Mixing variable types

Analysis of Covariance (ANCOVA)

Mixing variable types

- Before you learned about the GLM, ANOVA and regression seemed to be distinct approaches
  - ANOVA for grouped data
  - Regression for numeric predictors
- Now that you know ANOVA and regression are the same thing, why not mix variable types?
- What happens when you include categorical variable and a numeric variable together in a GLM?

Generality of GLM

Table 6.1 Comparing word equations with traditional tests

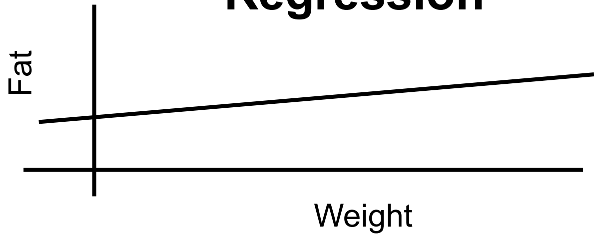
<table>
<tr><th>Example</th><th>Traditional test</th><th>GLM word equation</th></tr>
<tr><td>Comparing the yield between two fertilisers</td><td>Two sample t-test</td><td>YIELD = FERTIL</td></tr>
<tr><td>Comparing the yield between three or more fertilisers</td><td>One way analysis of variance</td><td>YIELD = FERTIL</td></tr>
<tr><td>Comparing the yield between fertilisers in a blocked experiment</td><td>One way blocked analysis of variance</td><td>YIELD = BLOCK + FERTIL</td></tr>
<tr><td>Investigating the relationship between fat content and weight</td><td>Regression</td><td>FAT = WEIGHT</td></tr>
<tr><td>Investigating the relationship between fat content and sex, controlling for weight differences</td><td>Analysis of covariance</td><td>FAT = WEIGHT + SEX</td></tr>
<tr><td>Investigating which factors may influence the likelihood of spotting whales on a boat trip</td><td>Multiple regression</td><td>LGWHALES = CLOUD + RAIN + VIS</td></tr>
<tr><td>Investigating the factors which affect the number of blooms on prize roses</td><td>Two way analysis of variance</td><td>SQBLOOMS = SHADE | WATER</td></tr>
</table>

Graphs and equations

ANOVA

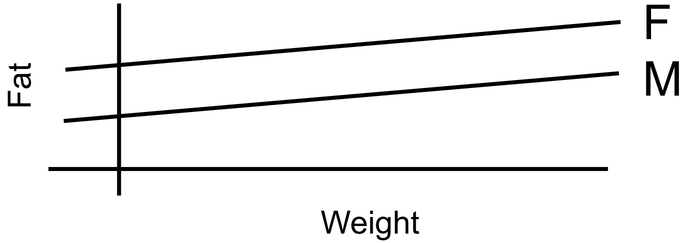
$$FAT = \alpha + \beta_1 Male$$

Regression



$$FAT = \alpha + \beta_2 WEIGHT$$

ANCOVA



$$FAT = \alpha + \beta_1 Male + \beta_2 WEIGHT$$

Regression with parallel lines, one for each level of the categorical variable

Three reasons to do ANCOVA

- Experimentally interesting question is the regression line, but we need to account for a categorical variable (block)
  - Example: fat vs. weight, accounting for sex
- Experimentally interesting question is the comparison of means, but we need to reduce the noise to increase effect sizes
  - Example: leprosy bacteria, accounting for initial bacterial density
- Experimentally interesting question is the comparison of means, but we need to adjust the means to account for the effect of the covariate
  - Example: comparing wing chords between sexes of birds, adjusting for sex differences in mass

Blocking on a categorical variable

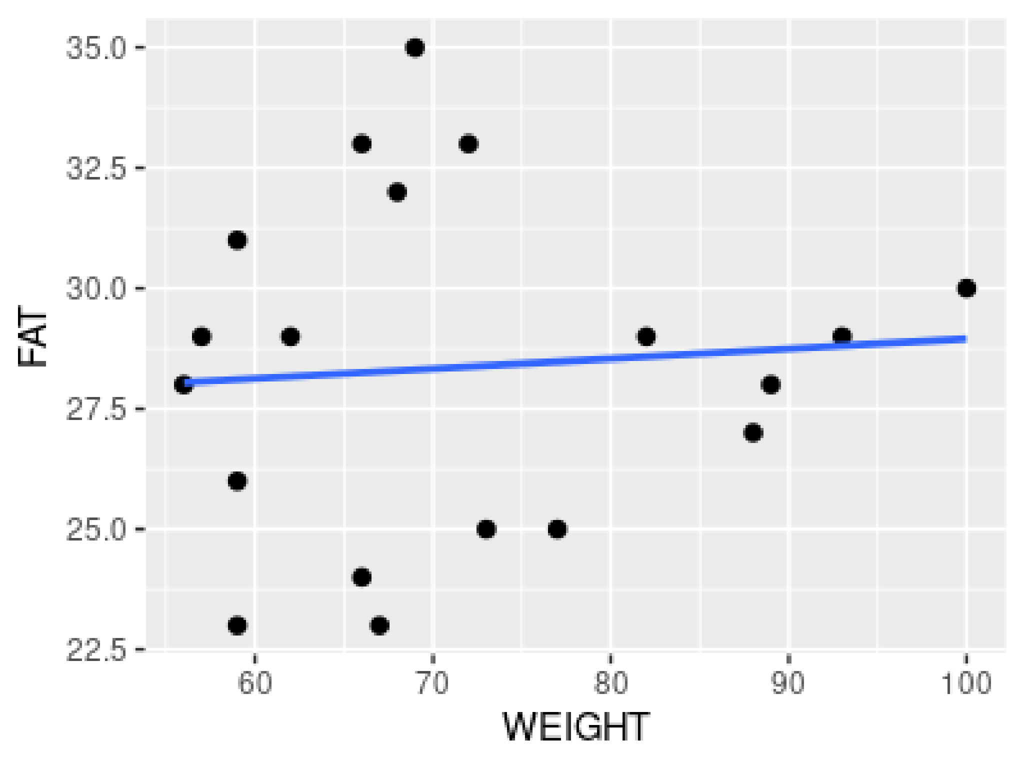
- The regression question is the interesting one, but there are groups in the data
  - Sexes
  - Age classes
  - Location samples are housed (greenhouse, chamber)
- For the regression to properly represent the numeric relationship between predictor and response, the grouping needs to be accounted for
- Example: fat mass/body mass relationship

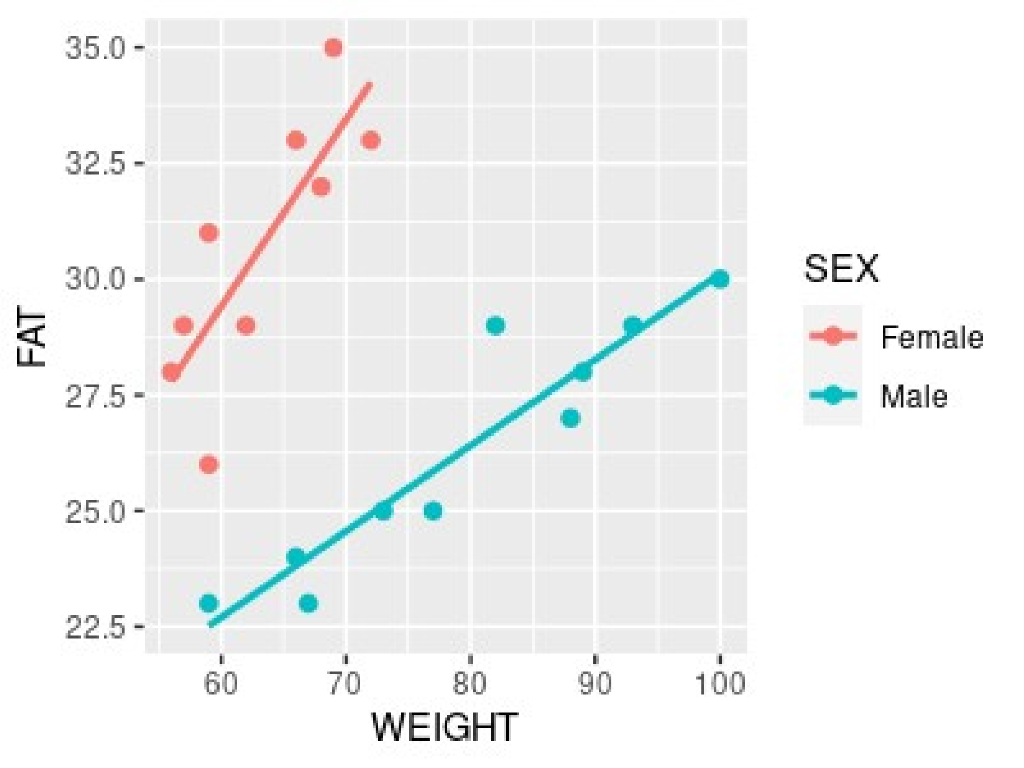
Fat mass vs. body mass relationship for a sample of people

```
Analysis of Variance Table
Response: FAT
          Df  Sum Sq Mean Sq F value Pr(>F)
WEIGHT     1   1.328  1.3282   0.104  0.751
Residuals 17 217.093 12.7702
```

Not a significant relationship between fat and weight

What's wrong with this picture?

