Thinking about statistics for radiation effects using some examples

Prof. Marian Scott

University of Glasgow

“The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.”

ASA statement on p-values, 2016

“Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting..”

ASA statement on p-values, 2016

General design pointers

1. Statistical inference (association or causal)

2. Statistical significance (real world relevance or not)

Statistical inference (association or causal)

Inference- generalising from a sample to a population-”Learn about what you do not observe (parameters) from what you do observe (data)”

“Correlation does not imply causation, and yet causal conclusions drawn from a carefully designed experiment are often valid. What can a statistical model say about causation?” (Holland, 1986).

There are specific models for causal inference.

Statistical inference (association or causal)

“The aim of standard statistical analysis, typified by regression, is to infer parameters of a distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer aspects of the data generation process.” (Pearl, 2003)

General design pointers

1. Statistical inference (association or causal)

The statistical modelling commonly reported is not causal.

General design pointers

2. Statistical significance (real world relevance or not)

A p-value helps determine statistical significance (i.e. the effect observed is unlikely to have occurred by chance).

It does not translate always to real world importance.

We should also be conscious of the power of the experiment of survey (typically related to sample size)

3. Independence or dependence

The nature of the statistical model will vary whether you have independent or dependent observations.

Classical forms of dependence include temporal dependence (ie observations made close in time are more similar than observations far apart in time) Spatial dependence (observations close in space are more similar than if they were far apart)

Ignore dependence at your peril

4. Nature of the variables

Counts- discrete
Presence/absence- binary
Continuous

All require different statistical models (although they can be described often in a general framework- GLM (generalised linear model)

5. Nature of a statistical model

- **Outcomes or Responses**
  these are the results of the survey or experiment and are sometimes referred to as '***dependent variables***'.
- **Causes or Explanations**
  these are the conditions or environment within which the ***outcomes*** or ***responses*** have been observed and are sometimes referred to as '***independent variables***' or covariates.

Statistical Models

covariates' may be aspects that the experimenter has no control over but that are relevant to the ***outcomes*** or ***responses***. In observational studies, these are usually not under the control of the experimenter but are recorded as possible explanations of the ***outcomes*** or ***responses***.

Every statistical model makes assumptions which you need to check.

A covariate may be continuous or categorical (often then called a factor)

Definitions

- A *factor* is a discrete variable used to classify experimental units. For example, ”Gender” might be a factor with two levels “male” and “female” and “Diet” might be a factor with three levels “low”, “medium” and “high” protein. The levels within each factor can be discrete, such as “Drug A” and “Drug B”, or they may be quantitative such as 0, 10, 20 and 30 mg/kg.
- Fixed effects factors, are variables which can be controlled by the investigator. These include gender, dose, diet, and any treatment which can be administered to the animals. Most experiments are designed to study the fixed effects.

Definitions

- **Fixed Effect:**
  a factor, levels of which are chosen in a non-random way and therefore cannot be expected to represent the population as a whole.
- **Random Effect:**
  a factor, levels of which are chosen randomly from all the possible levels and the levels used can be expected to represent the population as a whole; random effects are used where individuals are measured multiple times or there is nested sub-sampling.

- A model combining fixed and random effects is called a mixed effects model.

Sample Size

Greater numbers of experimental units will result in:

- more evidence for your conclusions
- more powerful tests
- narrower confidence intervals
- an better opportunity to test assumptions

Guard against :-

- pseudo-replication,
  ignoring correlations
- Increasing some treatment combinations more than others

Sample sizes should be determined from a knowledge of the variability of the material that one is working with and the size of the effects that one is looking for.

linear models

- For the linear model, we write the response as a linear function of the explanatory variables.
- For the simplest case (linear regression) we write
  - $Y = \beta_0 + \beta_1 x + \varepsilon$

- β_0 and β_1 represents the unknown parameters, which we estimate. x is typically a fixed effect
- Output will be estimated coefficients for the slope and intercept and their standard errors,
- ε we assume has zero mean and typically has a Normal distribution
- If we have n observations (y_i, x_i)we commonly assume that they are independent

Generalised linear models (GLM)

- When we make different distributional assumptions, then we need to extend the simple linear model- this is done by changing how we link the response to the explanatory variables using a link function. For the simple linear model, the link function is the identity function
- For Y ~Poisson, so counts data, the link function is the log
- For Y~ Binomial, we use logit link function

- GLMs can be simply fit in the appropriate software, you typically need to identify the distribution that your response comes from and also the link function that you will use. The fitting is more complex but will be invisible to you.
- Output will still be estimated coefficients and standard errors,

Generalised linear models

- **Generalised** linear models are designed to deal with the situation that our response variable is not Normally distributed and may not even be continuous.
- A general linear model is one that is linear in the parameters
  - $y=\beta_0 +\beta_1 x + \varepsilon$
- A generalised linear model has a linear predictor

$$\eta=\beta_0 +\beta_1 x$$

and a link function that describes how the E(Y) = μ depends on η

$$g(\mu_i) = \eta_i$$

Generalised linear **mixed** models, (GLMM)

The basic regression models typically assume that observations are independent but we often have structure that introduces dependence which we need to deal with.

This is often done by introducing some random effects into the models.

Some examples follow

GLMM: a practical guide for ecology and evolution Bolker et al, 2009

“problems often involve random effects, whose purpose is instead to quantify the variation among units. The most familiar types of random effect are the blocks in experiments or observational studies that are replicated across sites or times.

