Data Analysis in Evol/Evol - HW Week7

Jing-Yi Lu
Reference R code see: github.com/jingyilu/Data-analysis-ecoevo

1. Use linear regression models to identify the important factors affecting number of species and number of native species observed:

i. Number of species

	Partial Coefficient Estimates	*t*	*P*
(Intercept)	(11.4848)	0.448	0.6591
Area	-0.0285	-1.141	0.268
Elevation	0.3301	5.270	$\mathbf{4.37\times10^{-5}}$
Distance from nearest island	-0.1545	-0.134	0.894
Distance from Santa Cruz	-0.2600	-1.068	0.299
Area of adjacent island	-0.0796	-3.903	**0.0009**

ii. Number of native species

	Partial Coefficient Estimates	*t*	*P*
(Intercept)	(8.2567)	1.516	0.1468
Area	-0.0086	-1.708	0.105
Elevation	0.0823	6.412	$\mathbf{4.91\times10^{-6}}$
Distance from nearest island	0.0075	0.032	0.974
Distance from Santa Cruz	-0.0722	-1.446	0.165
Area of adjacent island	-0.0177	-4.273	**0.0005**

Elevation and the area of adjacent island contribute significantly to both the number of species and the number of native species, especially for elevation. The elevation contributes positively to the number of species/native species, that is, increasing the highest elevation of an island leads to higher species number. On the contrary, the area of adjacent island contributes negatively to the number of species. The larger the adjacent island is, the fewer plant species were observed.

2.

a) Hypotheses testing for
H_0 = There is no association between the C and D loci (Two loci are independent.)
H_1 = There is association between the C and D loci (Two loci are not independent.)
Use Fisher's exact test for each population (2x2 contingency table, test for independence.)
Dutch: P-value = 2.4×10^{-100}
Poles: P-value = 2.2×10^{-13}
Greeks: P-value = 7.2×10^{-73}
We can reject the null hypotheses that two loci are independent for all populations. There is association between the C and D loci for every population.

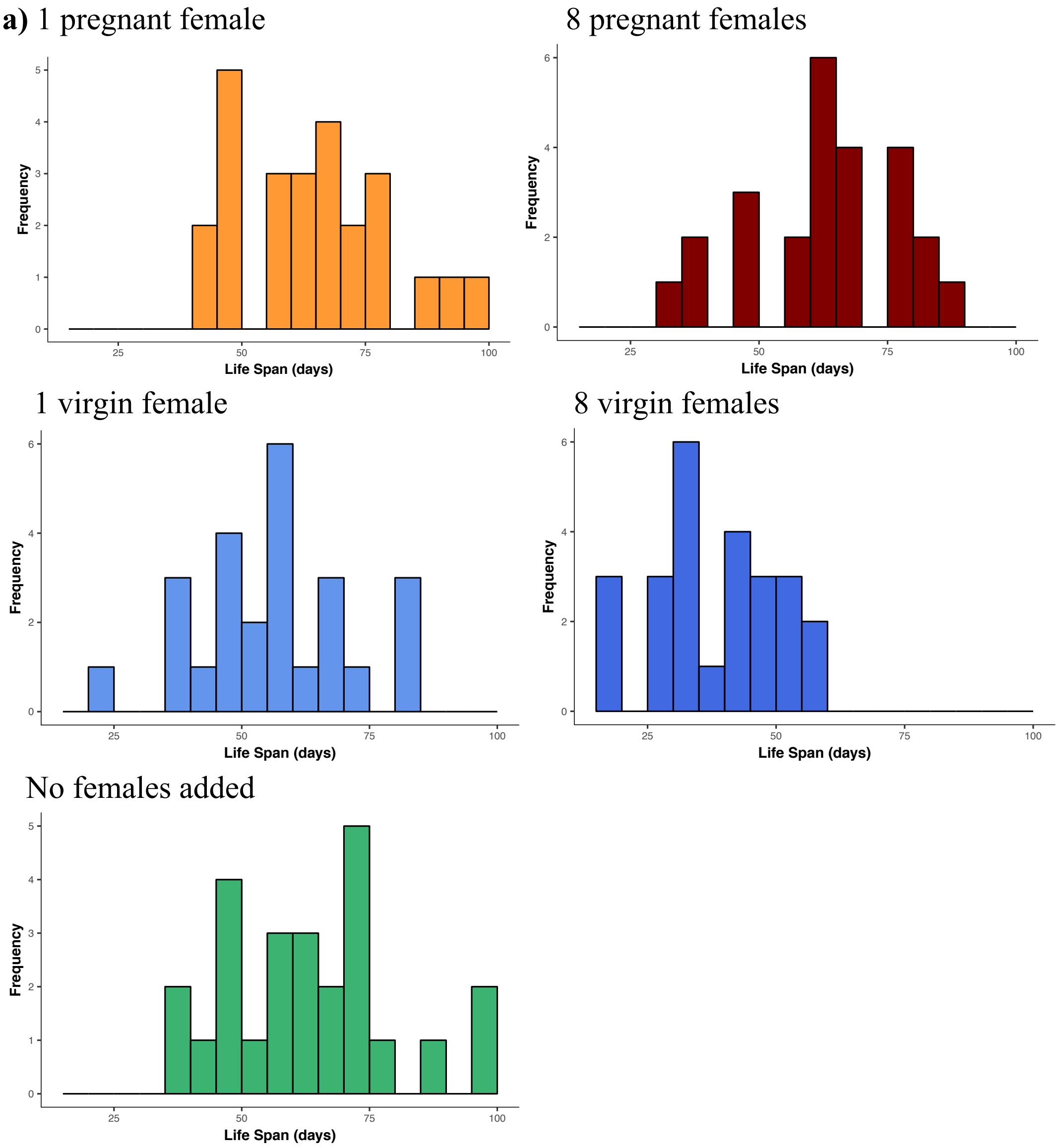
Considering correction multiple comparison by Bonferroni correction as the same loci (C and D) are tested in three populations.