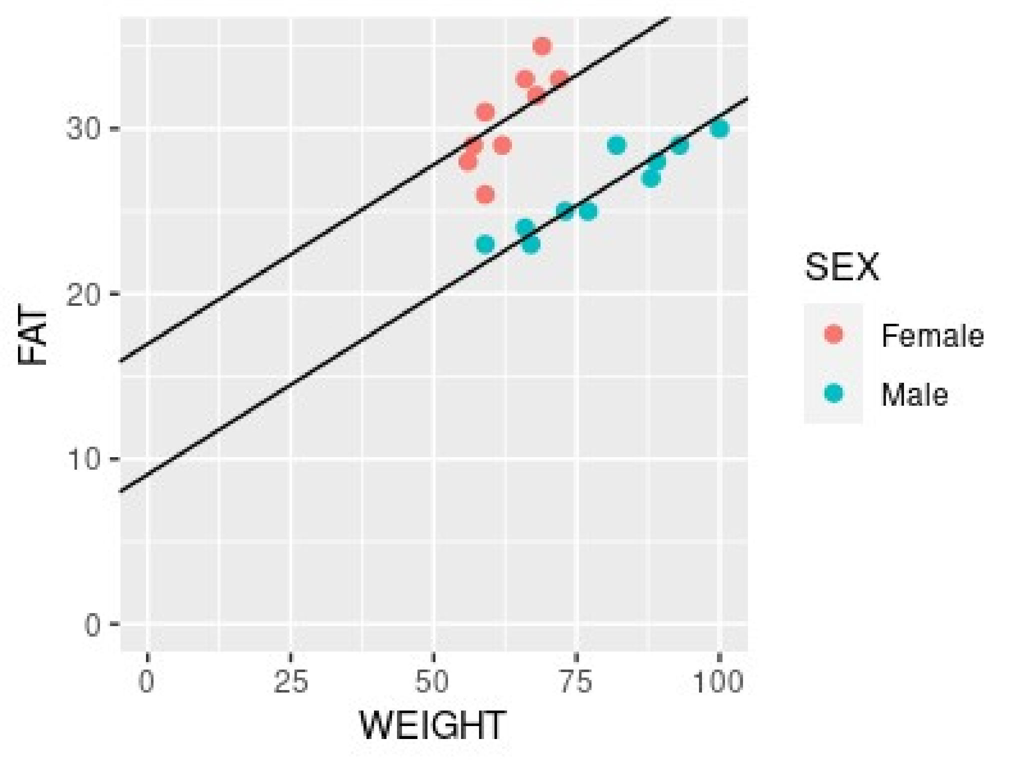




The sexes seem to have a similar fat vs. weight relationship, but women have higher fat percentages at a given weight

Model is FAT ~ WEIGHT + SEX

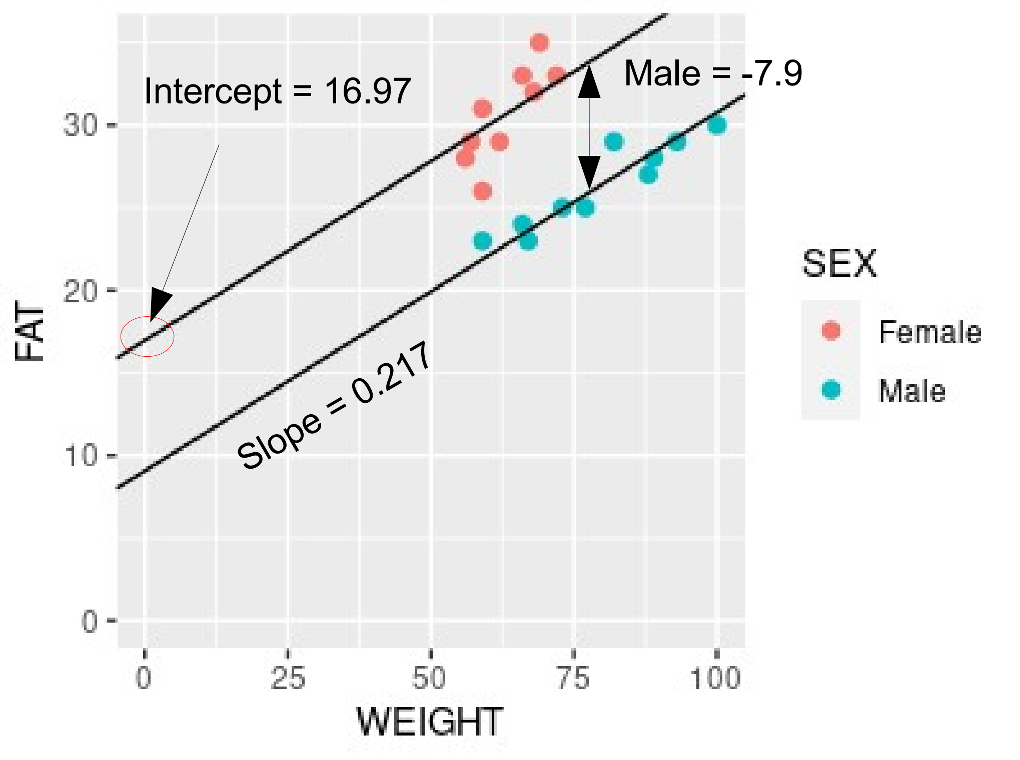
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.96228    2.41021   7.038 2.80e-06 ***
WEIGHT       0.21715    0.03724   5.831 2.56e-05 ***
SEXMale     -7.90375    0.95337  -8.290 3.48e-07 ***
```



- What does the intercept mean?
- What is the male coefficient?
- Is the slope the same or different for males and females?

$$FAT = 16.96 + 0.217 \times WEIGHT - 7.904 \times Male$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.96228    2.41021   7.038 2.80e-06 ***
WEIGHT       0.21715    0.03724   5.831 2.56e-05 ***
SEXMale     -7.90375    0.95337  -8.290 3.48e-07 ***
```



$$FAT = 16.96 + 0.217 \times WEIGHT - 7.90 \times Male$$

$$FAT_{female} = 16.96 + 0.217 \times WEIGHT - 7.90 \times 0$$

$$FAT_{female} = 16.96 + 0.217 \times WEIGHT$$

$$FAT_{male} = 16.96 + 0.217 \times WEIGHT - 7.90 \times 1$$

$$FAT_{male} = 9.06 + 0.217 \times WEIGHT$$

Slope is the same for both sexes

Intercepts are different

The SEXMale coefficient is the vertical distance between the lines at a given weight

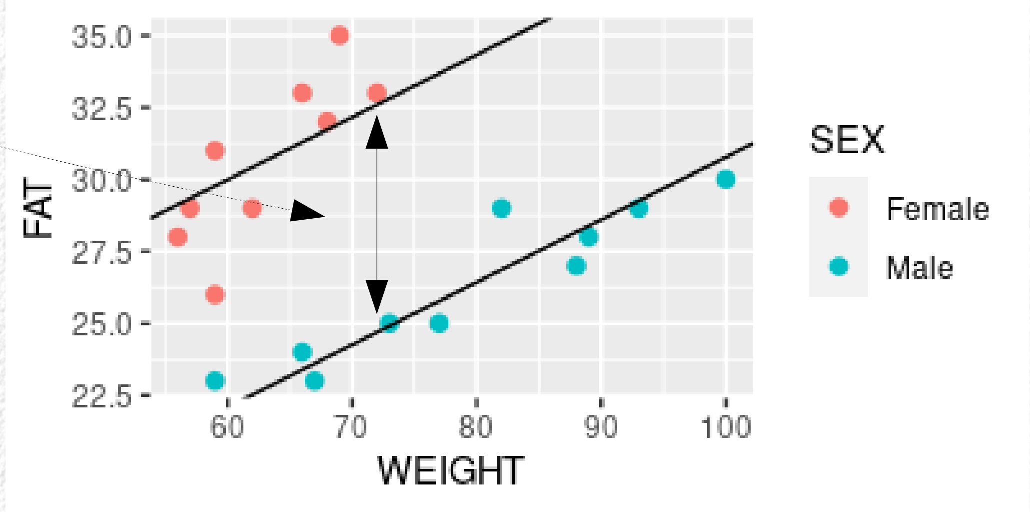
ANOVA table

```
Anova Table (Type II tests)

Response: FAT
           Sum Sq Df F value    Pr(>F)
WEIGHT     87.105  1  33.996 2.556e-05 ***
SEX       176.098  1  68.729 3.482e-07 ***
Residuals  40.995 16
```