Random effects also encompass variation among individuals (when multiple responses are measured per individual, such as survival of multiple offspring or sex ratios of multiple broods), genotypes, species and regions or time periods”.

GLMM: a practical guide for ecology and evolution Bolker et al, 2009

- **Fixed Effect:**
a factor, levels of which are chosen in a non-random way and therefore cannot be expected to represent the population as a whole.

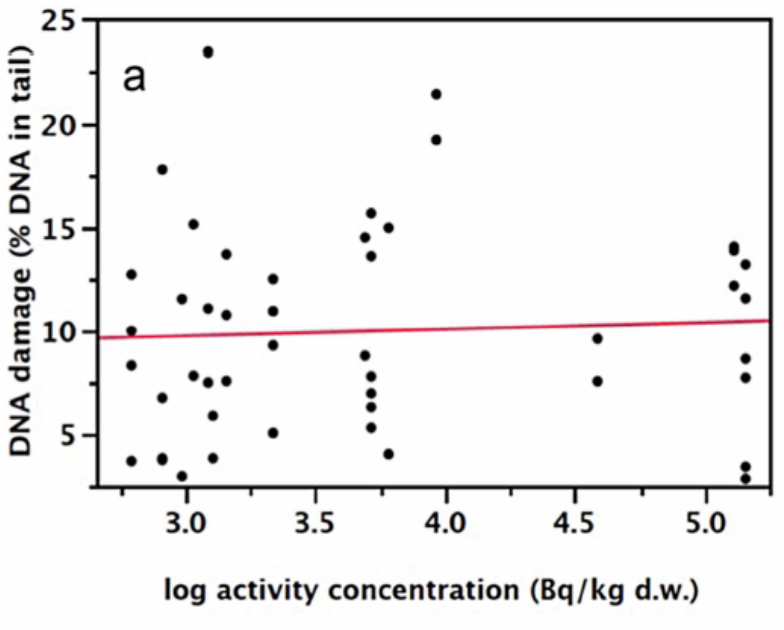
- **Random Effect:**
a factor, levels of which are chosen randomly from all the possible levels and the levels used thus can be expected to represent the population as a whole; such effects are frequently found in experiments where individuals are measured multiple times or there is nested sub-sampling.

Linear mixed models

- The challenge of defining the model in the notation of the software being used. (the fixed part of the model is straightforward).
- You can have more than one random effect

- Output includes parameter estimates for the fixed effects and a variance estimate for the random effect.

•Abundance and genetic damage of barn swallows from Fukushima**(2015) Bonisoli Alquati et al**



•we attached (TLDs) to the inner and outer rim of 55 barn swallow nests from the Fukushima region . From 62 chicks from 16 nests, we also collected a small blood sample

This study would include a random effect for nest (since it is a common source of variation for the chicks).

Long term census data reveal abundant wildlife populations at Chernobyl. Deryabina et al. Current Biology *25*, 2015

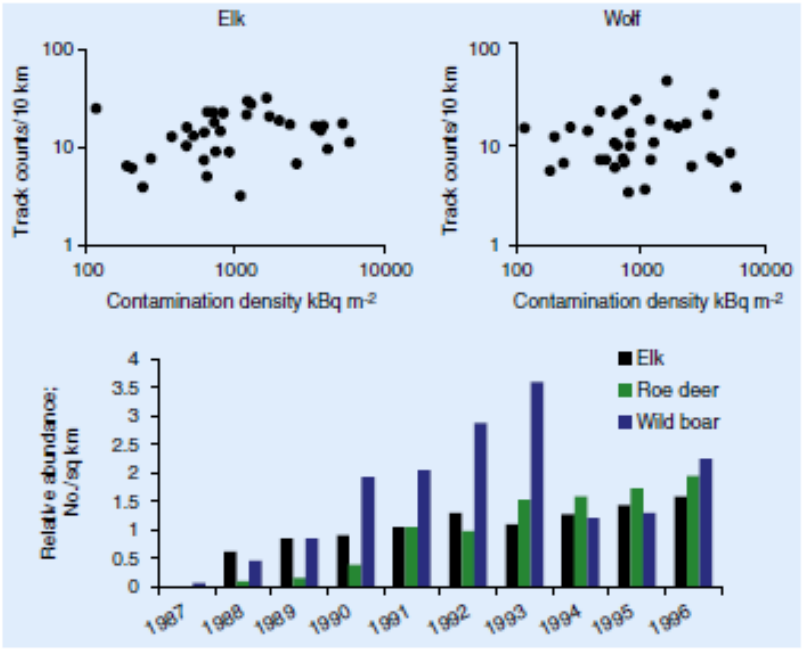
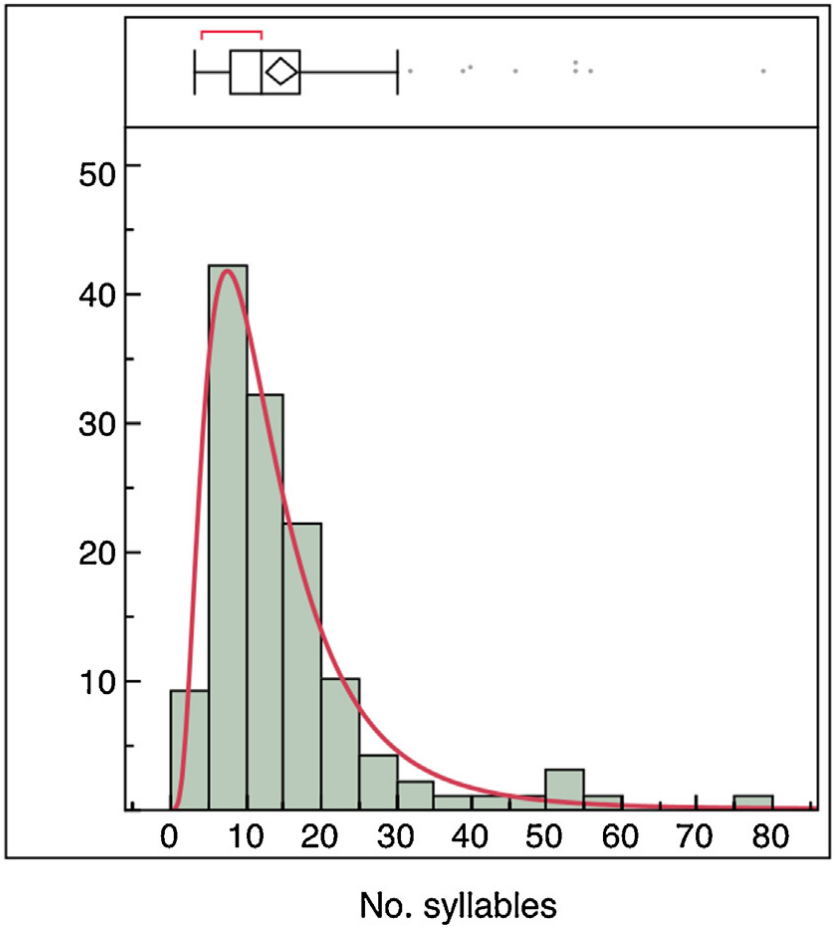


