

Presentation material for day three:  
When the classical model assumptions  
don't hold... doh!

SCIE4402

## A review of what we have covered



- R - a new tool (old friend) for data analysis
- t-tests: various flavours (assumption issues)
- Proportions tests looked at differences (one sided tests)
- Sample size, precision, power
- ANOVA (factor variables)
- ANCOVA (confounding effect)
- Linear regression (continuous variables)
- Dummy variable regression = ANOVA

Recap on the ANOVA model and the  
link to the regression model

## Motivation for the ANOVA test



- Way of identifying existence of a difference in the mean across groups
- Looks at two sources of variation in the data
  - Variation in the group means from the grand mean
  - Variation within each group around each group mean
- The hypothesis testing test set up
  - Null: All means are equal
  - Alternate: At least one of the means is different
  - Low p-values: reject the null of no difference in the group means
- The test statistic is an F-statistic
  - There is a unique critical value for each sample size
- If we reject the null we move to pair-wise t-tests
  - Unequal or equal variance assumption

## The ANOVA statistic



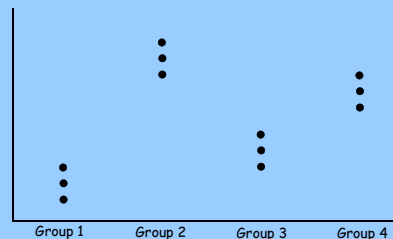
If the between group variation is large and the within group variation is low the F value is large. This is evidence against the null

$$F\text{-value} = \frac{\sum_i n_i (\bar{X}_i - \bar{X})^2 / (K-1)}{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2 / (N-K)} = \frac{\text{(standardised) between group variation}}{\text{(standardised) within group variation}}$$

$$F\text{-value} = \frac{\text{Large number}}{\text{Small number}} = \text{Large number}$$

Large F-value implies there are true differences  
(we can look for p-value < 0.05)

Yield ha<sup>-1</sup>



Yield  $\text{ha}^{-1}$

Just push the observations around about so we can see what is going on

Group 1 Group 2 Group 3 Group 4

Yield  $\text{ha}^{-1}$

Group means — Grand mean —

Group 1 Group 2 Group 3 Group 4

Yield  $\text{ha}^{-1}$

Between Group variation  
A measure of the variation due to differences in the group means

Group 1 Group 2 Group 3 Group 4

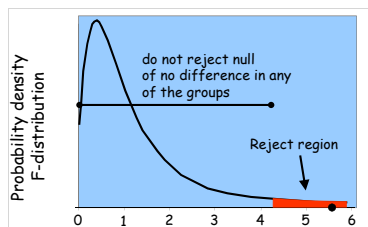
Yield  $\text{ha}^{-1}$

Within Group variation  
A measure of the variation of data within each group

Group 1 Group 2 Group 3 Group 4

## The F-statistic distribution

Large F-value reject the null that there are no differences (look for low p-value)



## Regression is also an ANOVA test

Yield  $\text{ha}^{-1}$

Model 1 explanation of the data

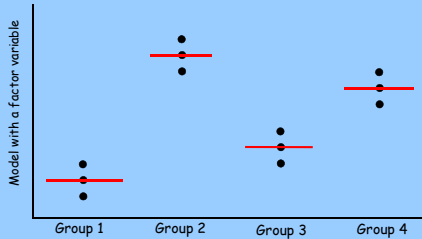
Intercept only model

Group 1 Group 2 Group 3 Group 4

The best explanation of the data is a straight line at the grand mean

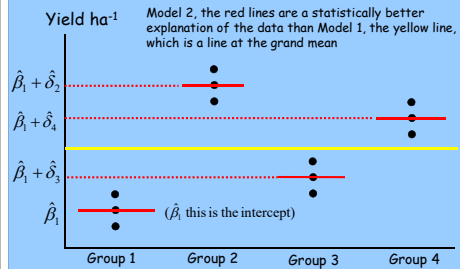
## Regression also an ANOVA test

Yield ha<sup>-1</sup> Model 2, explanation of the data



The best explanation of the data is a different mean for each group

## F-test to select between models

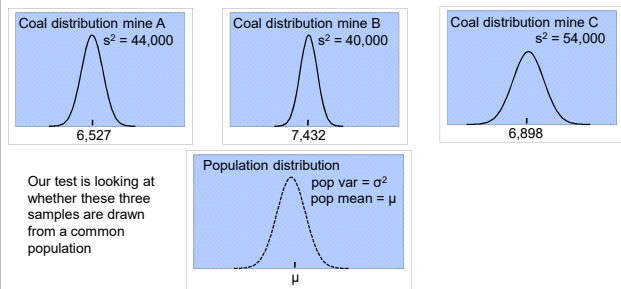


The F-statistic in the summary(.) output gives us this test statistic. Slightly more complex for many factors.

## ANOVA and regression are the same

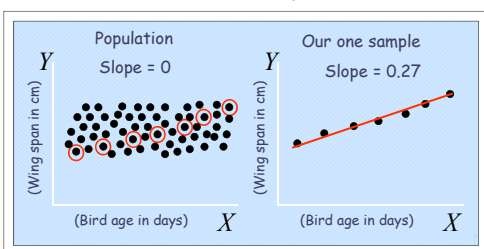
- This just helps for the way I set things up the discussion
  - I can largely talk about (linear) regression models without any loss of generality. The ANOVA model and the regression model are the same.
  - Note both use the `lm()` command in R
- We have the linear regression model
 
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$
- We have the ANOVA model
 
$$Y_{ij} = \mu + \tau_j + e_{ij}$$
- Talk about the error term assumptions

## A recap on what we are trying to do



## A recap on what we are trying to do

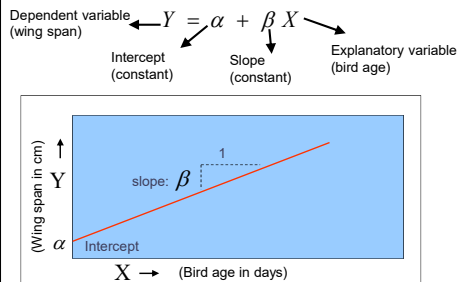
Make statements about the relationship between X and Y



Our test is looking at whether the one sample we have could come from a population with no relationship between X & Y.

## Have a kitkat... it's time for a break

## Recall the equation of a straight line

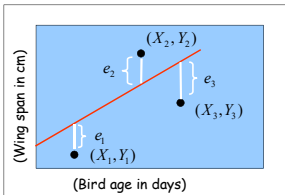


## How do we get estimates of alpha & beta

- In classical statistics there are three basic options
  - Generalised method of moments
  - Maximum likelihood
  - Least squares (ordinary and general)
    - Least absolute deviation is another, less common option
- Conceptually least squares works as follows
  - The error is the difference between the fitted value of our trend line and the actual observed value
  - Minimise the sum of the squared residuals

$$\text{Minimise} = e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2 = \sum_i e_i^2$$

## The least squares framework



The rule we use is to minimise the squared vertical error distances

Positive and negative errors cancel out

The regression line we fit will:

$$\text{Minimise} = e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2 = \sum_i e_i^2$$

## The model assumptions

- For each X value Y is given by a formula, for example:  $Y_i = \alpha + \beta X_i + e_i$
- Expected value of the error:  $E(e_i) = 0$
- The variance of the error:  $\text{Var}(e_i) = \sigma^2$
- The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
- The  $X_i$ s can be considered fixed in repeated samples
- No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$ 
  - The assumption is not about the data but the error
  - We don't really need the assumption

