87039	1221.16

- AIC analysis confirms that the interaction term can't be dropped. This is our preferred model (must keep components).

Interpreting the model

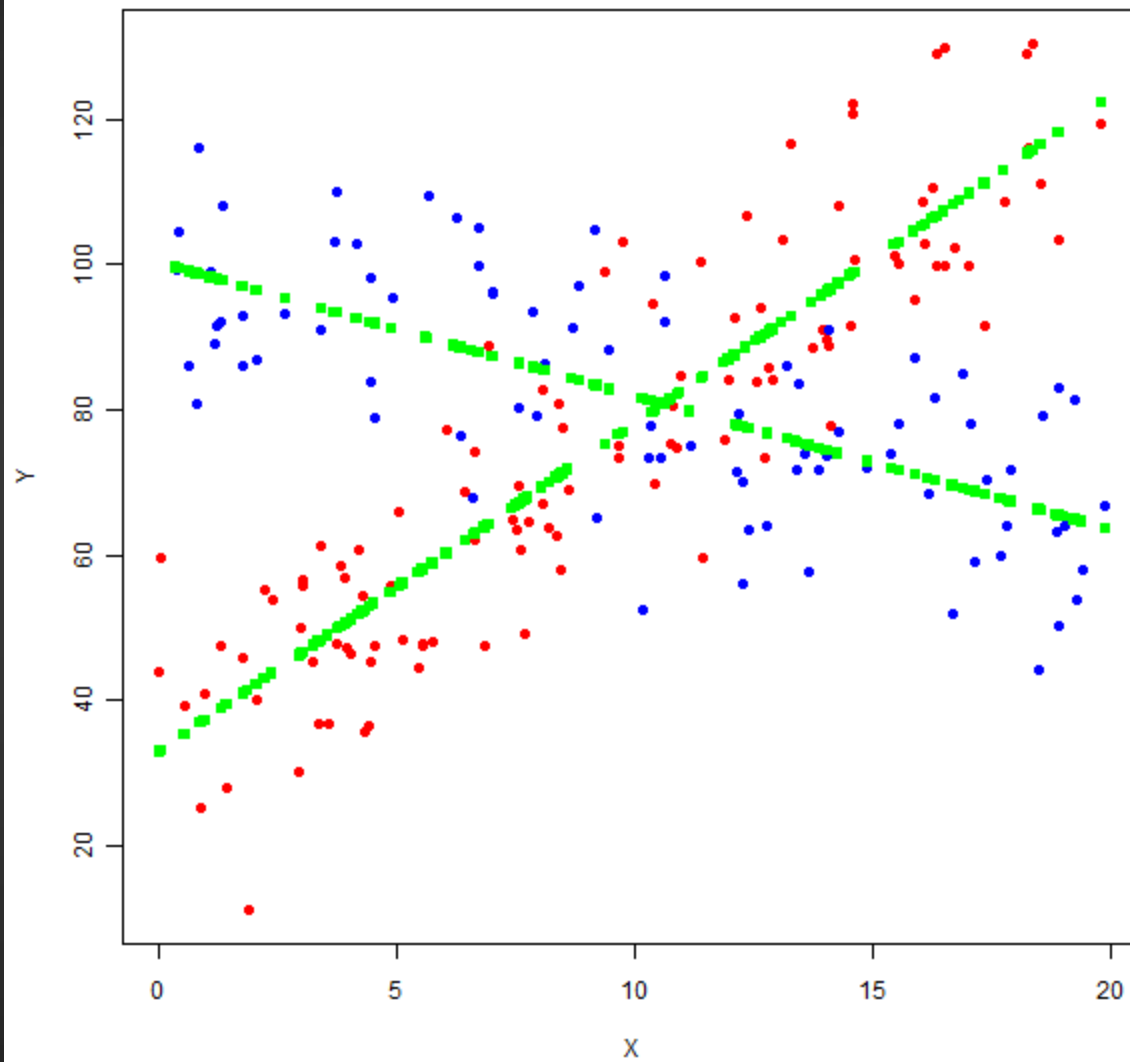
- If you're in group A, your prediction line is

$$Y = 33.48 + 4.38X.$$

- If you're in group B, we change the intercept by adding in the group B effect, and we change the slope by adding the X-by-Group interaction term to get:

$$Y = 100.77 - 1.93X.$$

A plot of Y on X, with fitted interaction model



Additional material

- All the graphs and fictional data sets in this lecture were generated in R.
- Here is the [R script](#).
- Run it
`with:source("transformScript.txt",echo=TRUE)`