Figure 1. Animal abundances in the Chernobyl exclusion zone.
(Top) Mean number of track counts per 10 km (in 2008–2010) for elk and wolf plotted against mean ^{137}Cs contamination density of each route. Analysis using linear mixed models including habitat variables found no evidence of correlation between mammal density and contamination density. See Supplemental Information for other species studied. (Bottom) Change in relative abundance of three species in the 10 years after the Chernobyl accident.

linear mixed models with repeated measures

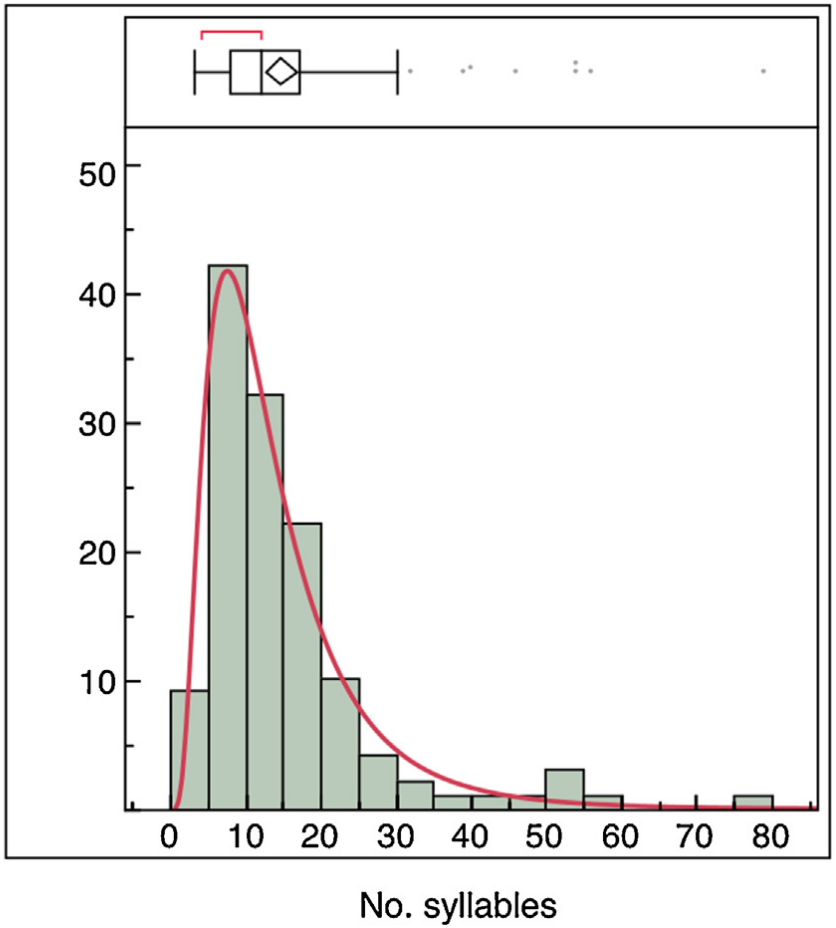
Mixed models- we have fixed and random effects, repeated measures- observations are not independent, so repeated measures over time or space. The random effect is used to capture this, and we also need to specify the covariance or dependence

The number of syllables in Chernobyl cuckoo calls reliably indicate habitat, soil and radiation levels. Anders Pape Møller, Federico Morelli, Timothy A. Mousseau, Piotr Tryjanowski (2016)



Mixed effects models refer to a variety of models, which have as a key feature both fixed and random effects. In this study, Generalized Linear Models were used to study variation in the number of syllables in common cuckoo call in relation to environmental variables (habitat and soil type) modeled as fixed effects

The number of syllables in Chernobyl cuckoo calls reliably indicate habitat, soil and radiation levels. Anders Pape Møller, Federico Morelli, Timothy A. Mousseau, Piotr Tryjanowski (2016)



•Statistical analysis. For the analysis of genetic integrity of nestlings we used general linear mixed models (GLMMs) where we included radiation exposure (either logtransformed radioactivity of the nest material or radiation dose as inferred from the TLDs) as a covariate, and the nest of origin as a random effect. In both analyses we included duration of exposure as a covariate.