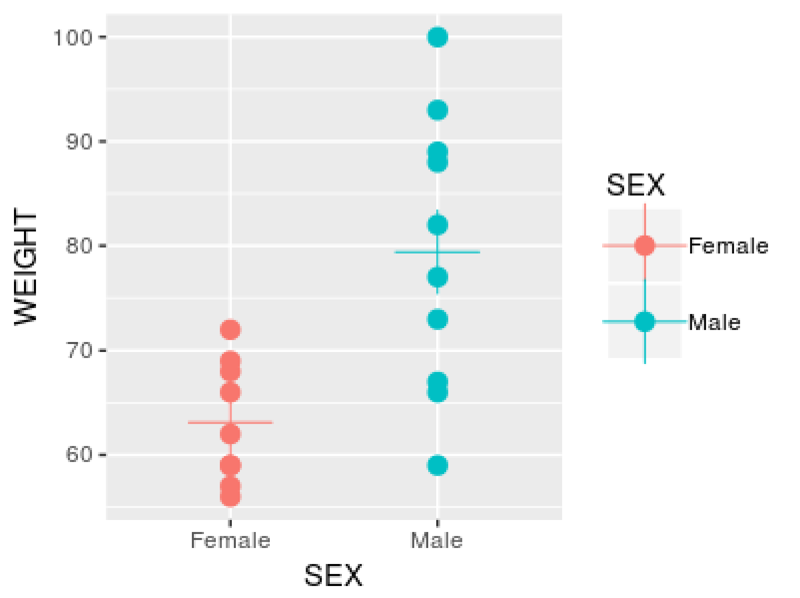
Test of fat vs. weight relationship, allowing for sex differences in intercept

Accounting for the different groupings in the data



Lack of independence of predictors

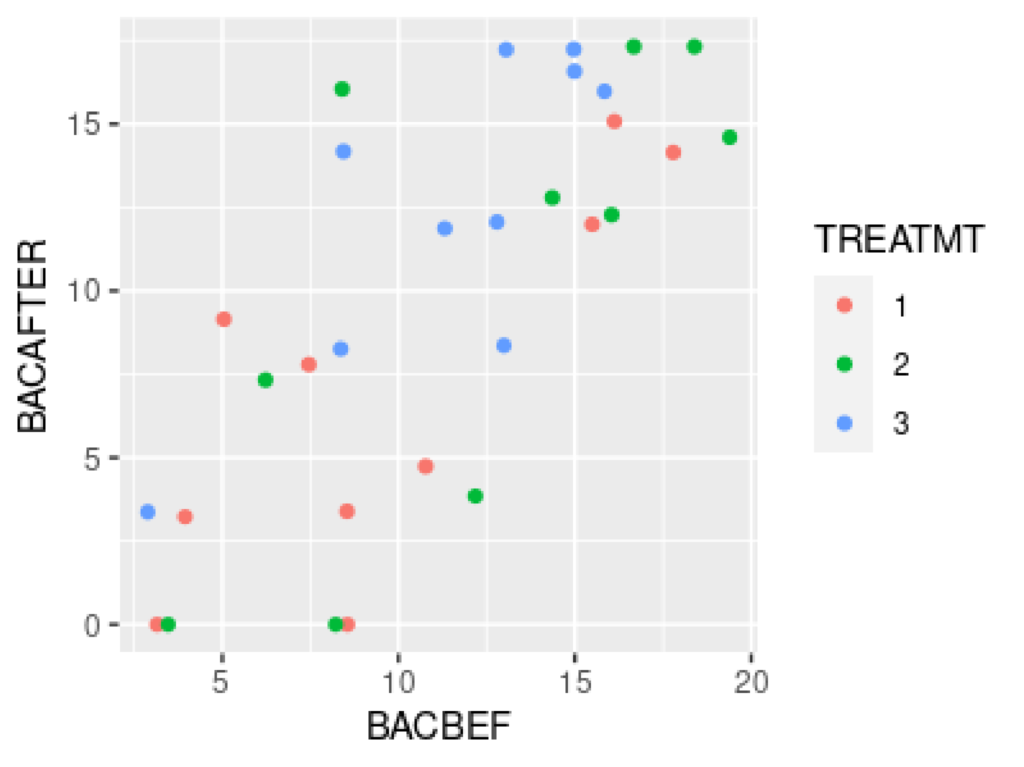
- Sex and weight are not orthogonal = not independent
  - The numeric variable is different on average between the categories → sexes differ in mean weight
- r^2 = 0.40, equivalent to correlation of 0.63
- This means that:
  - Type II and Type I SS will be different
  - Order of entry of sex and weight will matter in Type I
- Solution: either enter the nuisance first, or use Type II



Classic ANCOVA

- Question of interest is comparison of group means
- But, there is a numeric variable that is a nuisance = a covariate
- Include the covariate in the model to:
  - Account for random variation caused by the covariate → statistical elimination, increase effect size of the treatment variable
  - Make comparisons between the covariate-adjusted means – equivalent to setting the groups to the same value of the covariate

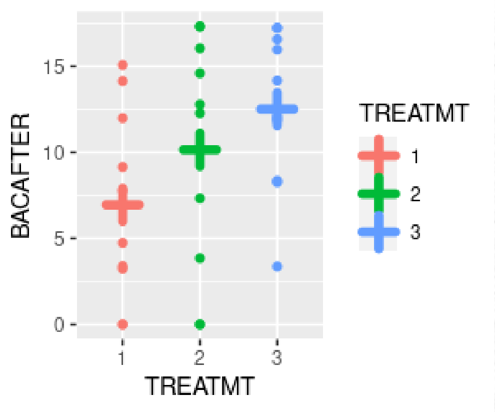
Cleaning up noisy data – leprosy experiment



- Leprosy caused by bacteria
- Testing effects of three different treatments (levels 1, 2, 3)
- Amount of bacteria after treatment is partly due to initial bacteria levels, before treatment
- We're asking: is there an effect of treatment, once initial bacteria levels are accounted for?

Statistically eliminating BACBEF from the test of treatment on BACAFTER

No adjustment for BACBEF

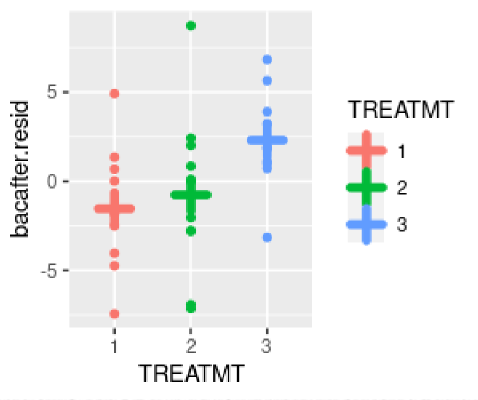


Analysis of Variance Table

Response: BACAFTER

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TREATMT	2	155.81	77.904	2.3502	0.1146
Residuals	27	894.99	33.148		

