Short case study 3:
Three non-identifiable models, two of which are vital in biology

What is non-identifiability?

There are many kinds of non-identifiable models in statistics, and many methods that have been developed to spot them, but here we define non-identifiability as "when all parameters in a model cannot be uniquely identified with infinite data". A classic case of non-identifiability could be defined as:

$y \sim normal(\alpha \cdot \beta, \sigma^2)$

Where we are attempting to estimate three parameters (α, β, and σ) in order to model a continuous response variable (y). There are multiple combinations of α and β that could lead to the same prediction of y (e.g., compare the predictions of $\alpha = 1$ and $\beta = 1$ vs. $\alpha = 2$ and $\beta = 0.5$) and so α and β are non-identifiable.

These kinds of non-identifiability would most likely be caught by Step 2 of our approach: the ability of a given model to detect such parameters would be so poor it would warrant further investigation. However, many other off-the-shelf statistical tools detect such issues (and, indeed, you cannot easily fit a model such as described above using tools such as `R`‘s `lm`). So what about cases of non-identifiability that 'standard' tools can't detect, but that our workflow can?

Non-identifiability in regression models

Perhaps the most commonly observed form of non-identifiability comes from correlation or redundancy among explanatory variables. While such cases are often straightforward to identify when the variable are linear combinations of each other, if they are not linear combinations, or if sufficient error is introduced that the correlation is not obvious, it can be difficult to spot.

Consider a case where there are three variables, each of which sequentially affects the other. This is a common consideration in path-analysis/structural-equation modelling type scenarios, and can be simulated by drawing a variable (`a`), making another explanatory variable that depends upon `a` (`b`), and then finally simulating a response variable that depends on both `a` and `b`:

```r
set.seed(123456)
a <- rnorm(1000)
b <- a + rnorm(1000, sd=.01)
y <- a + b + rnorm(1000)
```

This is not an unusual situation to occur in the real-world: anyone who studies plants will be concerned about those plants' environmental humidity and temperature, two variables whose links are at least as complicated as the above. While we are using somewhat absurdly correlated data here, importantly *a standard regression model does not flag any issues to the user* when it is fit to these data:

```r
summary(lm(y ~ a + b))
```

```
##
## Call:
## lm(formula = y ~ a + b)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.0056 -0.7173 -0.0004  0.6173  3.4362
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.05125    0.03189   1.607    0.108
## a           -1.18560    3.13248  -0.378    0.705
## b            3.21517    3.13218   1.026    0.305
##
## Residual standard error: 1.008 on 997 degrees of freedom
## Multiple R-squared:  0.7999, Adjusted R-squared:  0.7995
## F-statistic:  1992 on 2 and 997 DF,  p-value: < 2.2e-16
```

Where the user would be left under the impression that neither `a` nor `b` were notable correlates of `y`. There is non-identifiability in the sense that the statistical model is unable to distinguish or identify the true values of the coefficients in the model. Because the model cannot distinguish between the impact of either `a` or `b`, it assigns no importance to either.

Properly data exploration would hopefully detect extreme cases of this problem (see also our 'lasso' example), but where the issue is inherent to the model or the correlation is not as profound as in this demonstration, Steps 2 (checking our model with simulated data, where this non-identifiability would be noticed) and 4 (checking the impact of the parameters of the model) would identify it. Notably, our approach would be of use even if the researcher were "only" carrying out a linear regression model, and would be flagged in both steps even if only this were being used. This would highlight the implausibility of generating a model that explains such a high proportion of the variance and yet has not a single statistically significant term.