Fitness costs of increased cataract frequency and cumulative radiation dose in natural mammalian populations from Chernobyl. Lehmann et al. Nature Scientific reports, 2015

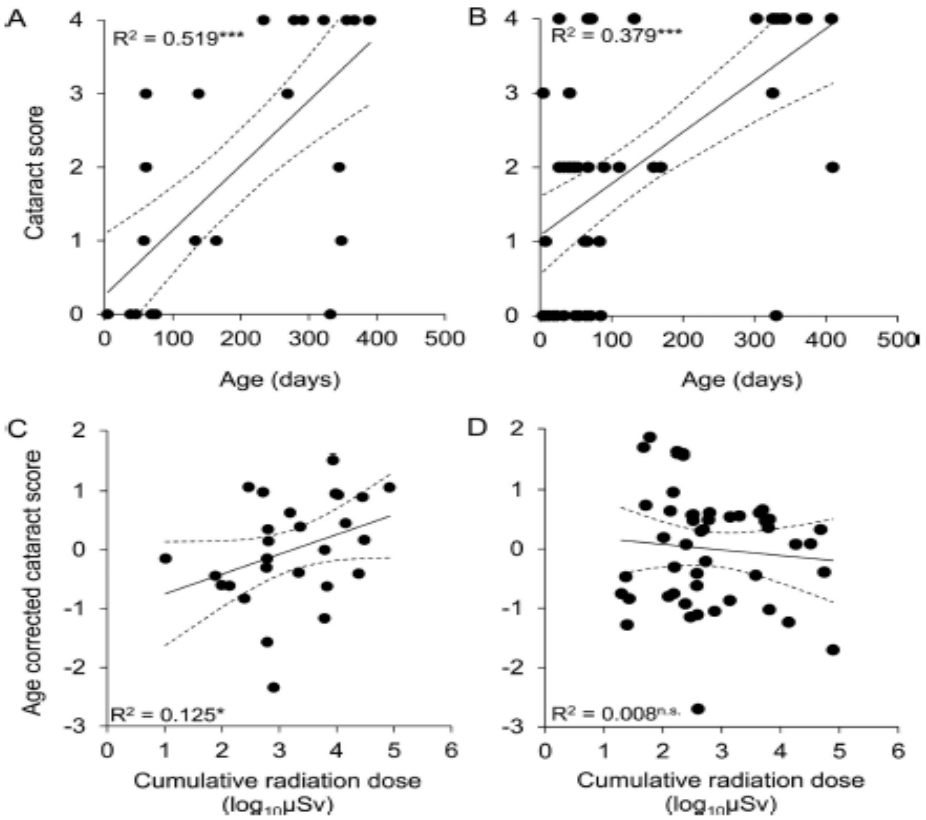


Figure 1. The upper panels show cataract scores regressed against the age estimate for (A) female and **(B)** male bank voles collected from Chernobyl. The lower panel shows cataract scores corrected for age (standardized residuals from a generalized linear regression on the data split by sex), regressed against the logarithm of lifetime accumulated radiation dose, for **(C)** female and **(D)** male bank voles. The symbol after the R^2 value denotes the significance level of the regression (n.s. = $P > 0.05$; *$P < 0.05$; **$P < 0.005$; ***$P < 0.001$). Dashed lines refer to the 95% confidence interval limits of the regression (solid) line.

a generalized linear mixed model with cataract score as dependent variable, sex as factor and age (in hours) and the logarithm of accumulated radiation dose (log10µ Sv) as covariates. The main effects as well as the interactions age*radiation and sex*radiation were included. Collection year and location were added as random (block) factors.

Fitness costs of increased cataract frequency and cumulative radiation dose in natural mammalian populations from Chernobyl. Lehmann et al. Nature Scientific reports, 2015