With BACBEF accounted for



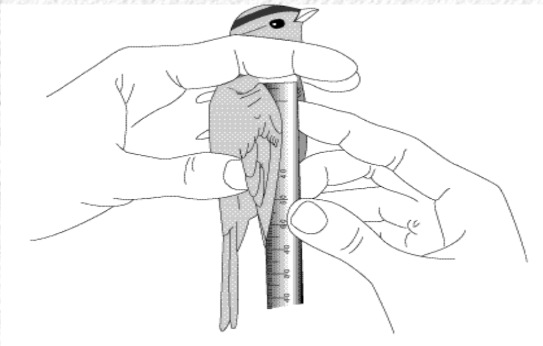
Analysis of Variance Table

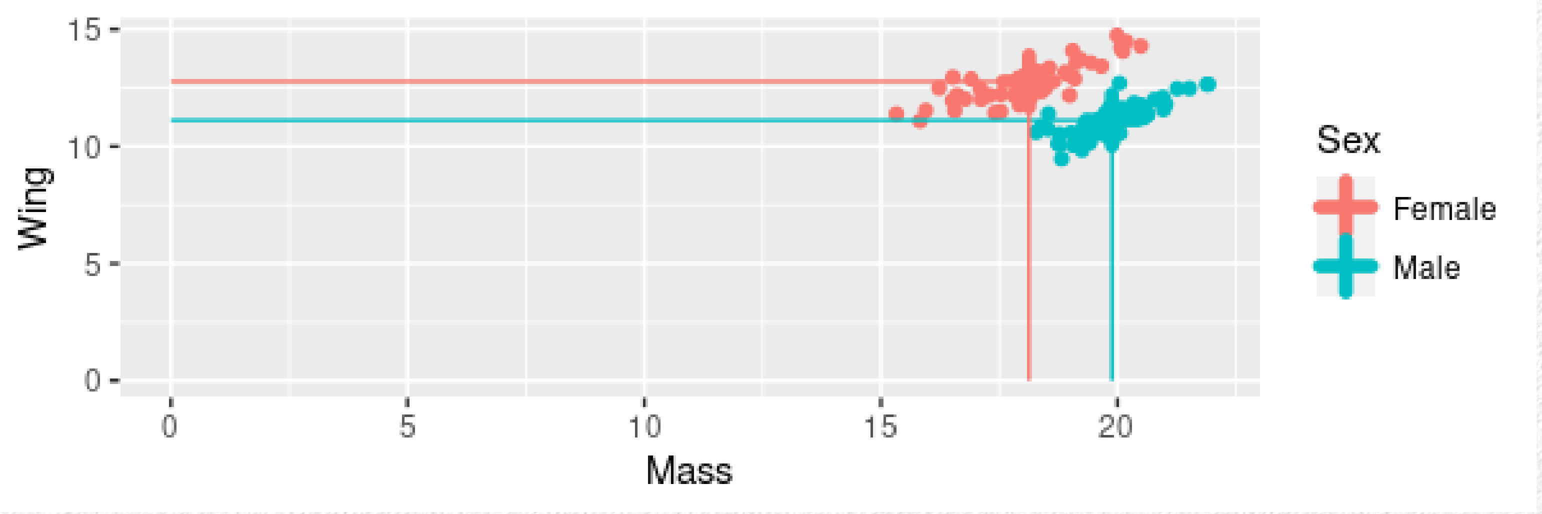
Response: BACAFTER

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BACBEF	1	587.50	587.50	40.1988	1.03e-06
TREATMT	2	83.31	41.66	2.8502	0.076
Residuals	26	379.99	14.61		

Making covariate-adjusted comparisons

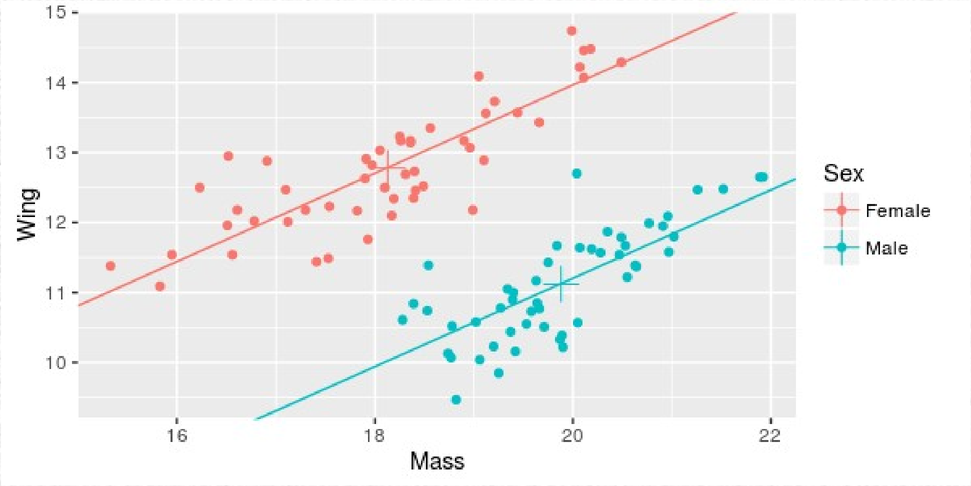
- Good examples come from study of shapes and sizes or organisms = morphometrics
- In a species of bird we are studying, females have bigger wing chords
- But, the sexes are also different in mass – males are heavier
- Is the wing difference really just a size difference (i.e. a difference in mass)?
- If the sexes were the same mass, what would the difference in wing chord be?



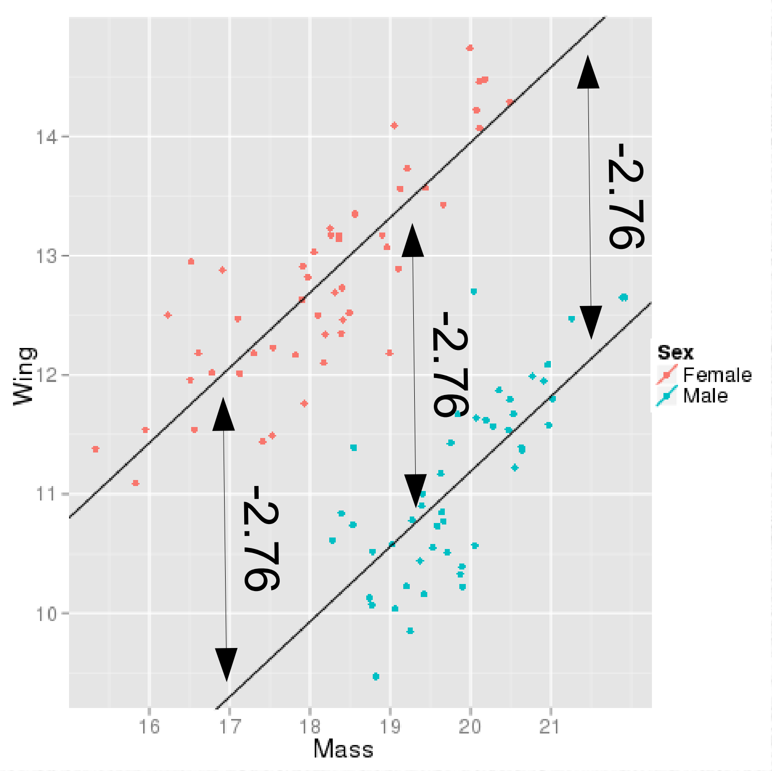


The goal of the analysis

- We will fit two parallel lines through the data
  - Same slope
  - Different intercepts
- We will use the regressions to find the wing chord for each sex at a common mass (called the **least squares means**)



Fitted model



$R^2 = 0.82$

Males: Wing =

1.34 – 2.76 (1) + 0.63 Mass

-1.42 + 0.63 Mass

Females: Wing =

1.34 – 2.76 (0) + 0.63 Mass

1.34 + 0.63 Mass

Slope is the same for both sexes

The coefficient for Male is the difference between the lines at any point along the x-axis