## The assumptions always hold right?

- Generally in undergraduate courses the assumptions do hold
  - Yes we made up the data!
- What do you do when the assumptions don't hold
- Focus is on a tool kit of solutions
- Look at multiple options
  - No one right solution
  - Sometimes there is no solution
- Balance of testing and visual inspection
- Approaches that are generally useful

## The model assumptions

- For each X value Y is given by a formula, for example:  $Y_i = \alpha + \beta X_i + e_i$
- Expected value of the error:  $E(e_i) = 0$
- The variance of the error:  $\text{Var}(e_i) = \sigma^2$
- The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
- The  $X_i$ s can be considered fixed in repeated samples
- No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$ 
  - The assumption is not about the data but the error
  - We don't really need the assumption

## Technical formula underlying our work



### Estimates and estimators

- Least squares estimators are general formulas
  - They represent random variables

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

- Estimates are specific values

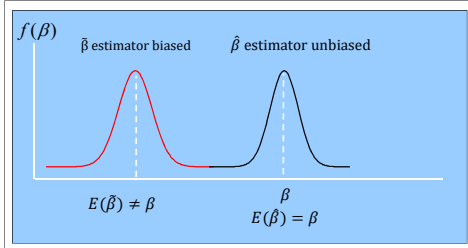
$$\hat{\alpha} = 7.60 \text{ cm} \quad \hat{\beta} = 2.70 \text{ cm}$$

- The formulas for the estimates don't include the error term
  - Normality assumption is not needed for getting estimates
- Estimates will still be:
  - Unbiased, Consistent, Minimum variance

## Unbiased estimates



- Average of many samples approaches true values for  $\alpha$  and  $\beta$



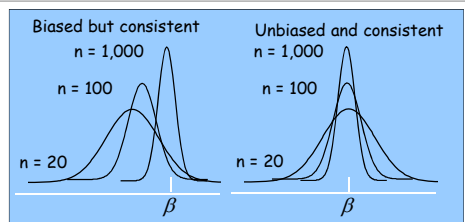
Does not say any one set of estimates are close to the true value

## Consistent



- Consistency means the sampling distribution becomes more concentrated on the true parameter value as sample size increase

So with more data we get more and more precise estimates of the true value

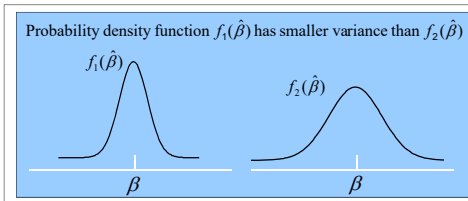


## Minimum variance



- The variance of a random variable is the average squared distance between the values and the mean

For a given sample size, if we repeat the experiment many times the distribution is smallest



## So where does the error term appear?



## Variance formula (hypothesis testing)



The variance of the slope

$$\text{Var}(\hat{\beta}) = \left[ \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \right]$$

The variance of the error term

Sum of squares of the X values about their mean

- Variance decreases with

- Lower error variance
  - Less uncertainty in the statistical model
- Greater variation in the X values
  - Increase in the sum of squares
- Increase in the number of observations
  - Via the sum of squares impact

## Same for the intercept (less interesting)



The variance of the intercept

$$\text{Var}(\hat{\alpha}) = \sigma^2 \left[ \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} \right]$$

The variance of the error term

Sum of squared X values

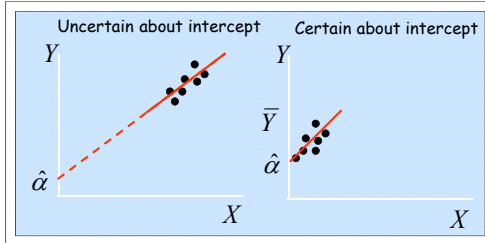
Number of observations

Sum of squares of the X values about their mean

## Visual of these properties



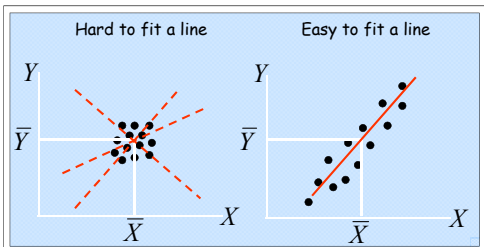
- Intuition about uncertainty (variance)



## Visual of these properties



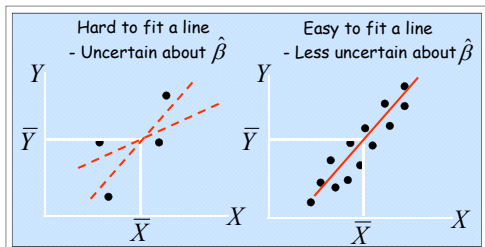
- Intuition about uncertainty (variance)



## Visual of these properties



- Intuition about uncertainty (variance)



## The role of the normality assumption

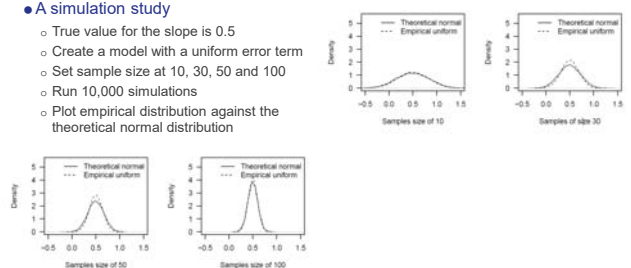


- We can show that if  $e_i \sim N(0, \sigma^2)$  then (conditional)  $Y_i$  is normal
- We can also show that our estimators (formula) can be written as a linear function of the  $Y_i$ , so that means the least squares estimators are normally distributed
 
$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2}\right) \quad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$$
- This property can then be used to derive confidence intervals and all the information we need for null hypothesis testing
  - but what if error is not normal? The CLT solves the problem

## The impact on hypothesis testing



- A simulation study
  - True value for the slope is 0.5
  - Create a model with a uniform error term
  - Set sample size at 10, 30, 50 and 100
  - Run 10,000 simulations
  - Plot empirical distribution against the theoretical normal distribution



## Normality assumption summary




- The assumption relates to the error term not the raw data
  - Raw data that is not normal is not a reason to transform data
    - Zero, one data, count data use a different model
- The normality assumption matters for hypothesis testing with small samples
  - But then low power is a bigger problem
- The assumption does not matter if you have a large sample
  - 1, 2, 3, 4... 30, infinity is how I think of it...
- Tests to detect departures from normality have low power when the sample size is small
- If a small sample do not obsess over the p-value
- If a large sample do not obsess over the p-value!