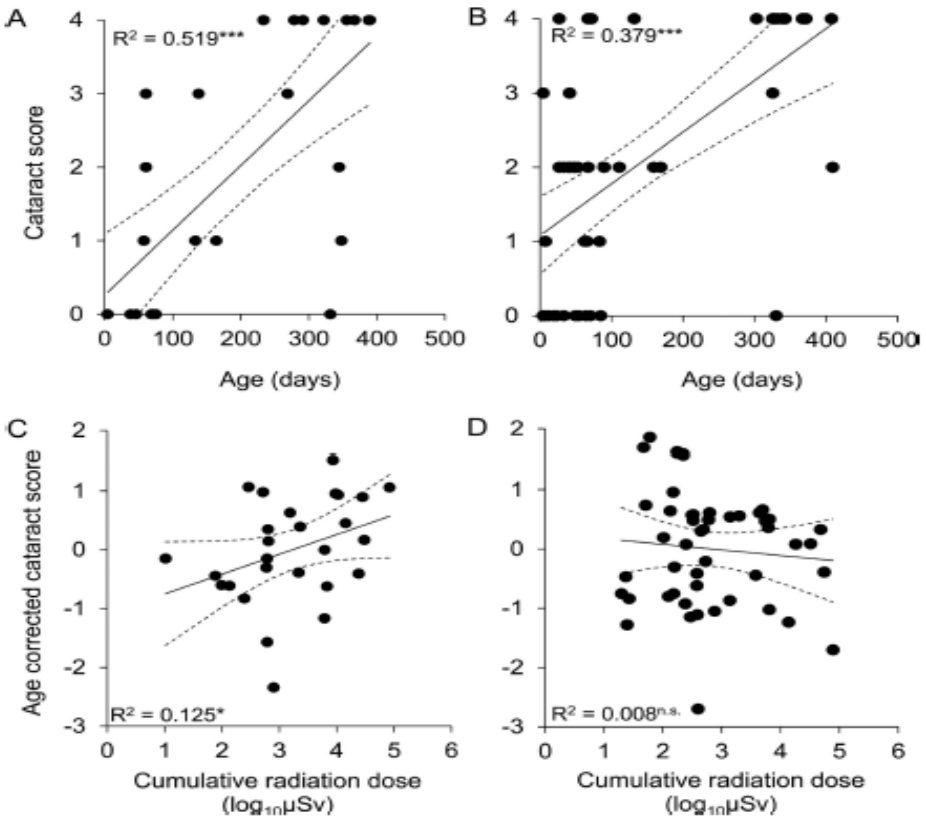


Figure 1. The upper panels show cataract scores regressed against the age estimate for (A) female and **(B)** male bank voles collected from Chernobyl. The lower panel shows cataract scores corrected for age (standardized residuals from a generalized linear regression on the data split by sex), regressed against the logarithm of lifetime accumulated radiation dose, for (C) female and (D) male bank voles. The symbol after the R^2 value denotes the significance level of the regression (n.s. = $P > 0.05$; *$P < 0.05$; **$P < 0.005$; ***$P < 0.001$). Dashed lines refer to the 95% confidence interval limits of the regression (solid) line.

a generalized linear mixed model with cataract score as dependent variable, sex as factor and age (in hours) and the logarithm of accumulated radiation dose (log10µ Sv) as covariates.

Generalised- because the distribution of the response variable is not normal

Mixed because repeated measurements in location

Choosing which explanatory variables to include, and interaction terms

- when fitting a linear model (or indeed any model), then we need to be careful in our selection of explanatory variables and also how correlated the explanatory variables are (known as multi-collinearity). If there are high correlations amongst the explanatory variables, then this can cause problems in the model fitting and testing.

- There are several approaches to model building, one common one is to start with the most general (ie. full model including all potential explanatory variables) and remove terms as they prove to be statistically non- significant.

Choosing which explanatory variables to include, and interaction terms

- Interaction terms in models are used to describe the joint (common) effects of two or more explanatory variables.

- They are most commonly used in designed experiments eg with two fixed effect factors, we can imagine an interaction between the two, but they can also be used with a random effect.

- Statistical convention tends to be that is an interaction term is significant, then the main effects of the factors are not discussed.
- If the interaction term is not significant then it is removed and the model re-fitted.

Some basic regression features – outliers and influence

- when fitting a linear model (or indeed any model), then we need to be mindful of a) outliers and b) observations with considerable (excessive) influence on the parameter estimates

- What is an outlier?
- In general it is an unusual observation, ie one that seems unlike other data values in the data set (very big, very small), and 'far away' from the main bulk of the data

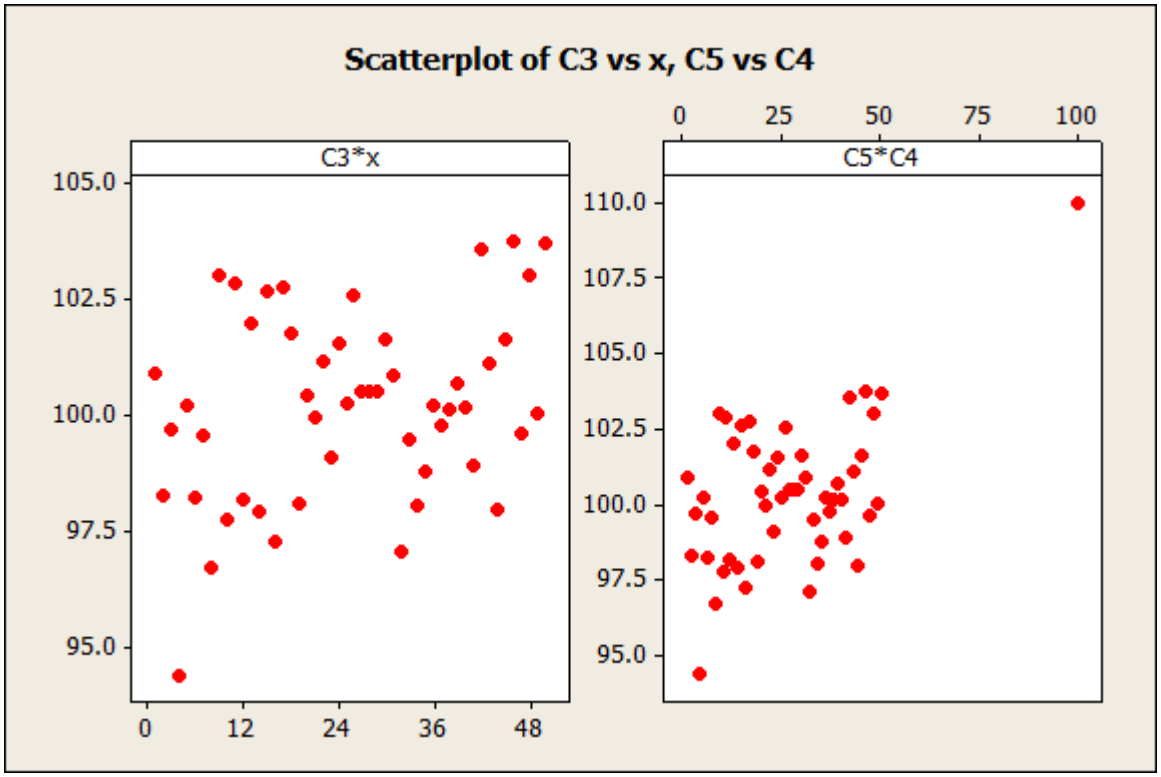
- With regard to a model, an outlier would be a data value that lies 'far away' from its modelled value.