Fitted model

- The test of Sex (dummy-coded) is based on mass-adjusted means
  - Vertical difference between the parallel lines
  - The SexMale coefficient
- Intercept is the mean wing chord for females that weigh 0 g

```
Call:
lm(formula = Wing ~ Mass + Sex, data = birds)
Residuals:
     Min       1Q   Median       3Q      Max
-1.14063 -0.31533  0.04671  0.30280  1.47952
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.34656    0.85452   1.576    0.118
Mass         0.63055    0.04697  13.423   <2e-16 ***
SexMale     -2.76222    0.12970 -21.297   <2e-16 ***
---
Residual standard error: 0.5011 on 97 degrees of freedom
Multiple R-squared:  0.8238,    Adjusted R-squared:
0.8202
F-statistic: 226.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

ANOVA tables

- Note that when Mass is entered first it has very low SS
- Type I SS assigns confounded variation to the first variable entered
  - The examples we've seen assign more SS when a variable is entered first
  - Here Mass gets a higher SS if it's entered second
- What happened here?

Type I SS

Analysis of Variance Table

Response: Wing

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mass	1	0.004	0.004	0.0155	0.9013
Sex	1	113.879	113.879	453.5571	<2e-16
Residuals	97	24.355	0.251		

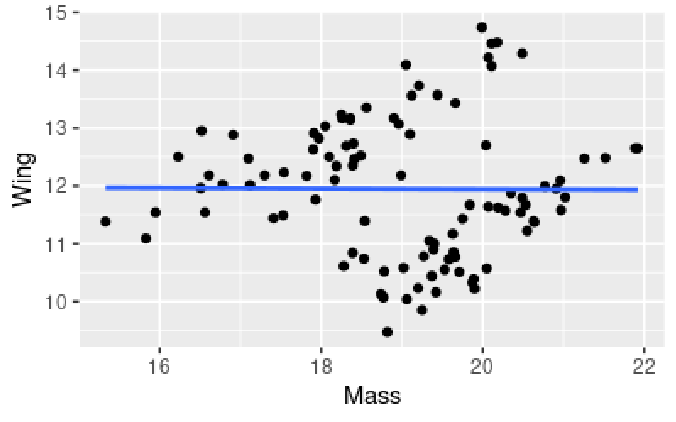
Anova Table (Type II tests)

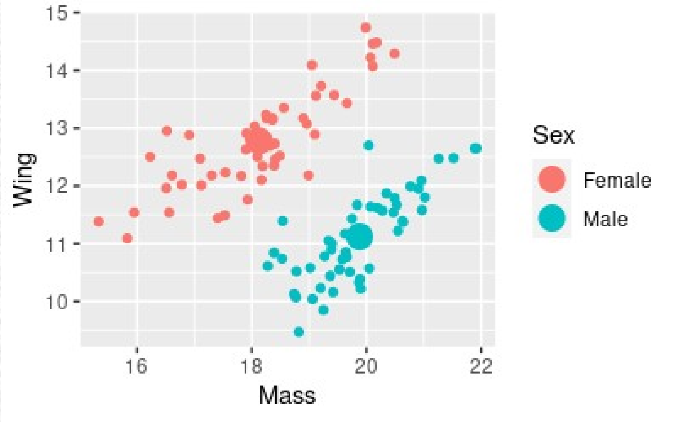
Response: Wing

	Sum Sq	Df	F value	Pr(>F)
Mass	45.241	1	180.19	< 2.2e-16
Sex	113.879	1	453.56	< 2.2e-16
Residuals	24.355	97		

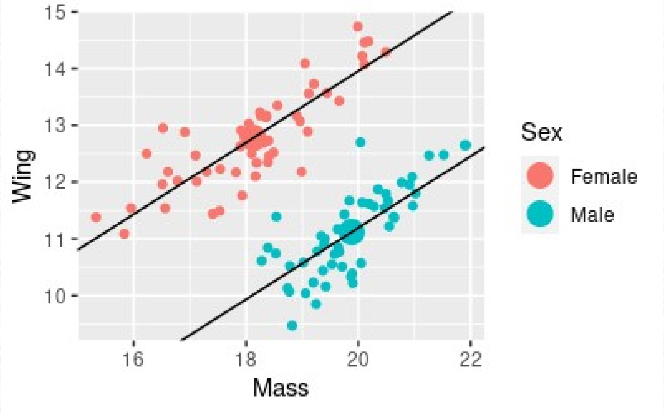
Type I SS – mass explains very little without sex

First →