## Time for a break



## The model assumptions



1. For each X value Y is given by a formula, for example:  $Y_i = \alpha + \beta X_i + e_i$
2. Expected value of the error:  $E(e_i) = 0$
3. The variance of the error is constant:  $\text{Var}(e_i) = \sigma^2$  
4. The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
5. The  $X_i$ s can be considered fixed in repeated samples
6. No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$ 
  - The assumption is not about the data but the error
  - We don't really need the assumption

## Homoscedasticity vs heteroscedasticity



- Homoscedasticity is the assumption of constant error variance
- Formally we wrote something like  $\text{Var}(e_i) = \sigma^2$  as a model assumption
- To estimate the error variance we have

$$\text{Var}(e_i) = \sigma^2 = E[e_i - E(e_i)]^2 = E(e_i^2)$$

We have expected value of error term equals zero

Expected value or average of the squared residuals

## Why does the error term variance matter



- Error variance does not feature in the formula for the point estimates
  - This will be a useful result for us we can make use when solving the problem
- Error variance does feature in the calculation of the estimate variance formula
  - If the assumption is wrong we make wrong conclusions... doh!
- Type of mistake for ANOVA
  - My experience is an increase the chance of a Type II error
    - There is a difference and we fail to detect it
  - Similarly for follow up pair-wise t-tests
- Type of mistake for regression models
  - Just know that we can make 'wrong' decisions
  - Hard to identify systematic effects

## The classic error variance estimator



The intuitive estimator  $\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$  - but the errors are unobserved

Replace the unobserved error with least squares error  $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$

Add a correction for the number of parameters estimated  
- intercept and a slope

Our revised estimator is:  $\hat{\text{Var}}(\hat{e}) = \hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-2}$

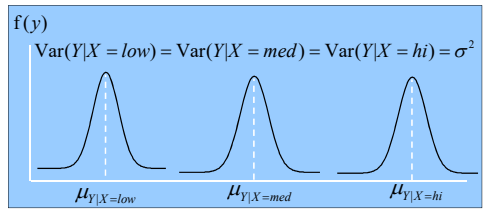
Unbiased estimator of the error variance

## The practicalities of the assumption



- For the ANOVA model, in practice the assumption means that the spread of the data is the same for different levels of the factor variable

The distribution conditional on the factor levels is the same

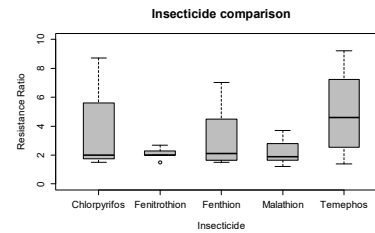


## Methods of detection ANOVA



- A visual inspection of the data or a formal test
  - Bartlett.test(); bptest(); leveneTest()
    - low p-values lead to the rejection of the null hypothesis of homoscedasticity

- Practical issue is that the pooled variance will be large – we fail to detect differences

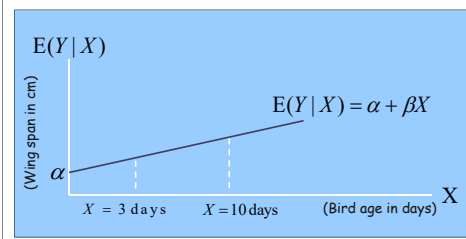


## The issue and regression models



- Visualise the linear model

- For each X value (bird age) the regression function gives the average value (conditional mean) for wing span (Y)

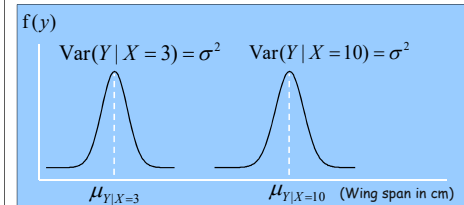


## Conditional distribution



- On day three we get some measurements and on day ten we get some measurements, the means will be different but the uncertainty surrounding the estimates is the same

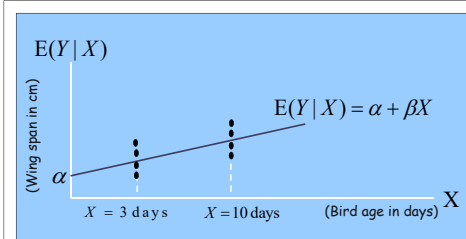
Conceptually think of an ANOVA with many measurements at age 3 and at age 10



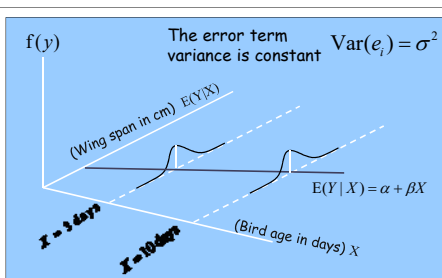
## The formal regression model



- What the data looks like with constant error variance



## When the error variance is constant

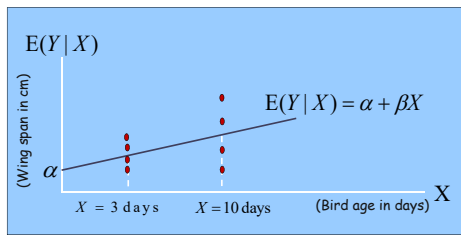


- Equal uncertainty for all observations



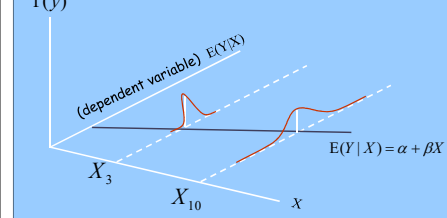
## The formal regression model

- What the data may look like with unequal error variance



## When the error variance is not constant

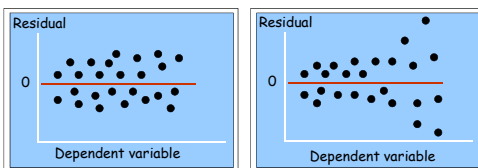
Violation of the assumption of constant error term variance  $\text{Var}(e_i) \neq \sigma^2$



- Less certainty about observations at 10 days than at 3 days

## Test for heteroscedasticity: regression

- Informally, we can look at residual plots (we often have many X variables)



- Formal test is the bptest
  - With low p-values we reject the null of homoscedasticity

## Hetero consequences and solutions

- Estimators for slopes and intercept do not use the error variance
  - That means we can use these values
- The variance formulas do use the error variance
  - Hypothesis testing will be incorrect
- Use what is called a robust method to get the variances
  - With the robust method hypothesis testing will be correct
  - In some cases slight differences in test format
    - Practical p-value decision rule is unchanged
- 1. Use the point estimates from `lm()`
- 2. Use the robust formula available in R
  - Modern version of Welch type adjustments for t-tests

## Example of the steps ANOVA

- Load an additional R package
  - `library(AER)`
- Estimate the linear model (point estimates are valid)
  - `lm.wheat <- lm(yield~fertiliser, data=wheat)`
- Check the variance assumption
  - `leveneTest(lm.wheat)` (if we fail the test)
- Conduct the robust ANOVA
  - `anova(lm.wheat, vcov = vcovHC)`
    - The big A anova test is different to the little a anova test
    - The 'vcov = vcovHC' bit means conduct the test with the Hetero robust covariance matrix
    - Interpret the results as you would a standard anova test
- Move on to robust multiple comparisons

## Example of the steps Mult. comparison

- Pairwise t-tests for multiple comparisons
  - Set the pooled variance option equal to FALSE
- Tukey test is perhaps more common
  - Load another package `library(multcomp)` (may need to download first)
- Classic Tukey approach with constant error variance (two steps here)
  - `tuk.input <- aov(yield~fertiliser, data = wheat)`
  - `TukeyHSD(tuk.input)`
- Robust Tukey approach with non constant error variance (two steps)
  - `tuk.wheat <- glht(lm.wheat, mcp(fertiliser = "Tukey"), vcov=vcovHC)`
  - `summary(tuk.wheat)`

## Example of the steps for Regression



- Load an additional R package (or download first of not on a uni computer)
  - library(AER)
- Estimate the linear model
  - `lm.lobster <- lm(size~distance, data = lobster.data)`
- Check the variance assumption
  - `bptest(lm.lobster)` (more than one test type)
- Obtain the estimates, SE, t-value, and p-value
  - `coefest(lm.lobster, vcovHC(lm.lobster))`
- Get the other model details (mainly R<sup>2</sup>)
  - `summary(lm.lobster)`
  - Note: it is possible to put all the information together
    - Example code in the script files

## So we use vcovHC what does it do



- In a standard model we assume the error variance is constant
  - We use the same error variance estimate for each data point
    - This estimate is the (n-k) average of the squared error terms

$$\text{Var}(e_i) = \sigma^2 \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$$