Non-identifiability in mixed models and comparative analysis

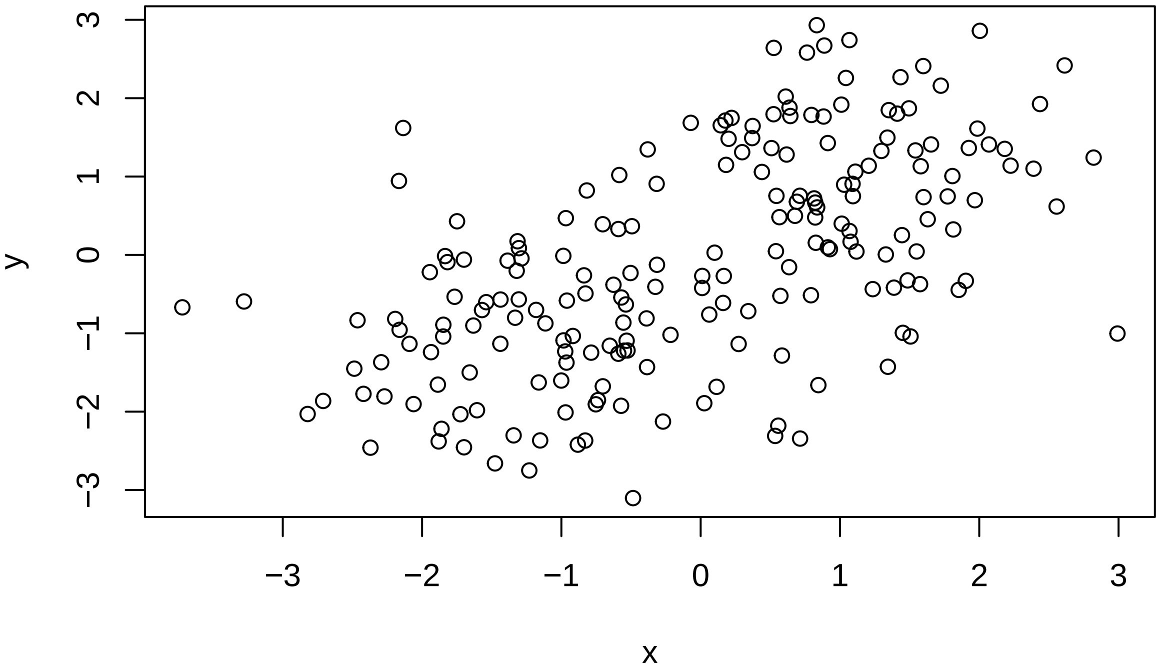
A slightly less contrived case of non-identifiability is essentially the reason for the entire existence of the field of comparative analysis. In 1985, Joe Felsenstein (DOI: 10.1086/284325) observed that two lineages, evolving separately, could generate data that might look as though they were correlated, when in fact they were not and simply represented two lineages evolving separately with no apparent correlation.

```r
# Simulate one lineage
apples.x <- rnorm(n=100, mean=-1, sd=1)
apples.y <- rnorm(n=100, mean=-1, sd=1)

# Simulate another lineage
oranges.x <- rnorm(n=100, mean=1, sd=1)
oranges.y <- rnorm(n=100, mean=1, sd=1)

# Merge the data
data <- data.frame(x=c(apples.x,oranges.x),
  y=c(apples.y,oranges.y), fruit=rep(c("apples","oranges"),each=100))

# They look correlated...
with(data, plot(y~x))
```



```r
# ...and indeed a simple statistical test shows they are
cor.test(c(apples.x,oranges.x), c(apples.y,oranges.y))
```

```
##
##  Pearson's product-moment correlation
##
## data:  c(apples.x, oranges.x) and c(apples.y, oranges.y)
## t = 9.873, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4733454 0.6604753
## sample estimates:
##       cor
## 0.5743664
```

Joe Felsenstein's insight, which led to him writing a paper that has been cited over 11,000 times and has spawned an entire field, was to recognise that fitting this statistical model without considering the generative process, as our approach encourages the reader to do, is flawed. There is no correlation between `x` and `y` in these data, but rather `x` and `y` are each drawn from one of two separate distributions ($normal(1, 1)$ or $normal(-1, 1)$) that we have 'mixed' together.

By carrying out Step 1 of our process (working through what an evolutionary process, in words, might look like) Felsenstein was able to identify that two separate models are non-identifiable (we cannot tell the difference between statistical correlation and two separate distributions). In his seminal paper, he then demonstrated this analytically (which is somewhat equivalent to Step 2 in our workflow), and then laid the groundwork for the field of comparative analysis of species' traits.