Second →



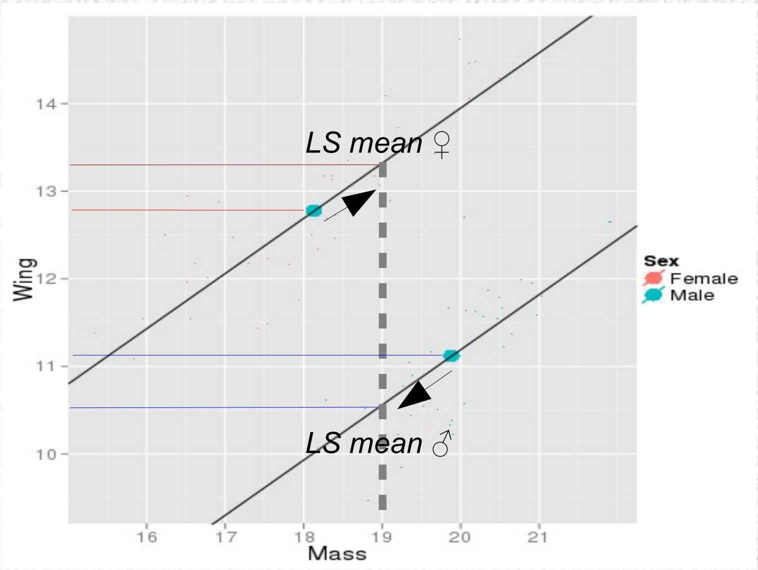
Estimating mass-adjusted mean wing chord

- It is **very important** to base your biological interpretations on what the statistical analysis is actually testing
- If we are testing mass-adjusted means, then mass-adjusted (least squares) means should be interpreted
- Obtained by predicting the mean of the response variable for each category at a selected value of the covariate
  - Usually done at the covariate mean – the location of minimum SE
  - The difference in LS means is the same at any mass, as long as the same mass is used for both sexes
- May be closer together than the actual means or further apart depending on the data

Least-squares means

Vertical distance between the lines is a mass-adjusted measure of difference in wing chord

Predicted values for each sex at the same mass gives "least squares means"



Mean mass = 19

Males: Mass adjusted mean wing chord

1.34 – 2.76 (1) + 0.63 Mass

-1.42 + 0.63 (19) = ***10.55***

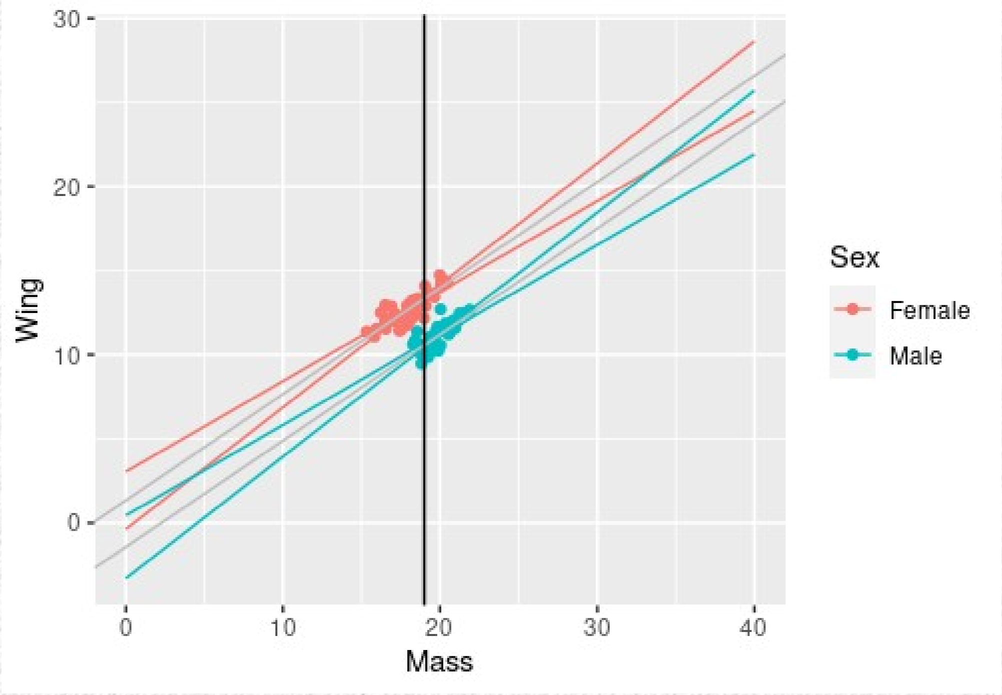
Females: Mass adjusted mean wing chord

1.34 – 2.76 (0) + 0.63 Mass

1.34 + 0.63 (19) = ***13.31***

Why predict LS means at the mean of x?

- Lines are parallel, so vertical distance is the same at any mass
- But, standard errors smaller near he middle of the data
- At mean of mass they will be as small as possible for both sexes at once

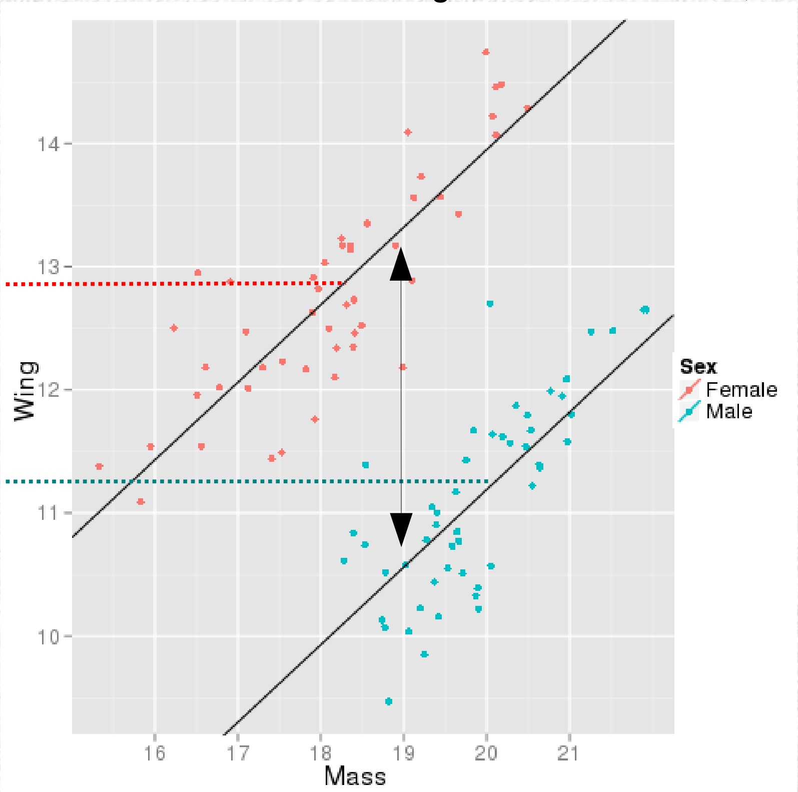


LS means are not always more different than raw means

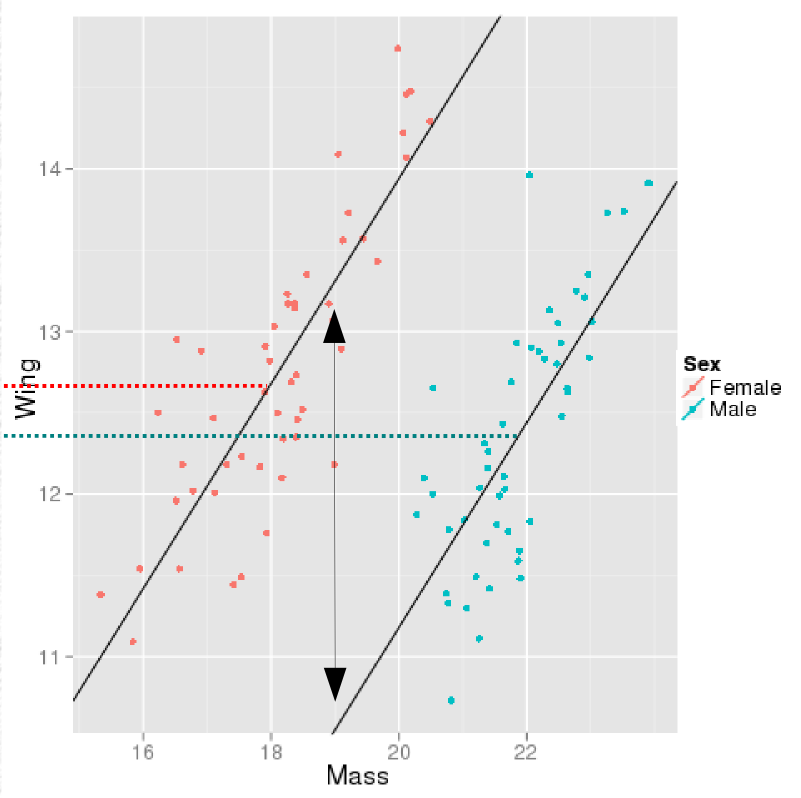
- In this first example, there is more difference between sexes when mass is accounted for