- In the robust model rather than use the average squared error term we use the actual squared error term (with adjustment)
  - We use a different variance error estimate for each data point
    - $\text{Var}(e_i) = \sigma_i^2 = e_i^2 \times \text{adjustment}$
    - There are at least four adjustment options. Each adjustment slightly inflates the estimate. An example adjustment is (n/n-k)

## Alternative methods: GLS



- The robust method is very general
  - The approach can always be considered valid
- Generalised least squares or similar via maximum likelihood can be more 'efficient'
- Method requires some input from the researcher about the nature of the hetero
- For ANOVA this is easy
  - The hetero is related to the groups
    - We have a specific solution we don't need a general solution
    - We model the hetero directly as a function of the factor levels
- For Regression the best option is less clear
  - Hetero could be related to many sources
    - Solid guess is that hetero related to dependent variable

## GLS: ANOVA implementation



- Use an R package (or download first of not on a uni computer)
  - library(nlme)
- Two factor ANOVA comparison
  - `fit.homo <- lm(Yield~Fert*Water, data = spuds)`
  - `fit.het <- gls(Yield~Fert*Water, weights=varident(form=~1|Fert*Water), data=spuds)`
    - Let the variance be different across groups, but the same within groups
  - `Anova(fit.het)`
    - Interpret the output the same as a standard anova()
- Multiple comparisons
  - Can use pairwise t-tests (no pooled variance)
  - Tukey multiple comparisons get a little complicated
    - The options in the multcomp package don't work
    - There is some code in relevant handout to make this work

## GLS: Regression implementation



- Use an R package (or download first of not on a uni computer)
  - library(nlme)
- Regression comparison
  - `food.lm <- lm(food.spend~weekly.income, data = Food.data)`
  - `food.gls <- gls(food.spend~weekly.income, weights = varPower(), data = Food.data)`
    - Each error variance is different & place greater 'weight' on estimates with low error variance
  - `summary(food.gls)`
    - Interpret the output the same as a standard summary()
  - Use `anova()` to compare \ select between models
- Worth seeing what is common in your discipline
  - Direct modelling of variance is more art than science
  - You just get a feel for it over time

## GLS what is going on



- This is complicated...
- GLS stands for generalised least squares
  - ... but this is not exactly what R does... ahhhhhh!
  - If you are interested consult the nlme guide or a statistics textbook
    - How you set up the likelihood function
    - You can also implement in other ways
- GLS and the link to mixed models is something worth knowing about
  - One the one hand it is very old skool
  - On the other computer power makes it a modern approach
  - The nlme package is my favourite R package

## Hetero summary




- Common problem with most (all) large data sets Regression and ANOVA
  - Model assumption violation (constant error variance)
- Today there is a fix for the issue
  - Relatively easy to implement
- Robust covariance matrix
  - Also called White's or Huber-White
  - Details in the R help guide
  - Details in the relevant handout
  - Can be some technical issues
    - If robust crashes ask, there are work arounds (hcc0)
- Direct modelling approach (gls)
  - Increase in efficiency, but more complex, and more 'art'

## Time for a break



## The model assumptions



- 1. For each X value Y is given by a formula, for example:  $Y_i = \alpha + \beta X_i + e_i$
- 2. Expected value of the error:  $E(e_i) = 0$
- 3. The variance of the error is constant:  $\text{Var}(e_i) = \sigma^2$
- 4. The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$ ,  $\text{Cov}(e_i, e_{i-1}) = 0$  
- 5. The  $X_i$ s can be considered fixed in repeated samples
- 6. No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$ 
  - The assumption is not about the data but the error
  - We don't really need the assumption

## Autocorrelation in the error term

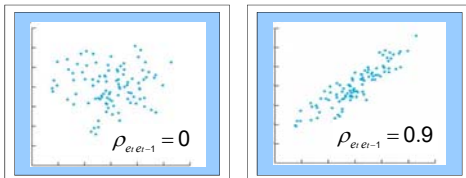


- The assumption about the relationship between error terms is  $\text{Cov}(e_i, e_j) = 0$ 
  - Knowledge of error term  $e_i$  tells you nothing about the error term  $e_j$
- With time series data the data is listed in a specific order
  - Here there can be a big issue with correlation in error terms in adjacent time periods
  - We usually change the subscript to identify this  $\text{Cov}(e_i, e_{i-1}) = 0$
  - Not a problem in cross section data
    - You can just change the order
    - Spatial correlation can also be an issue
- When to worry about autocorrelation and what to do
  - All time series models so not relevant for ANOVA
    - Generalisation from robust HC to HAC
    - Similar gls solution

## What does the problem look like



- Visual inspection of the error terms – is there a pattern (also look at the ACF)



- Formal test – bgtest
  - Low p-value evidence against the null of no autocorrelation

## What does it mean in practice?



- As with heteroscedasticity we know that the estimators (which are the formulas we use) do not rely on this assumption
  - That means the point estimates can be used, Yay!
- As with heteroscedasticity there will be a problem with the estimate SE, hence the t-values and, hence the p-value is wrong
  - Our hypothesis testing will be wrong
  - Generally there is an increase in Type I error rate
  - Many textbook will show the AR(1) formulas
- There is a generalisation from the hetero robust option
  - Need the AER package
- There is an option to use the GLS command
  - Need the nlme package

## Robust regression implementation



- Load an additional R package (or download first of not on a uni computer)
  - library(AER)
- Estimate the linear model
  - `lm.lobster <- lm(log(weight)~log(week), data = Profit)`
- Check the autocorrelation assumption
  - `bgtest(lm.lobster)`
- Obtain the estimates, SE, t-value, and p-value
  - `coefest(lm.lobster, vcovHAC(lm.lobster))`
- Obtain other model details (mainly R<sup>2</sup>)
  - `summary(lm.lobster)`
  - Note: it is possible to put all the information together
    - Example code in the script files

## GLS: Regression implementation



- Use an R package (or download first of not on a uni computer)
  - library(nlme)
- Regression comparison
  - `lob.lm <- lm((log(weight)~log(week), data = Profit)`
  - `lob.gls <- gls(log(weight)~log(week), corr = corARMA(p=1, q= 0), data = Profit)`
    - *p* denotes the autocorrelation terms and *q* the moving average terms
    - AR1, in practice solves pretty much everything (no MA terms for us)
  - `summary(lob.gls)`
    - Interpret the output the same as a standard summary()
  - Use `anova()` to compare \ select between models
- Worth seeing what is common in your discipline
  - Direct modelling of variance is more art than science
    - You just get a feel for it over time

## What you are actually doing HAC



- When you fail a `bgtest()` it means there is a pattern in the error terms
  - The model assumption is that there is no pattern
- The robust method generalises from the HC method
  - The HAC method starts by allowing  $\text{Var}(e_i) = \sigma_i^2$   $\hat{\sigma}_i^2 = \hat{e}_i^2 \times \text{adjustment}$ 
    - Exactly the same as the HC option
  - Then also allows  $\text{Cov}(e_i, e_{i-1}) = \text{Cov}(\hat{e}_i, \hat{e}_{i-1}) \times \text{adjustment}$ 
    - Then imposes some pattern of quickly declining correlation
    - Different structures are possible
- So anytime you address autocorrelation you also must allow for hetero (whether needed or not)
  - Imposes structure on the covariance matrix
  - Hard to write without matrix notation
    - You can look this stuff up

## GLS what is going on



- This is complicated...
- GLS stands for generalised least squares
  - ... but this is not exactly what R does...
  - If you are interested consult the nlme guide or a statistics textbook
    - How you set up the likelihood function
    - Can be important in some disciplines

