ANCOVA, transformations and some first thoughts on experimental design

https://xkcd.com/1507/

Lecture 6
Biological statistics III
Ayco Tack

Outline

- ANCOVA
  - *Controlling for a quantitative variable*
- *Model criticism*
  - *Checking assumptions for linear models*
- *Transformations*
  - *Improve agreement with model assumptions*
- *Introduction to experimental design*

ANCOVA

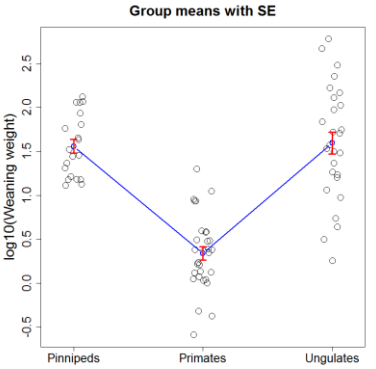
Controlling for the effect of a quantitative variable

A quantitative x-variable is called a **covariate** when it is used to reduce residual variation in a response variable

Example: We want to know if there are differences in offspring weight at weaning among three taxonomic groups of mammals

Graph: The graph shows that primates are smaller at weaning than either ungulates or pinnipeds

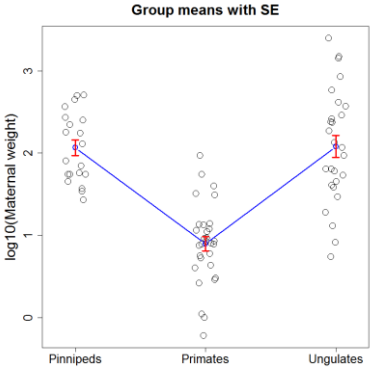
Additional question: Is this simply an effect of primates being small mammals?



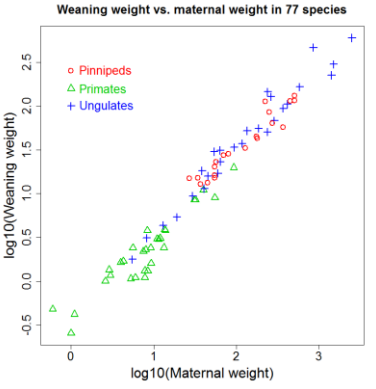
Weaning weights (kg) for 20 pinnipeds, 30 primates and 27 ungulates

Looking for a covariate

Check if maternal size differs among the groups

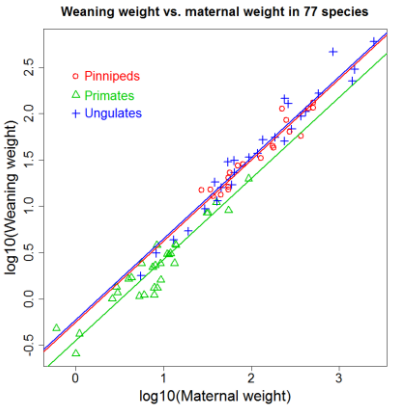


Yes, primates seem to be smaller on average, which could affect weaning weight



Weaning weights (kg) for 20 pinnipeds, 30 primates and 27 ungulates

The resulting ANCOVA



Primates have the smallest intercept. The fitted lines are forced to be parallel so there is only one slope

What to do:

- Test for differences in log weaning weight between the groups, controlling for log maternal weight (the covariate)

ANCOVA for $\log_{10}$ weaning weight

Source	df	SS	MS	F	P
Group	2	0.397	0.199	10.65	**<0.0001**
Covariate	1	17.349	17.349	931.51	**<0.0001**
Error	73	1.360	0.019		
Total	76	19.106			

Conclusions:

- The effect of taxonomic group on weaning weight is significant, even when controlling for maternal weight.
- From the smaller intercept it appears that primate mothers wean their offspring at a smaller relative size, i.e. when one takes into account allometry.
- In addition, there is a highly significant effect of maternal weight on weaning weight: bigger species have bigger babies.