- But, accounting for a significant covariate could:
  - Enhance the difference between sexes
  - Reduce difference between sexes
  - Make it impossible to tell if there is a difference or not

When accounting for the covariate enhances differences

Wing: Females > Males, Mass: Females < Males

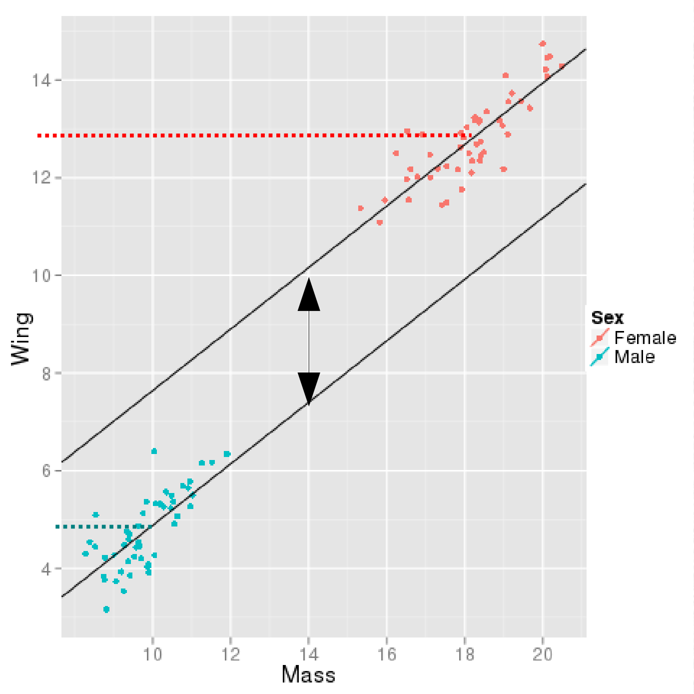


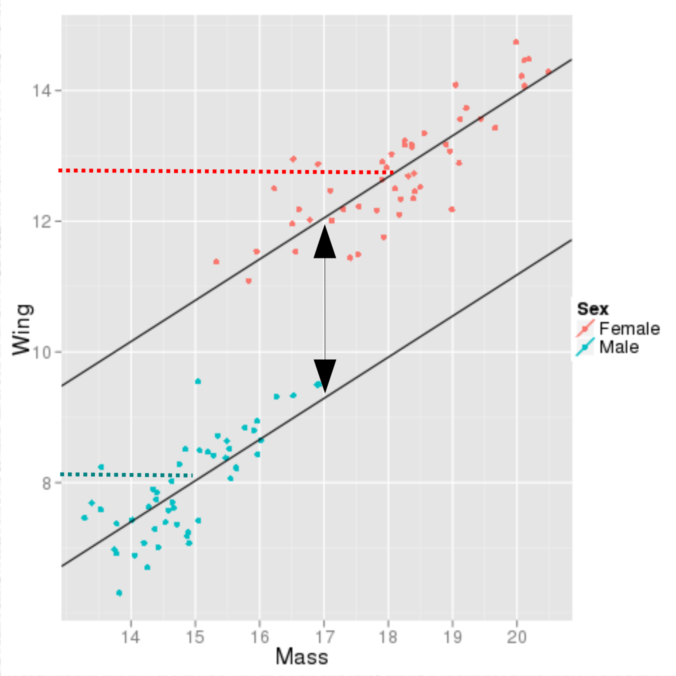
There is a difference in shape, and size is obscuring it



When accounting for the covariate reduces differences

Wing: Females > Males, Mass: Females > Males

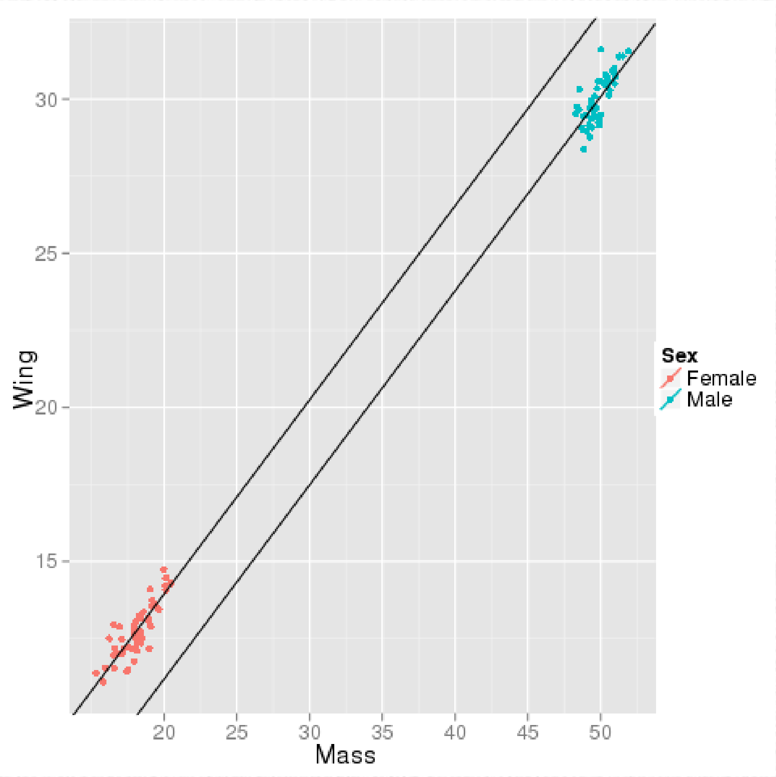


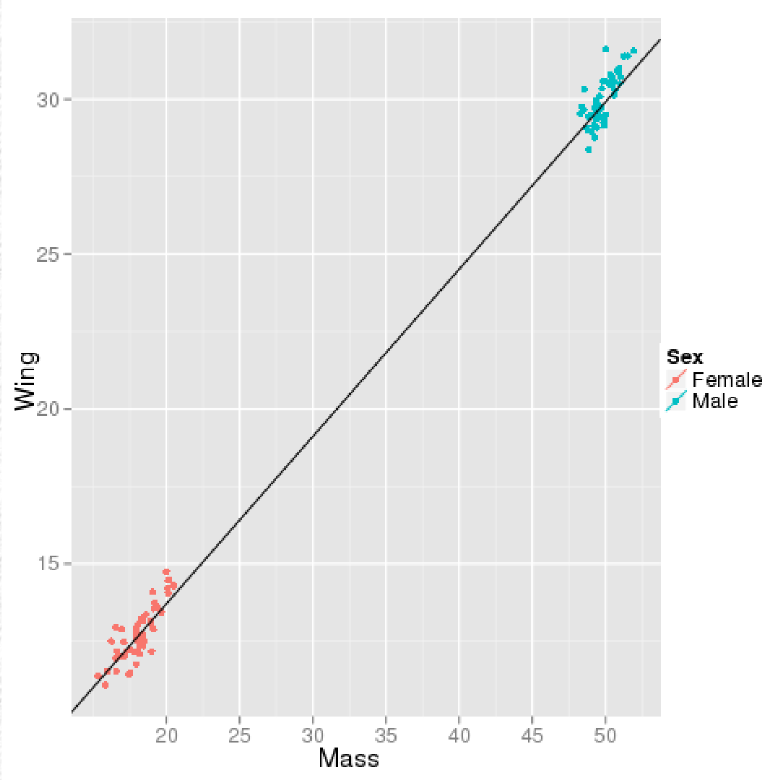


The difference in wing chord is in part due to a difference in size

Size is making the shape difference look bigger than it really is

When mass and sex are not statistically distinguishable





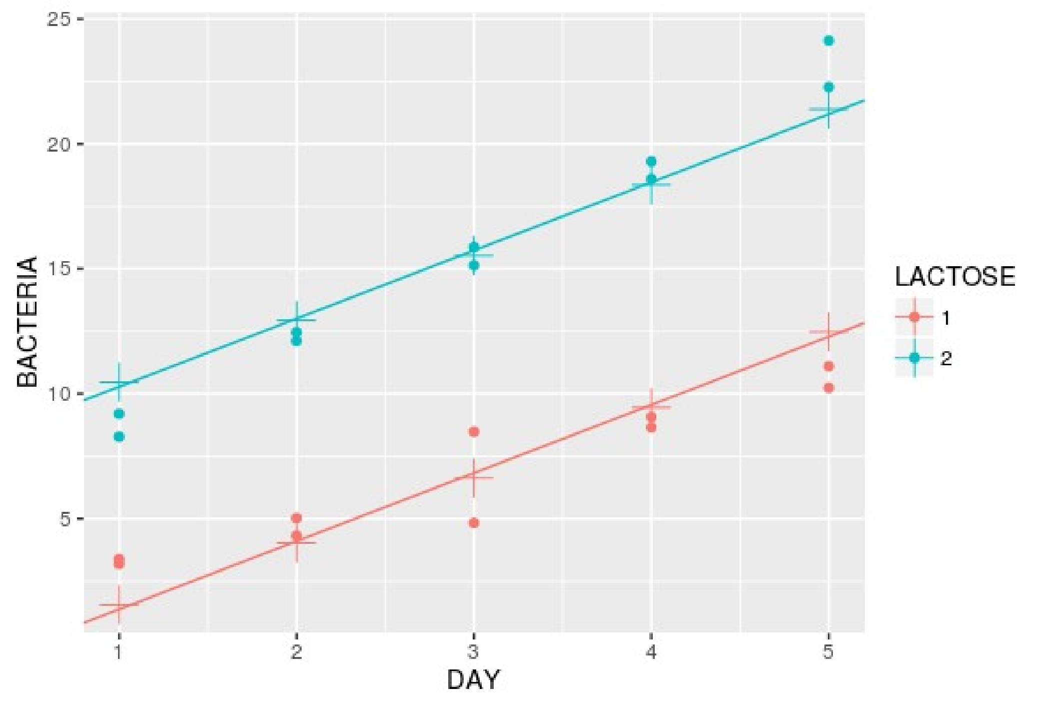
Sex is no longer significant, because a single line fits nearly as well as two parallel lines

Conclude that the difference in wing shape is entirely due to differences in size

Sometimes a variable can be treated as either continuous or categorical

- Some variables can either be expressed as numeric values, or can be expressed as categories
- Examples
  - Change over time – days since treatment (numeric), day of the week after treatment (category)
  - Dose – milligrams of dosage (numeric), or high, medium, or low dose (category)
- We could:
  - Treat the variable as categorical and block on it
  - Treat the variable as numeric and use it as a covariate
- How to choose?

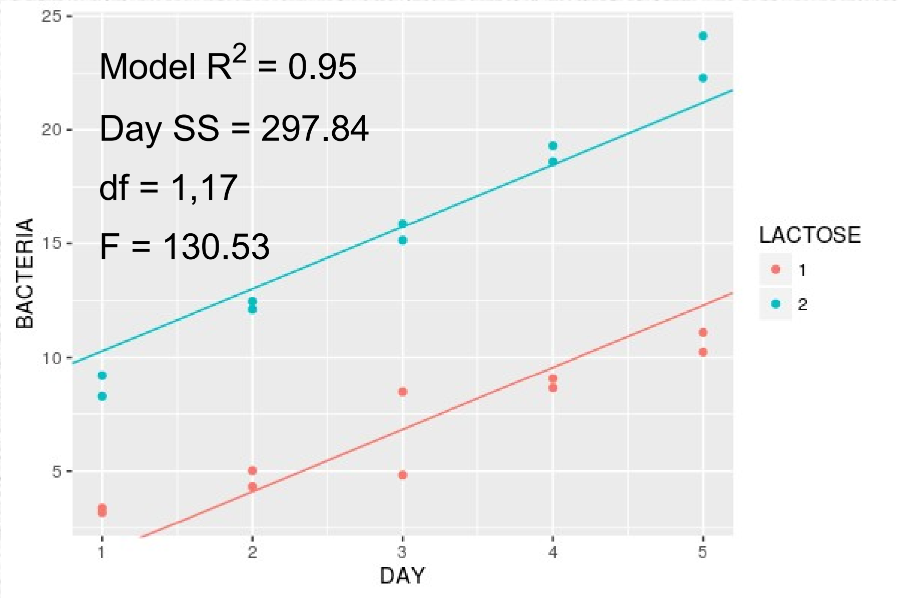