## Time for a break



## The model assumptions

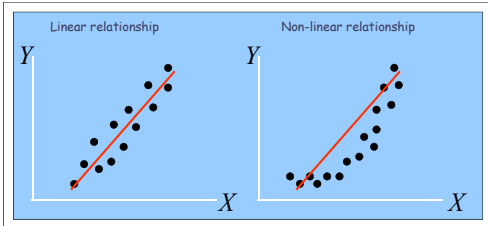


1. For each X value Y is given by a formula, for example:  $Y_i = \alpha + \beta X_i + e_i$
2. Expected value of the error:  $E(e_i) = 0$
3. The variance of the error is constant:  $\text{Var}(e_i) = \sigma^2$
4. The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
5. The Xs can be considered fixed in repeated samples
6. No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$



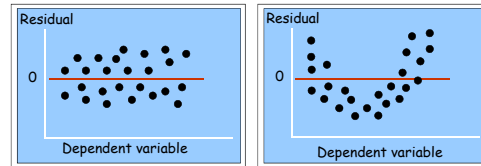
## Model functional form

- For each X value of the Y value is given as:  $Y_i = \alpha + \beta X_i + e_i$ 
  - Nonlinearities or omitted variables



## Checking the model functional form

- Informally, we can look at residual plots (we often have many X variables)
  - No problem
  - Functional form problem



- A pattern in the residuals means we have a problem

## Testing for nonlinearity (and omit. var.)

- There is a formal test for checking model functional form (Ramsey RESET)
- The test can be implemented in many ways one option is as follows
  - First estimate the model, say:  $Y_i = \alpha + \beta X_i + e_i$ 
    - Could have many explanatory variables
  - Save the fitted values from this model:  $\hat{y}$
  - Estimate a second model  $Y_i = \alpha + \beta X_i + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + u_i$ 
    - Which looks odd
    - There are other versions
  - Use an F-test for the null  $H_0 : \hat{\gamma}_1 = \hat{\gamma}_2 = 0$
- Why does this work?
  - The  $\hat{y}_i^2$  and  $\hat{y}_i^3$  are non-linear functions of the Xs
  - So if we reject the Null we have some nonlinear relationships
  - Or we have missing variables (or hetero actually)

## Good news and bad news

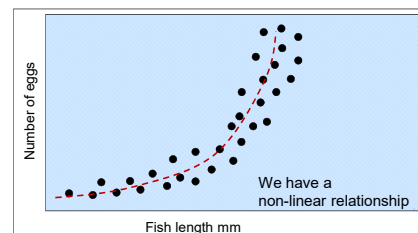
- It is important to check your model
  - Issue has appeared in some legal cases
    - Value of recreational assets
- The good news
  - The test is automated in R as the RESET test
    - Safe to use the defaults but you can edit if interested
  - Low p-values are evidence against the null
    - Low p-values mean you have a problem
- The bad news
  - Can't discriminate between models
    - Two functional forms can pass the test
  - Does not provide guidance on how to fix
    - Transformation, omitted variable, hetero

## A process to think about

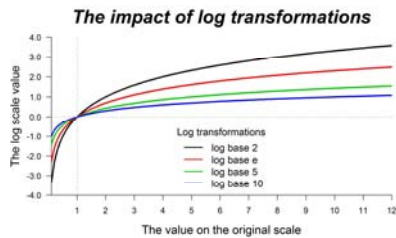
- Explore the data via different plots
  - Hard with many X variables, but sometimes you see something
- Guess an appropriate functional form (focus on logs and power terms)
- Estimate the model form you guess
- Use the RESET test to check your guess
  - Low p-values we reject the null (we have a problem)
- If the model passes it is possibly ok, but check other options
  - Passing the test does not mean you have the right model
- If the model fails the test look at other options
  - Try transformations and is there a missing variable
  - If nothing works, check hetero as that could be the problem
  - Be aware of 'overfitting' including rubbish to pass the test

## Nonlinear data examples

Large female fish produce a lot more eggs



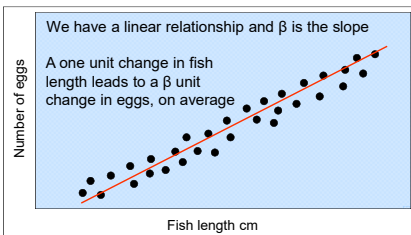
## Recall log transformations



## Data transformation options

- No transformation  $Y_i = \alpha + \beta X_i$
- log (Y) transformation  $\log(Y_i) = \alpha + \beta X_i$
- log (X) transformation  $Y_i = \alpha + \beta \log(X_i)$
- log (Y) and log (X)  $\log(Y_i) = \alpha + \beta \log(X_i)$

## Linear-linear scale plot

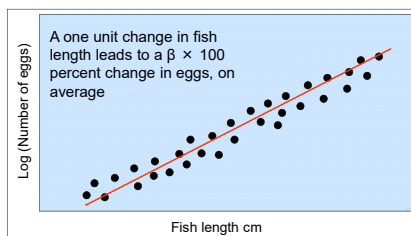


if  $\beta = 6500$  we say a one cm increase in fish length leads to an increase in egg production of 6500

## Interpretation of the slope

- Log Y transformation  $\log(Y_i) = \alpha + \beta X_i$
- How to interpret the slope coefficient  $\beta$ 
  - A one unit change in X leads to a  $\beta \times 100$  percent change in Y
- Example
  - Let Y measure fish egg production (no.)
  - X measure fish length (cm)
  - $\beta = 0.095$   $\log(Y_i) = \alpha + 0.095X_i$
  - Then we say: A one cm increase in fish length results, on average, in an increase in egg production of 9.5%

## Log-linear scale plot

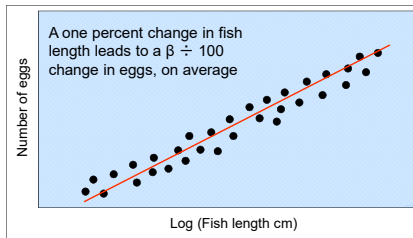


if  $\beta = 0.095$  we say a one cm increase in fish length leads to an increase in egg production of 9.5%

## Interpretation of the slope

- Log X transformation  $Y_i = \alpha + \beta \log(X_i)$
- How to interpret the slope coefficient  $\beta$ 
  - A one percent change in X leads to a  $\beta \times 100$  unit change in Y
- Example
  - Let Y measure fish egg production (no.)
  - X measure fish length (cm)
  - $\beta = 32000$   $\log(Y_i) = \alpha + 32000X_i$
  - A one percent increase in fish length results, on average, in an increase in egg production of 320 eggs

## Linear-log scale plot

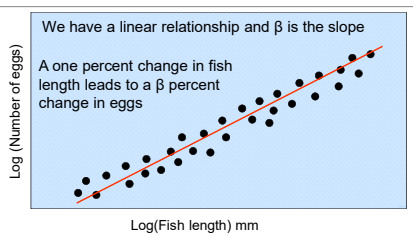


if  $\beta = 32,000$  we say a one percent increase in fish length leads to an increase in egg production of 320 eggs

## Interpretation of the slope

- Log X & Y transformation  $\log(Y_i) = \alpha + \beta \log(X_i)$
- How to interpret the slope coefficient  $\beta$ 
  - A one percent change in X leads to a  $\beta$  percent change in Y
- Example
  - Let Y measure fish egg production (no.)
  - X measure fish length (cm)
  - $\beta = 4.8$   $\log(Y_i) = \alpha + 4.8X_i$
  - A one percent increase in fish length results, on average, in an increase in egg production of 4.8 percent