Model criticism for ANCOVA

Before accepting a conclusion from an ANCOVA, one should perform some model criticism. Important things to check for are:

- **Dependence between data points**: As in the other models so far, the data points need to be independent; this is mostly a matter of how the data were collected
- **Normally distributed residuals**: but note that these models are quite robust
- **No relationship between the residuals and fitted values:**
- **Variance heterogeneity**: As we discussed for ANOVA and regression, we want to avoid a higher variance in some groups than in others, or shot-gun patterns or outliers in regressions. In ANCOVA, you may get some insights in the fit by inspecting a figure with the slope for each group, and visually inspect normality across the covariate range for the raw data plot; furthermore, inspect the plot of the residuals against the fitted values (which should look like 'stars at the night sky'). (Don't use Bartlett or Fligner test on the raw data to test for unequal variances among groups, as the level of the covariate may differ among the groups, which affects the variance within each level of the factor.)
- **Non-linear relationships:** As for regressions, check whether the relationship with the covariate is linear (and not curved; and remember we can use a polynomial)

It is also important to test for interactions: You can test for non-parallel regression lines; in fact, you may have a theory why different groups respond differently to the covariate?

In case the model appears unacceptable, there are several possibilities one can consider:

- Variable transformations (e.g., logarithm)
- Non-parametric methods (e.g. function `sm.ancova` in the `sm` package)

Transformations

Purposes of transformations of the *y*-variable:

- **Get homogeneous variances**
- **Make residuals more closely normally distributed**
- **Improve linear fit in regression** (note that one often transforms the *x*-variable as well)

The most common case is variance heterogeneity where the variances are larger in groups with larger means

Situations where transformations might not help:

- The variances are heterogeneous in a way that is not related to the means in a simple manner (for instance, one group is extremely variable)
- The within group distributions are very far from normal, for instance multimodal, or with many zero-values

In that case some other approach than linear models is needed, perhaps non-parametric methods

Common types of transformations

The basic idea is to try to achieve homogeneous variances by making the variance in subsets of the data independent of the mean y-value of the subset. This can also help to 'straighten up' skewed distributions.

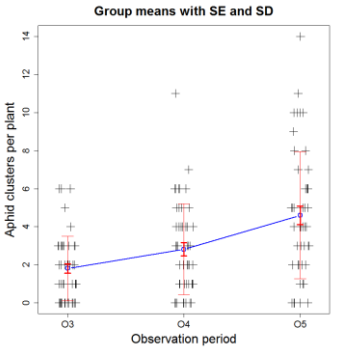
- **Logarithm:** Useful when the standard deviation in subsets of the data is proportional to the mean; traits related to body size are typical examples where the logarithm might be a suitable transformation.
- **Square root:** Useful when the variance in subsets of the data is proportional to the mean; counts of the type that might follow a Poisson distribution are typical candidates for a square root transformation.
- **Inverse:** If either of the two above are insufficient to make the variance independent of the mean, the inverse has a 'stronger kick'; transforming life length (which can have a very right-skewed distribution) to rate of death is an example where the inverse may be suitable.
- **Logit-transformation:** (even though the arcsine transformation, also called the arcsine square root transformation, was often used in the past): Useful for proportions, for which the variance must be small when the mean is close to zero or close to one but can be bigger in between these extremes.

For the logarithm, the square root and the inverse, if the data contain zeroes one can add a small number, for instance 0.5, to each data point before transforming. In R *we can use the function* `boxcox` *in package* `MASS` *to find a suitable transformation*

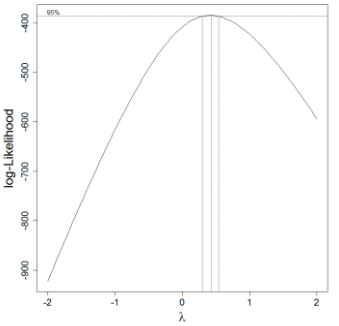
Example: transforming aphid cluster data

The number of *Aphis fabae* clusters were counted on 48 potted plants, *Matricaria perforata*, at 3 different points in time

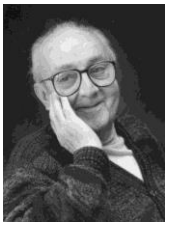
Let us use the `boxcox` function to look for a transformation that makes the variances homogeneous (we first add 0.5 to the *y*-variable, to avoid zeros)



The variance is larger in `O4` and `O5` and the Fligner-Killeen test rejects variance homogeneity ($p = 0.001$)

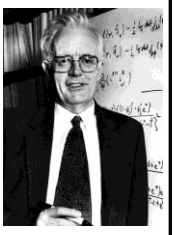


`sqrt(NoClust + 0.1)` might be an acceptable transformation



Box-Cox power transformations

The statisticians George Box and David Cox developed a method to estimate the exponent (lambda λ) to which we can raise the data to reach a normal distribution.