Some basic regression features – outliers and influence

- when fitting a linear model (or indeed any model), then we need to be mindful of observations with considerable (excessive) influence on the parameter estimates

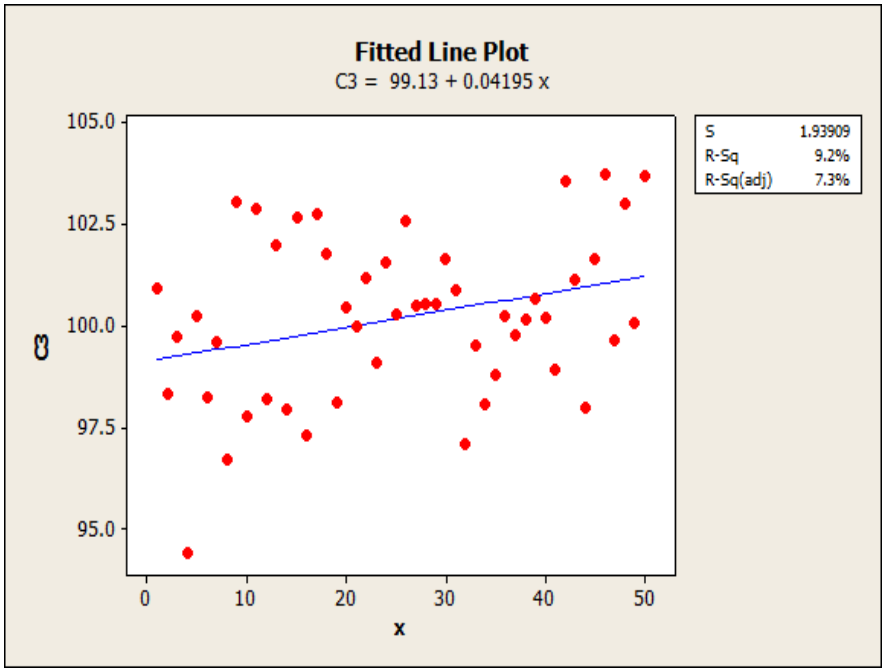
- What is an influential observation?
- In general it is an observation, which contributes strongly to a parameter estimate (it does not need to be an outlier)

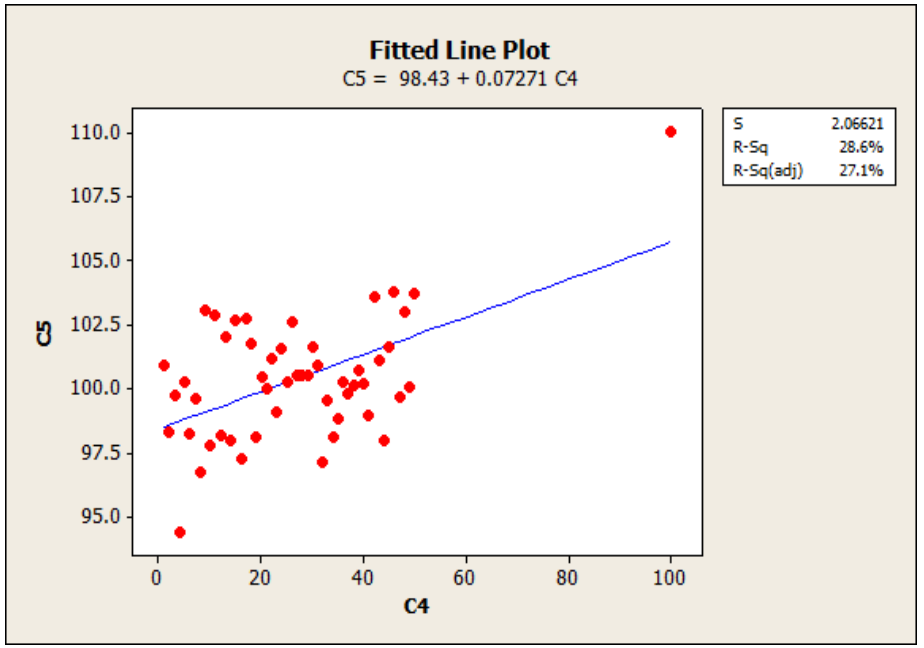
- A classic example in a straight line fitting is an x-value (the explanatory variable) which is extreme



•An influential observation added to the data in the scatterplot on the left is shown on the right.