## Log-log scale plots

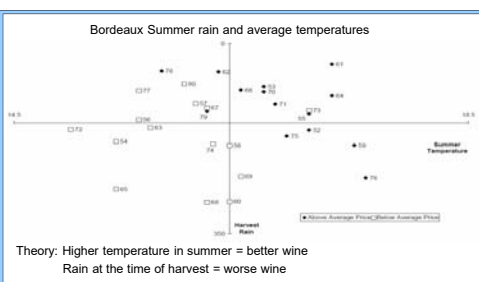


if  $\beta = 4.8$  we say a one percent increase in fish length (L) leads to a 4.8 percent increase in eggs (E)

## Regression models

- Fit a trend line to the data and describe the intercept and the slope
- If the data is not linear we have some transformations
  - Log of the Y value
  - Log of the X value
  - Log of the X value and the Y value
- We can just check all the possible examples
- Also check other options
  - Quadratic and other terms

## My honours research project ...so long ago



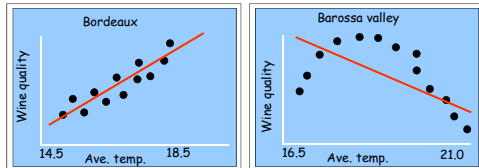
## So what kind of model

- Based on Bordeaux  $Q_t = \beta_1 + \beta_2 R_t + \beta_3 T_t + e_t$ 
  - Where,  $Q_t$  is wine quality,  $R_t$  is rainfall at harvest, and  $T_t$  is ave. monthly temperature during the growing season
- What did I think should happen
  - wine quality decreases with rain fall at harvest so  $\beta_2$  should be negative
  - wine quality increases with higher average temperatures so  $\beta_3$  should be positive
- Actually found  $\beta_2$  negative but also found  $\beta_3$  negative
  - What was going on?
- Model specification problem
  - - Bordeaux is colder than the Barossa valley

## A model specification problem



- For the Barossa Valley I needed a quadratic  $Q_i = \beta_1 + \beta_2 R_i + \beta_3 T_i + \beta_4 T_i^2 + e_i$ 
  - RESET test helped – something wrong but not what was wrong
  - Viticulture textbook also helped – stats is no substitute for discipline knowledge



## Time for a break



## Problems and almost problems



## Multiple linear regression model



- Theory is that peanut yield depends on nitrogen and phosphate
  - Dependent variable is peanut yield in kg per ha
  - Explanatory variables are Nitrogen applied in kilograms per ha and Phosphate based fertiliser applied in kilograms per ha

- Set up the basic model

$$\text{Yield} = \text{intercept} + \beta_2(\text{Nitrogen}) + \beta_3(\text{Phosphate})$$

$$Y = \beta_1 + \beta_2 N + \beta_3 P$$

## What do the parameters tell us ?



$\beta_1$  = Yield when you apply no nitrogen and no phosphate  
- Does not have to have a meaning

$\beta_2$  = Change in peanut yield (kg per ha) when you increase nitrogen by one unit (kg per ha)  
and  
the amount of phosphate fertiliser (P) is held constant

$\beta_3$  = Change in peanut yield (kg per ha) when you increase phosphate fertiliser by one unit (kg per ha)  
and  
The amount of nitrogen applied (N) is held constant

## The regression model



$$Y_i = \underbrace{\beta_1 + \beta_2 N_i + \beta_3 P_i}_{\text{Systematic part}} + \underbrace{e_i}_{\text{Random part}}$$

Farm	Yield ( $Y_i$ ) kg/ha	Nitrogen ( $N_i$ ) kg/ha	Phosphate ( $P_i$ ) kg/ha
1	72.1	5.2	12.8
2	62.8	4.1	12.0
...	...	...	...
80	75.4	6.8	11.4



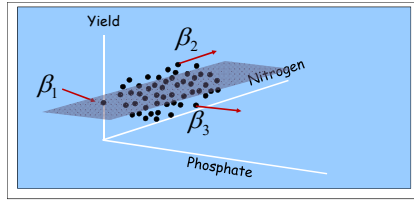
## A 3D representation of relationships



$\beta_1$  Where the plane cuts the vertical axis

$\beta_2$  Slope of the plane in the nitrogen axis

$\beta_3$  Slope of the plane in the phosphate axis



## General regression model



General model  $Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_K X_{iK} + e_i \quad i = 1, \dots, n$

-  $X_{i1}$  always equal to one for the intercept

Any model is just a specific case of the general model

-  $Y_i$  = peanut yield per hectare -  $X_{i1}$  = to one for the intercept

-  $X_{i2} = N_i$  = nitrogen applied per ha

-  $X_{i3} = P_i$  = phosphate applied per ha

$Y_i = \beta_1 + \beta_2 N_i + \beta_3 P_i + e_i$  - Just a specific case

Use generalised notation for assumptions  $X_i$

## The model assumptions



- 1. For each set of X values Y is given by a formula, for example:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_K X_{iK} + e_i$$

- 2. Expected value of the error:  $E(e_i) = 0$
- 3. The variance of the error is constant:  $\text{Var}(e_i) = \sigma^2$
- 4. The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
- 5. The  $X_i$ s can be considered fixed in repeated samples
- 6. No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$



## Add some regions to a model



- Say we have data from Toodyay and Greenough and we want to control for this

$$Y_i = \beta_1 + \beta_2 N_i + \beta_3 P_i + \beta_4 G_i + \beta_5 T_i + e_i$$

## What does the data structure look like



Farm	( $Y_i$ )	( $N_i$ )	( $P_i$ )	( $G_i$ )	( $T_i$ )	( $G_i$ ) + ( $T_i$ )
1	72.1	5.2	12.8	0	1	(0+1) = 1
2	68.1	4.2	10.5	0	1	(0+1) = 1
3	75.4	6.4	11.6	1	0	(0+1) = 1
4	71.4	6.4	11.6	1	0	(0+1) = 1
...	...	...	...	...	...	...
80	75.8	6.9	12.2	0	1	(0+1) = 1

-  $X_{i1}$  always equal to one for the intercept

We have  $X_i = (G + T)$ , which is an exact linear combination  
Model can not be estimated

## Model assumptions



$Y_i = \beta_1 + \beta_2 N_i + \beta_3 P_i + \beta_4 G_i + \beta_5 T_i + e_i$  becomes

$$Y_i = \beta_1 + \beta_2 N_i + \beta_3 P_i + \beta_4 G_i + e_i$$

$\beta_1$  = Intercept for farms in Toodyay

$\beta_1 + \beta_4$  = Intercept for farms in Greenough

$\beta_4$  = Difference in intercept between Toodyay farms and Greenough farms

Dummy variable trap for exact collinearity. Hard to follow in the summary table

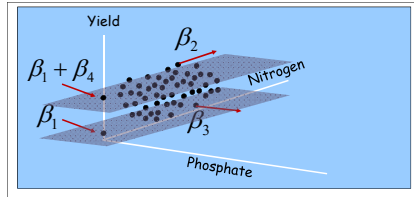
## A 3D representation of relationships



$\beta_1$  Where the plane cuts the vertical axis

$\beta_2$  Slope of the plane in the nitrogen axis

$\beta_3$  Slope of the plane in the phosphate axis

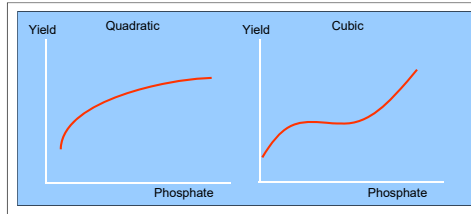


## Non-linear relationships



- We can use a quadratic to describe diminishing or increasing marginal changes in the response variate