- The lambda gives the power to which we should raise all values
- The box-cox function searches within the area $\lambda = -5$ to $\lambda = 5$
- Note that for $\lambda = 0$, the transformation is not Y^0 (which would be 1 for every value) but instead the logarithm of *Y*
- May not always work -> check probability plot for normal distribution
- the Box-Cox Power transformation only works if all the data is positive and greater than 0.

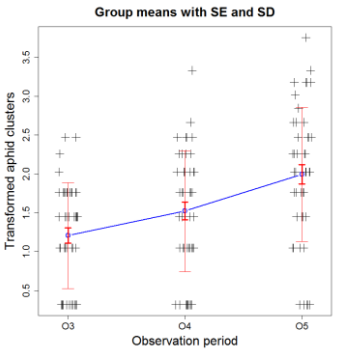
λ	Y'
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\text{Sqrt}(Y))$
0	log(Y)
0.5	$Y^{0.5} = \text{Sqrt}(Y)$
1	$Y^1 = Y$
2	Y^2

Example: transforming aphid cluster data

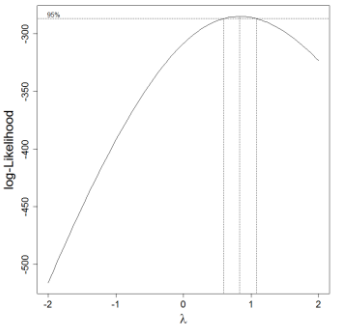
Try the square root transformation. In this case:

$$y_{transf} = \sqrt{y + 0.1}$$

We can also use `boxcox` again to see if the transformation achieved variance homogeneity (Just for illustration: this is not a method for testing this!)



This looks somewhat better and the Fligner-Killeen test no longer rejects variance homogeneity ($p = 0.26$)



Because $\lambda = 1$ is included in the 95% confidence interval in the plot, we conclude that the transformation is OK

Analysis of transformed data

Test: We now test the null hypothesis that the mean number of clusters is the same in the different observation periods

Source	df	SS	MS	F	P
Period	2	14.995	7.497	12.38	<0.0001
Error	141	85.411	0.606		
Total	143	100.406			

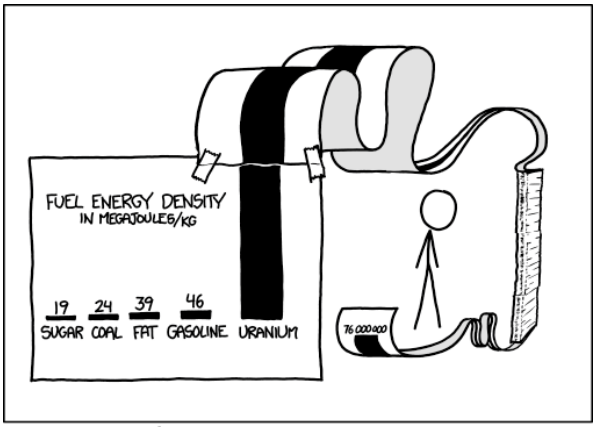
Conclusion: We reject H_0 and conclude that the number of clusters varies between observation periods

Tukey HSD test: *P*-values for pair-wise comparisons

Between O3 and O4	Between O3 and O5	Between O4 and O5
0.12	**<0.0001**	**0.010**

Conclusion: More aphid clusters per plant in O5 than in the other two periods, but there is no significant difference between O3 and O4

https://xkcd.com/1162/



(Most likely) Some wise words on interpreting and misinterpreting log-transformation:
https://www.nature.com/articles/s41559-018-0610-7