Growth of bacterial cultures over time under two different lactose treatments



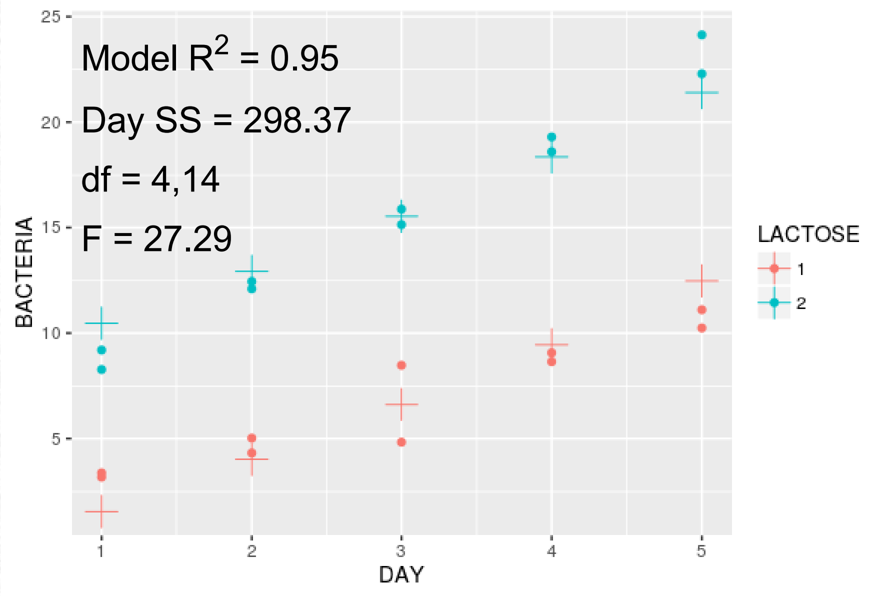
We can treat DAY as either categorical or numeric

Does it matter? If so, which is best?

If BACTERIA has a linear relationship with DAY...



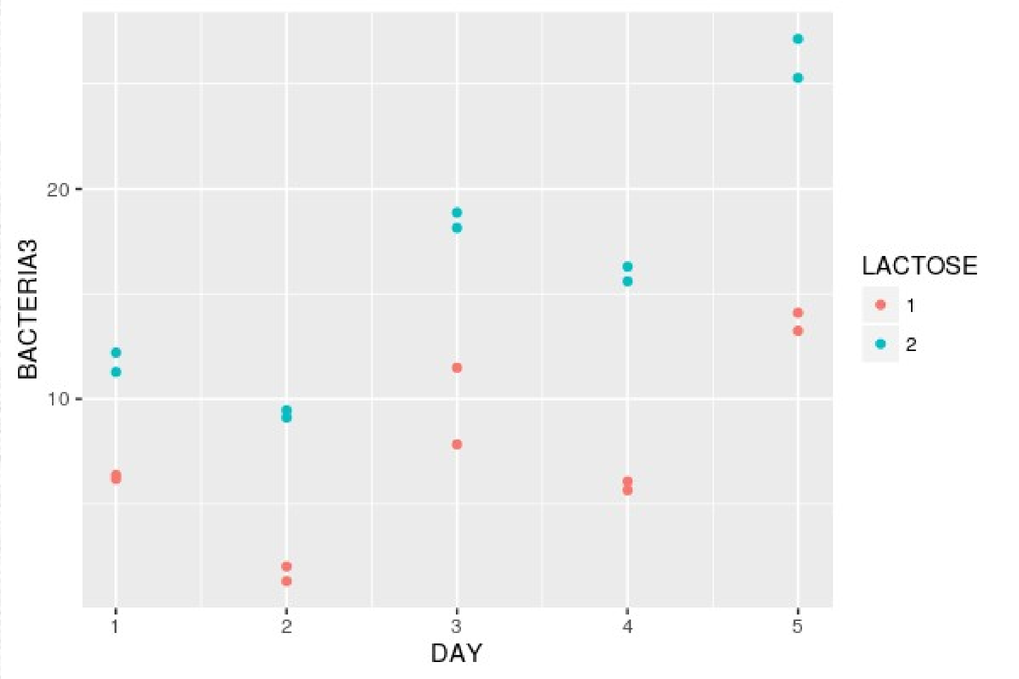
Regression gives a bigger F, because of the greater residual df → greater power

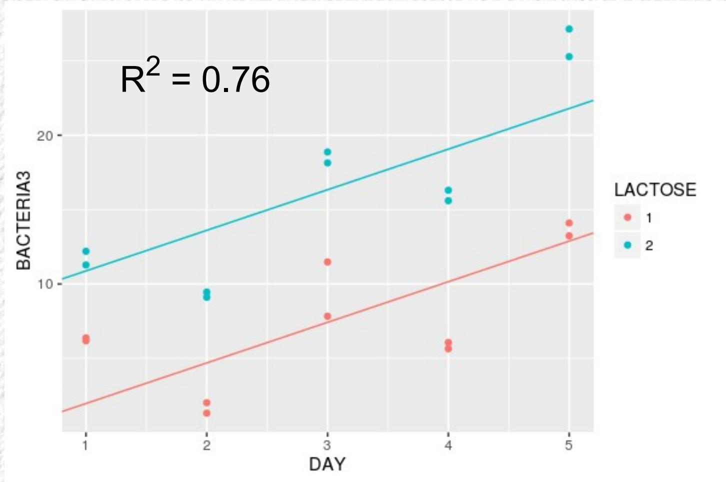


Block ANOVA has smaller F, because of 4 model df → lower residual df

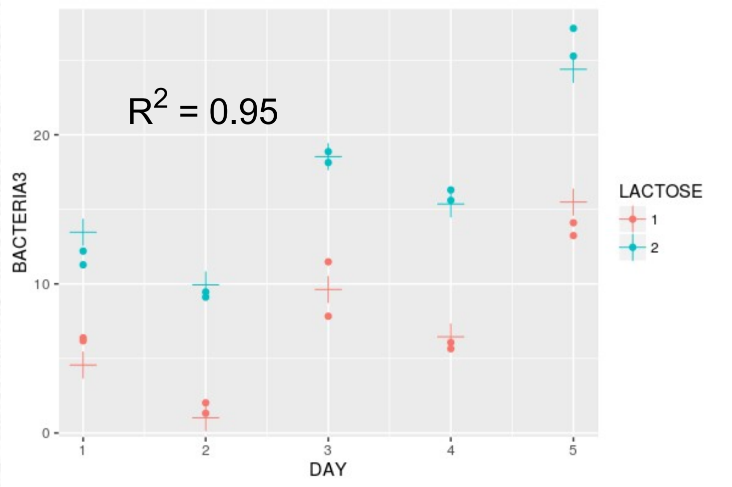
Lower power

If the pattern isn't linear...





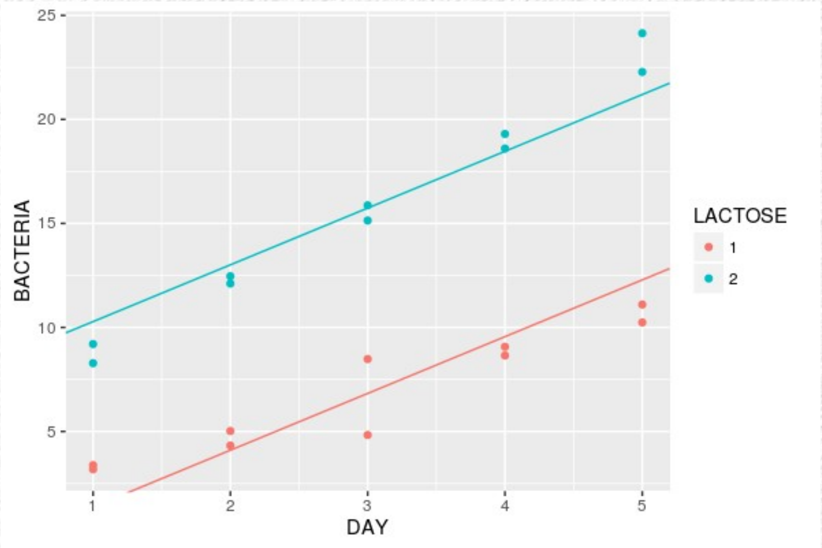
Straight line predicts poorly, can't follow the day to day oscillations



Means can stay close to the data

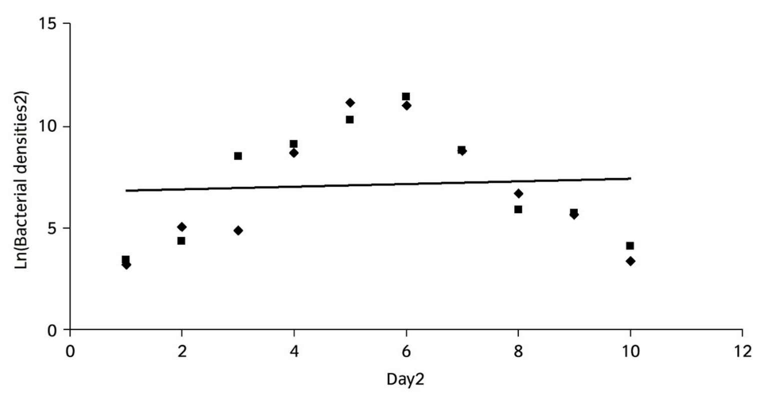
Use continuous variables when...

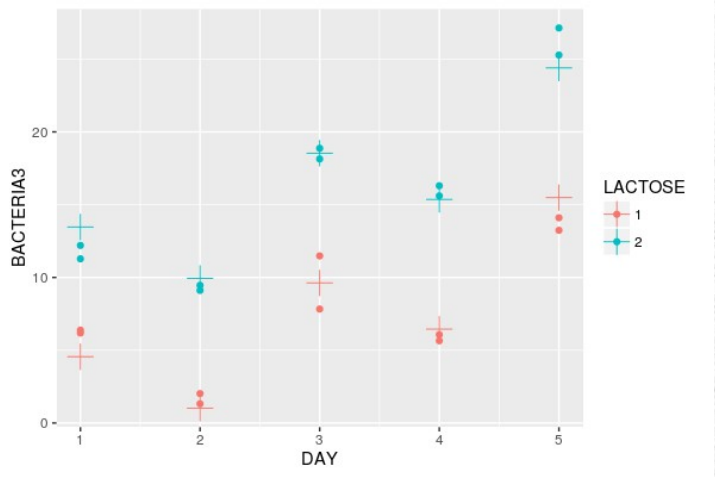
- There is good reason (supported by graphs) to expect a linear relationship
- Replication at each level would be low or absent if treated as a categorical grouping variable
- The question is appropriate (is there an increase or decrease over time?)
- A predictive equation is needed



Use categorical variables when...

- A linear relationship is not evident, such that the poor fit is a bigger problem than the loss of df
- Replication at each level sufficient
- A predictive regression equation is not needed
  - Can use orthogonal polynomials to test for trends





What's the model?

Response variable?

Continuous predictor?

Categorical predictor?

Which would be significant?

(a) Acclimatization of peak metabolic rate without insulatory acclimatization in deer mice