But we can also use,  $X^3$  or  $X^4$  depending on the shape we wanted to approximate



## The model assumptions



- 1. For each set of X values Y is given by a formula, for example:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_K X_{iK} + e_i$$

- 2. Expected value of the error:  $E(e_i) = 0$
- 3. The variance of the error is constant:  $\text{Var}(e_i) = \sigma^2$
- 4. The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
- 5. The  $X_i$ s can be considered fixed in repeated samples
- 6. No exact relationships between explanatory variables
- (optional) The error normally distributed:  $e_i \sim N(0, \sigma^2)$



## Hypothesis testing and power terms



- Individual coefficient t-tests

- might not pick up significance due to multicollinearity

- Joint test of significance for phosphate (polynomial terms)

- Use the F-test approach

- restricted model has all terms of interest equal to 0

Unrestricted model  $Y_i = \beta_1 + \beta_2 N_i + \beta_3 P_i + \beta_4 P_i^2 + e_i$

Restricted model  $Y_i = \beta_1 + \beta_2 N_i + e_i$

## The general F-test: what's going on



- Compares the explanatory power of the base model against the more complex model
- Model 1 excludes phosphate Model 2 has P and  $P^2$
- Model 2 has P and  $P^2$  are correlated so t-tests might break down
  - Both P and  $P^2$  could show as not statistically significant
- The F-test checks whether P and  $P^2$  jointly have an impact
  - We rely on theory / lit review for this kind of checking

## Formally what happens



The model  $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$

Variance formula  $\text{Var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{(1 - \rho_{23}^2) \sum (X_{2i} - \bar{X}_2)^2}$

Variance formula  $\text{Var}(\hat{\beta}_3) = \frac{\hat{\sigma}^2}{(1 - \rho_{32}^2) \sum (X_{3i} - \bar{X}_3)^2}$

So if  $X_2$  and  $X_3$  are correlated the variance increases

## Multicollinearity



The model  $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$

Variance formula  $\text{Var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{(1 - \rho_{23}^2) \sum (X_{2i} - \bar{X}_2)^2}$

Variance formula  $\text{Var}(\hat{\beta}_3) = \frac{\hat{\sigma}^2}{(1 - \rho_{32}^2) \sum (X_{3i} - \bar{X}_3)^2}$

So if  $X_2$  and  $X_3$  are correlated the variance increases

## The implication of higher variance



- Recall the t-test statistic for these tests

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_2)}} = \frac{\hat{\beta}_2}{\sqrt{\frac{\hat{\sigma}^2}{(1 - \rho_{23}^2) \sum (X_{2i} - \bar{X}_2)^2}}}$$

- If  $t$  is small (we have a large p-value) and we do not reject the null that  $\beta_2$  is equal to zero.
- When variables are correlated our t-test can not detect the difference

## Where might this be a problem



- Experiments ?
  - Can control the levels
    - We make sure there is no systematic correlation
- Observational data
  - Hard to implement controls

## Time for a break



## Review dummy variables & interactions



When might we want to use a dummy variable (factor)

- we have two (or more) plant species/ regions to consider (factor)

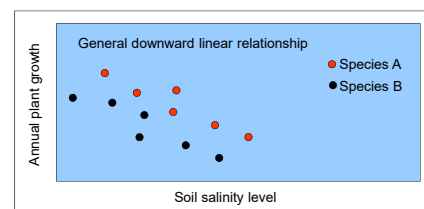
Option (i) estimate each relationship separately

Option (ii) estimate the relationship jointly

Joint estimation raises a general issue

- is there a common intercept to the species, groups, etc ?
- is there a common slope to the species, groups etc ?

## Pooling data



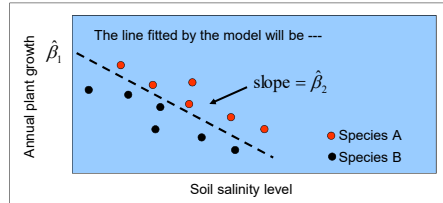
- Looks like a common slope but different intercept
- We only know this because we have plotted by group

## The simple regression model



Fit a simple linear regression model  $G_i = \beta_1 + \beta_2 S_i + e_i$

$G_i$  is annual plant growth, explained by the soil salinity  $S_i$  plus a random error term



## Model assumptions

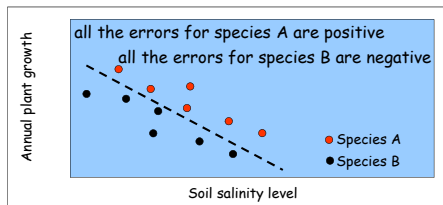


1. For each X value Y is given as  $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + e_i$
2. Expected value of the error:  $E(e_i) = 0$
3. The variance of the error:  $\text{Var}(e_i) = \sigma^2$
4. The covariance of the error:  $\text{Cov}(e_i, e_j) = 0$
5. The  $X$ s can be considered fixed in repeated samples
6. No exact relationships between explanatory variables

## We have a pattern in the error terms



We denote the error  $Y_i - E(Y_i) = e_i = Y_i - \hat{Y}_i$



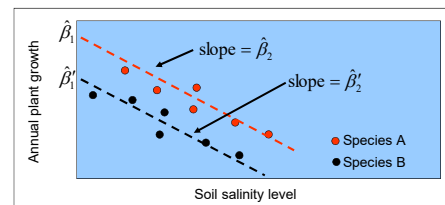
## Option 1. estimate two regressions



A model for Species A  $G_i = \beta_1 + \beta_2 S_i^a + e_i$

and

A model for Species B  $G_i = \beta'_1 + \beta'_2 S_i^b + e_i$



Each model has  $n = 6$  and  $K = 2$ , so its hard to be precise

## Option 2 is the 'dummy' variable



lm1 <- lm(Growth~Salinity +Plant, data = P)

Obs.	( $G_i$ )	( $S_i$ )	( $A_i$ )	( $B_i$ )	( $A_i$ ) + ( $B_i$ )
1	12.1	5.2	1	0	(0+1) = 1
2	18.1	2.2	1	0	(0+1) = 1
3	15.4	4.4	0	1	(0+1) = 1
4	9.4	7.2	0	1	(0+1) = 1
...	...	...	...	...	...
12	2.8	9.9	0	1	(0+1) = 1

-  $X_{i1}$  always equal to one for the intercept

We have  $X_{i1} = A_i + B_i$  for all  $i$  which is perfect multicollinearity

Note that R is formulating our data from words

R will only add  $A_i$  or  $B_i$  not both, here we add A

## Formally what happens



So  $G_i = \beta_1 + \beta_2 S_i + e_i$  becomes  $G_i = \beta_1 + \delta A_i + \beta_2 S_i + e_i$

lm1 <- lm(Growth~Salinity +Plant, data = P)

$$E[G_i] = \hat{G}_i = \begin{cases} (\beta_1 + \delta) + \beta_2 S_i & \text{when } A_i = 1 (B_i = 0) \\ \beta_1 + \beta_2 S_i & \text{when } A_i = 0 (B_i = 1) \end{cases}$$

We get a vertical shift in the intercept

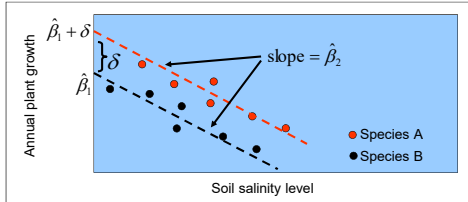
- Species B is in the base (Intercept) term