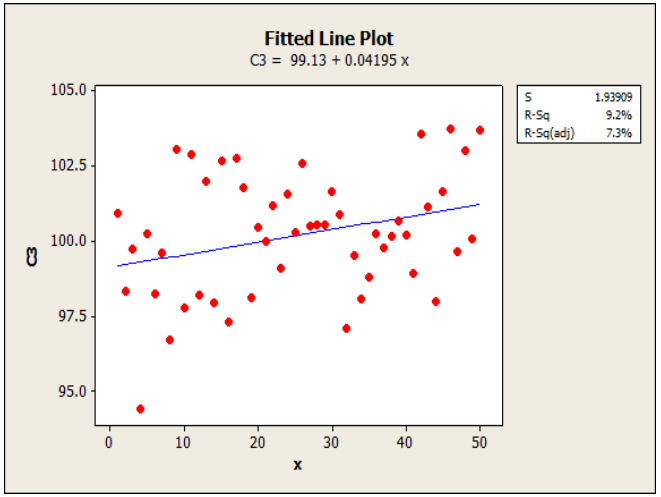
•Is there a relationship- is it statistically significant ?





- An influential observation added to the data in the scatterplot on the left is shown on the right.

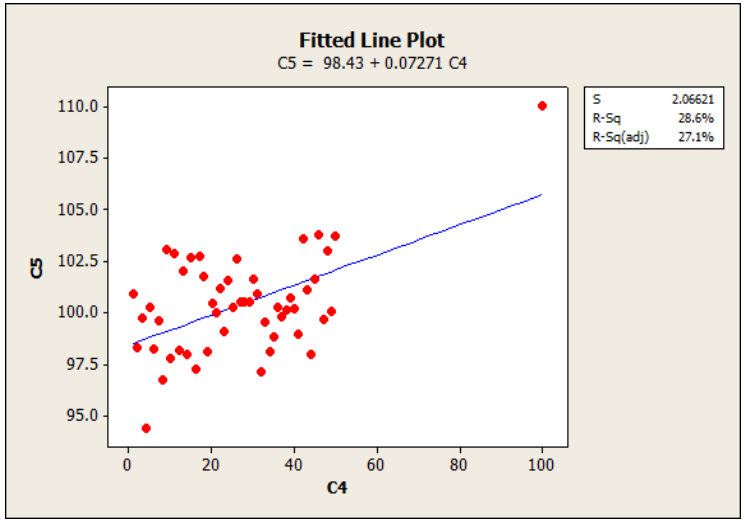
- Is there a relationship- is it statistically significant ?



- The regression equation is
- y = 99.1 + 0.0419 x
- The slope is 0.0419 with a standard error of 0.01900 and the p-value is 0.032

- R-Sq = 9.2% R-Sq(adj) = 7.3%

- So statistically significant, very little variation in y explained (7%), and the data were randomly generated with no relationship.



- The regression equation is
- y= 98.4 + 0.0727 x

- The slope is 0.07271 with se 0.01641 and p-value 0.000

- R-Sq = 28.6% R-Sq(adj) = 27.1%

- So statistically significant, more variation in y explained (27%), and the slope has changed from 0.0419 to 0.0727
- But remember, these data were simulated with no relationship between the y and x variables.

- Message to take home

- Be careful of unusual and influential observations

- Yes pay attention to statistical significance but also to model goodness of fit and assumptions.

- From simple linear model (R^2) (% variation of the response explained by the explanatory variable), to more complex measures of model fit (deviance, AIC)), these capture how well the model explains (the goodness of fit)

- Another issue- error in variables

- In most LM and GLM, we assume that the x-values are known (precisely, ie without uncertainty) but
- In many situations, where radiation dose in the x-variable, this is an assumption that is most likely not valid.

- Does it matter?
- Yes, is the simple answer, the parameter estimates if we ignore this effect will be biased.

Abundance and genetic damage of barn swallows from Fukushima**.** Bonisoli Alquati et al. Nature Scientific reports, 2015

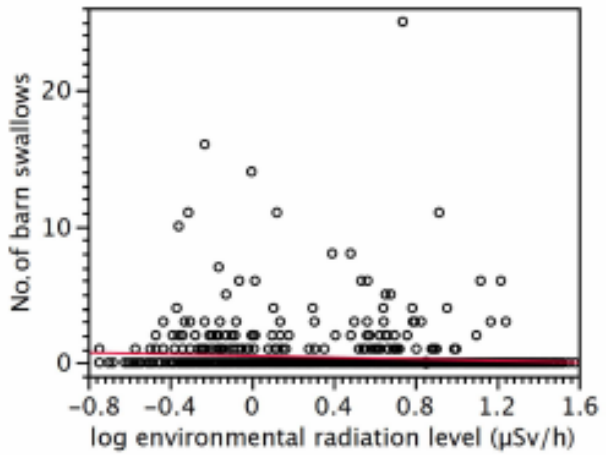


Figure 2 | Barn swallows abundance and radioactive contamination. The abundance of barn swallows declined with increasing levels of radioactive contamination as measured during our multi-year point-count census.