Recap of statistical power

Type 1 error:

- Reject a true H_0 (false positive)
- $\Pr[type\ I\ error] = \alpha$
- α is called the level of the test

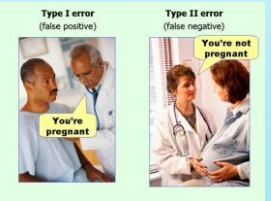
Type 2 error:

- Accept a false H_o (false negative)
- $\Pr[type\ II\ error] = \beta$
- $Power = 1 - \beta = \Pr[reject\ false\ H_0]$

(Quinn & Keough)

POPULATION SITUATION	STATISTICAL CONCLUSION: Reject H_0	STATISTICAL CONCLUSION: Retain H_0
Effect	Correct decision: Effect detected	Type II error: Effect not detected
No effect	Type I error: Effect detected; none exists	Correct decision: No effect detected; none exists

- Power is the probability of actually getting a significance given that some alternative hypothesis is true
- The power of a test depends on which particular alternative hypothesis that one considers
- It is bad practice to do tests with too little power (say, less than 50%)
- A good recommendation is to have at least 80% power



Sample size, detectable difference and power

Questions we might ask when comparing groups

- Which sample size is needed to detect a certain difference?
- How big a difference can we detect with our sample size?
  - Perhaps we can have an idea of the within-group standard deviation s
  - The standard error of the group mean is around $s/\sqrt{n}$
  - We are very likely to detect a difference that is about five times the standard error of a group mean
- What is the probability (power) to detect a certain difference?
  - We need a software package for this

It is a good idea to think about these questions while planning a study

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."
Sir Ronald Aylmer Fisher

Playing safe

Some insurance policies

- Conduct a study that is of interest regardless of results
- Look (also) for some "quick and easy" results
- Investigate (also) the strongest or most evident aspects of the phenomenon

Scientific strategies

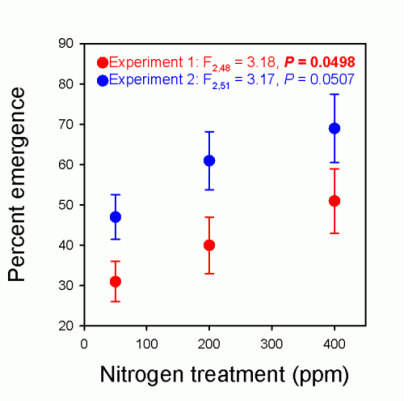
Some possible attitudes

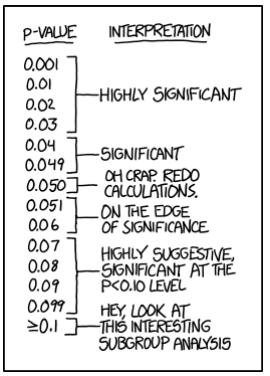
- Only go for long shots (gambler)
- Carry on blindly, hoping that things will somehow work out in the end (optimistic believer)
- Refuse to try something unless there is an absolute guarantee that it will succeed (control freak)
- Willing to try new things, but only if there is a realistic chance of success (mature scientist?)

When to give up?

- Concorde fallacy
- Is failure a possibility?

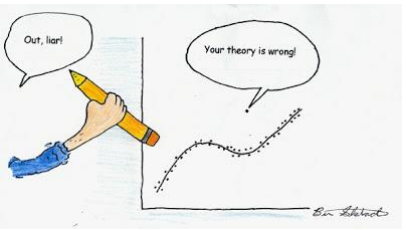
How to interpret p-values?





https://scientistseessquirrel.wordpress.com/2015/11/16/is-nearly-significant-ridiculous/
https://www.explainxkcd.com/wiki/index.php/1478:_P-Values
https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/
https://scientistseessquirrel.wordpress.com/2018/12/04/15th-century-technology-and-our-disdain-for-nearly-significant/

Related reading and information



- **Quinn & Keough:** Chapter 12.1 – 12.3
- **Crawley:** Sections 12.1, 9.11, p. 629

www.xkcd.com