If the  $\delta$  term is not statistically different from zero the intercept can be treated as equal

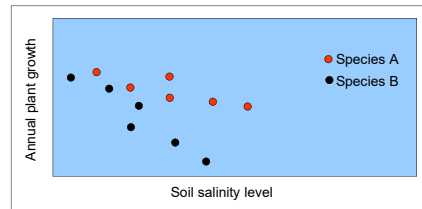
## We get more precision

The model now has  $n = 12$  and  $K=3$  (from 4 dof to 9)  
Readily extends to more groups (factor variable)

Model has a common slope but different intercept for each species



## Common intercept different slope



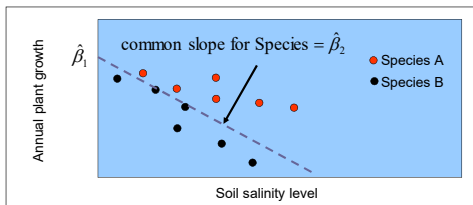
Looks like a common intercept but different slope

If we estimate two regressions we get imprecise estimates

## The simple model problem

$$G_i = \beta_1 + \beta_2 S_i + e_i$$

We again have a problem with the errors. Species A errors are positive Species B errors are negative



## A slight variation of the model

The base model  $G_i = \beta_1 + \beta_2 S_i + e_i$  becomes

$$G_i = \beta_1 + \gamma(A_i \times S_i) + \beta_2 S_i + e_i$$

The new variable  $(A_i \times S_i)$  is the product of soil salinity and the dummy variable for Species A

This kind of variable is called an interaction term or a slope dummy variable

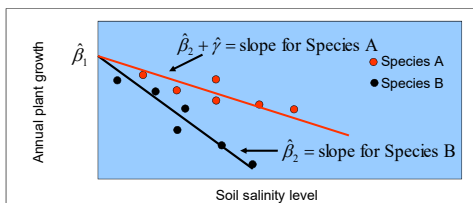
lm1 <- lm(Growth~Salinity +Plant\*Salinity - Plant, data = P)

## In practice what we have

lm1 <- lm(Growth~Salinity +Plant\*Salinity - Plant, data = P)

The regression fits two lines with common intercept but different slopes and we have  $n = 12$  and  $K=3$

- for separate regressions we have  $n = 6$  and  $K=2$



## Formally what we have

lm1 <- lm(Growth~Salinity +Plant\*Salinity - Plant, data = P)

Different slope model  $G_i = \beta_1 + \gamma(A_i \times S_i) + \beta_2 S_i + e_i$

$$\frac{\partial E[G_i]}{\partial S} = \begin{cases} (\beta_2 + \gamma) = \text{slope} & \text{when } A_i = 1 \\ \beta_2 = \text{slope} & \text{when } A_i = 0 \text{ ie } B_i = 1 \end{cases}$$

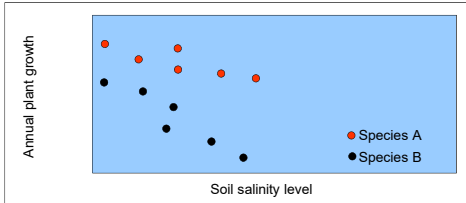
Slope for Species A is  $(\beta_2 + \gamma)$  and Slope for Species B is  $\beta_2$

- if  $\gamma$  is not statistically significant they have a common slope

The model has a common intercept  $\beta_1$

## Vary the slope and intercept

Assume both the intercept and the slope vary with Species  
We can still pool the data



## The pooled model

Simple model  $G_i = \beta_1 + \beta_2 S_i + e_i$  `lm0 <- lm(Growth~Salinity, data = P)`

Allow intercept to vary by Species  $G_i = \beta_1 + \delta A_i + \beta_2 S_i + e_i$   
`lm1 <- lm(Growth~Salinity + Plant, data = P)`

Allow slopes to vary with Species  $G_i = \beta_1 + \gamma(A_i \times S_i) + \beta_2 S_i + e_i$   
`lm2 <- lm(Growth~Salinity + Plant*Salinity, data = P)`

Combine  $G_i = \beta_1 + \delta A_i + \gamma(A_i \times S_i) + \beta_2 S_i + e_i$   
`lm3 <- lm(Growth~Salinity + Plant*Salinity, data = P)`

## The Pooled model

- Is there any advantage to the pooled model?
- Think about degrees of freedom per explanatory variable

## How you get the slope interpretation

## Interpretation of log changes

Traditional formula for calculating percentage changes (use  $P_1$  value as the base)

$$\frac{(101 - 100)}{100} = \frac{1}{100} \times 100 = 1.0\% \quad \frac{(110 - 100)}{100} = \frac{10}{100} \times 100 = 10.0\%$$

What about using the average of  $P_1$  and  $P_2$  as the base

$$\frac{(101 - 100)}{(100+101)/2} = \frac{1}{100.5} \times 100 = 1.00\% \quad \frac{(100 - 110)}{(100+110)/2} = \frac{-10}{105} \times 100 = -9.53\%$$

Log changes multiplied by 100 give % change where the average of  $P_1$  and  $P_2$  is the base

$$(\ln(101) - \ln(100)) \times 100 = 1.00\%$$

$$(\ln(P_2) - \ln(P_1)) \times 100 = \text{a percentage change}$$

$$(\ln(110) - \ln(100)) \times 100 = 9.53\%$$

## Linear-linear model

- No transformation  $Y_i = \alpha + \beta X_i$

$$\text{Slope} = \frac{\text{rise}}{\text{run}} = \frac{\Delta Y}{\Delta X} = \frac{\Delta Y}{1} = \beta$$

$$\Delta Y = \beta \times 1 = \beta = \text{change in } Y$$

A one unit change in the X variable results, on average, in a change in the Y variable  $\beta$  units

## Log-linear model



- Log Y transformation  $\log(Y_i) = \alpha + \beta X_i$

$$\text{Slope} = \frac{\text{rise}}{\text{run}} = \frac{\Delta \log Y}{\Delta X} = \frac{\Delta \log Y}{1} = \beta$$

$$= \Delta \log Y = \beta \times 1 = \beta$$

$$= (\Delta \log Y) \times 100 = \beta \times 100 = \text{percentage change in } Y$$

A one unit change in the X variable results, on average, in a change in the Y variable of  $\beta \times 100$  percent

## Linear-log model



- Log X transformation  $Y_i = \alpha + \beta \log(X_i)$

$$\text{Slope} = \frac{\text{rise}}{\text{run}} = \frac{\Delta Y}{\Delta \log X} = \frac{\Delta Y}{(\Delta \log X) \times 100} = \frac{\beta}{100}$$

$$= \frac{\Delta Y}{1\%} = \frac{\beta}{100}$$

$$= \Delta Y = \beta \times 1\% \div 100 = \text{the change in } Y$$

A one percent change in the X variable results, on average, in a change in the Y variable of  $\beta \div 100$  units

## Log-log model



- log X & Y transformation  $\log(Y_i) = \alpha + \beta \log(X_i)$

$$\text{Slope} = \frac{\text{rise}}{\text{run}} = \frac{\Delta \log Y}{\Delta \log X} = \frac{(\Delta \log Y) \times 100}{(\Delta \log X) \times 100} = \beta \times \frac{100}{100}$$

$$\frac{(\Delta \log Y) \times 100}{1\%} = \beta$$

$$= (\Delta \log Y) \times 100 = \beta \times 1\% = \text{percentage change in } Y$$

A one percent change in the X variable results, on average, in a change in the Y variable of  $\beta$  percent