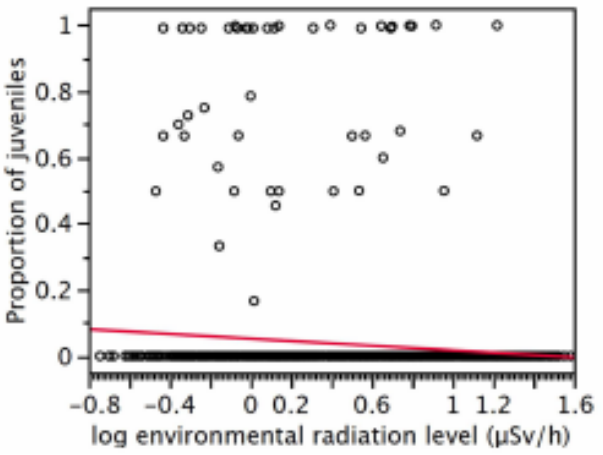


Figure 3 | Age ratio of barn swallows and radioactive contamination. The proportion of barn swallows being juveniles declined with increasing levels of radioactive contamination.

general linear mixed models (GLMMs) where we included radiation exposure (either log transformed radioactivity of the nest material) as a covariate, and the nest of origin as a random effect.

Number of swallows- count, so Poisson, and hence a generalised linear model, nest is random, since expect there to be variation across nests, and the nests represent the population of nests that could be sampled.

Some regression features-Poisson regression and modelling counts

- The GLM framework for modelling counts data uses the Poisson distribution

- Two features are important- theoretically for the Poisson the mean and variance are identical, so this should also hold roughly within the data and the number of zeros observed.

- If either there are too many zeros (zero-inflated) or the variance is larger than the mean (variance inflated) then the fitting needs to be adjusted to accommodate these

Some regression features-Poisson regression and modelling counts

- Zero- inflated: the model takes two parts, models the excess zeros (as a binary), models the counts (as a Poisson)
- A more complex model, that we can evaluate compared to the simpler poisson regression model (again informally check what % of observations would be zero if the Poisson model assumptions held)

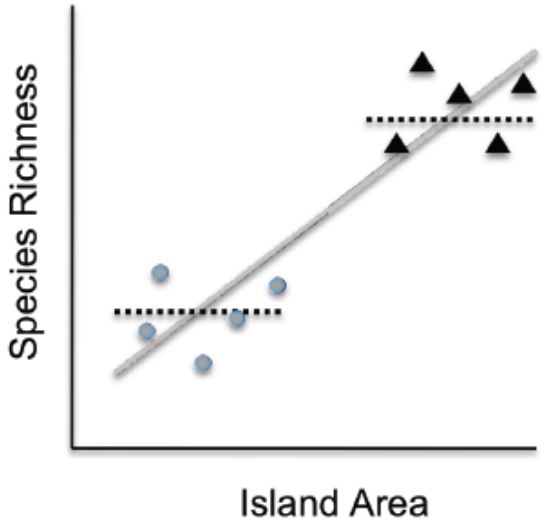
- Variance inflated, check the ratio of the sample mean and variance (is it approx 1?)

FORUM:Island, archipelago and taxon effects: mixed models as a means of dealing with the imperfect design of nature's experiments. Bunnefeld and Phillimore, 2012

- Hypothetical statistical model that aims to address island area as a predictor of the species richness of ten distinct taxa across all of the islands constituting ten different archipelagos. In this case there are at least three sources of non independence, namely island, archipelago and taxon,

- If we deal with the data (repeated measurements from each island, more than one island per archipelago) as independent, then we have fallen into the trap of pseudo replication

Figure 1. A hypothetical scenario illustrating how archipelago or taxon effects can mislead. In this figure there is no underlying relationship (dotted line) between island area and species richness but data points from the same archipelago (or taxon) tend to be more similar than those from different archipelagos (or taxa). In this case the slope estimated using a linear model (grey solid line) incorrectly identifies a positive relationship across all islands, which is due to differences between archipelagos.



FORUM:Island, archipelago and taxon effects: mixed models as a means of dealing with the imperfect design of nature's experiments. Bunnefeld and Phillimore, 2012

Linear mixed models

- Random effects describe the grouping (e.g. taxon) or the hierarchical structure(data points within islands within archipelagos) within the data.

- We estimate a single parameter, the variance across levels of the random effect.

Linear mixed models

- So we have species richness, S as the response, but we recognise the structure- a taxon p, on an island i, within an archipelago g
- All taxa on an island have a common source of variability (the island)
- All islands in an archipelago have a common source of variability (the archipelago)
- Model description S_{igp} = main effect+ island (random)+ archipelago (random) +taxon (random) + error
- For each random effect, we typically assume a Normal distribution, with a variance σ^2

references

Bunnefeld N, Phillimore A (2013) FORUM: Island, archipelago and taxon‘ effects: mixed models as a means of dealing with the imperfect design of nature’s experiments. Ecography.

Bonisoli Alquati et al. (2015). Abundance and genetic damage of barn swallows from Fukushima. Nature Scientific reports,

Wasserstein R L& Lazar N A (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. The American Statistician

Holland P (1986). Statistics and Causal Inference. JASA

Pearl J (2003). Statistics and Causal Inference: A Review. Test

Bolker et al (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Cell

Deryabina et al (2015). Long-term census data reveal abundant wildlife populations at Chernobyl. Current Biology

Lehmann et al. (2015). Fitness costs of increased cataract frequency and cumulative radiation dose in natural mammalian populations from Chernobyl Nature Scientific reports.

Moller et al (2016). The number of syllables in Chernobyl cuckoo calls reliably indicate habitat, soil and radiation levels. Ecological Indicators