Dutch: Adjusted P-value < 7.2×10^{-100}
Poles: Adjusted P-value = 6.6×10^{-13}
Greeks: Adjusted P-value < 2.2×10^{-72}
The correction does not affect the conclusion drown from the original Fisher's exact tests.

b) Hypothesis testing for
H_0 = There is no difference in gene (haplotype) frequency among different populations.
H_1 = There are differences in gene (haplotype) frequency among different populations.
Use chi-square test for the contingency between haplotype frequencies and populations.
$X^2 = 22.244$, df = 6, P-value = 0.001094
We can reject the null hypothesis of no difference in haplotype frequency among different populations.
There is no need for correction for multiple comparison as we only test one hypothesis.

3.

a)



1 pregnant female

8 pregnant females

1 virgin female

8 virgin females

No females added

b)

	Sum of square	**df**	***F***	***P***
Treatment	11939	4	13.61	**3.516×10^{-9}**
Residual	26314	120		

The experimental treatments have significantly influenced the lifespan of male fruit flies.

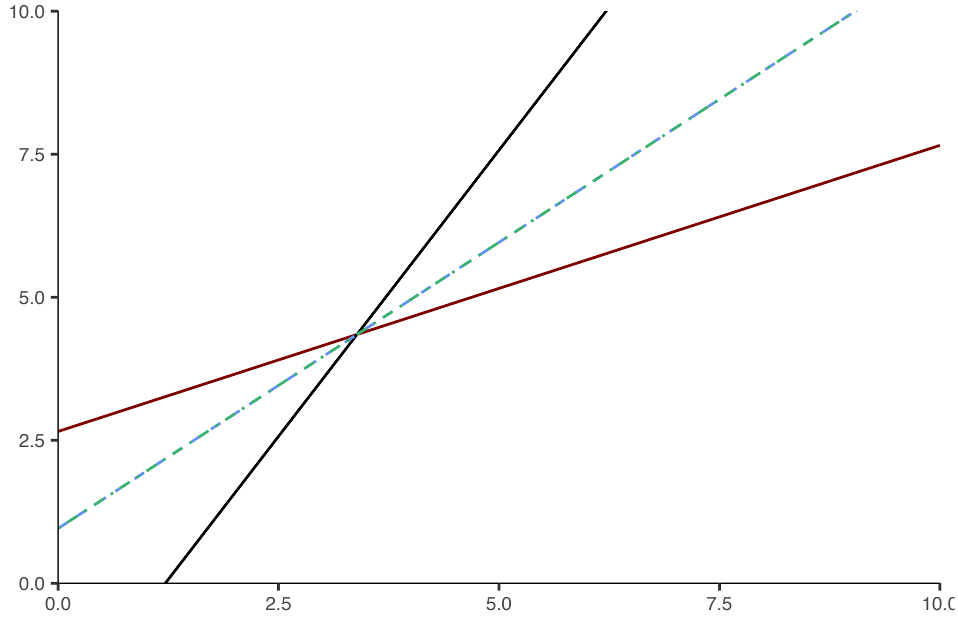
c) A fixed effect linear model. The treatment is a predetermined variable with different categories (actually the interacting term of number of added females and mating).

d) The assumption of equal variance for each group.
Use Levene's Test for homogeneity of variance across groups

	df	***F***	***P***
Groups = treatments	4	0.4916	0.7419

We cannot reject the null hypothesis that all groups have the homogeneous variance.

4.
Regression slope = cov(x,y)/var(x), as sd(x)=sd(y).
Correlation = 0.5 = cov(x,y)/sd(x)sd(y) = cov(x,y)/var(x) = regression slope of y~x.
The same for the regression slope of x~y (=0.5).
We randomly draw means for x and y (see R codes) and plot the regression y~x as the red line and the regression x~y as the black line.



Given a (predicted) y, we can find the future x predicted by y is closer to the mean (the point of the intersection in the plot) than the observed x. The variables, thus, will gradually regress toward to mean.
We can then set the correlation coefficient to 1 and calculate the regression slopes for y~x and x~y. The two regression lines overlap with the slopes equal to 1 (the green and blue lines on the plot). That is, there will be no regression effect when the correlation coefficient equals to 1.

5.

First hypothesis:

$$\begin{tikzcd}
\text{Nutrition} \arrow[dr, "\beta_1\ (>0)"] & \\
& \text{Fitness} \\
\text{Body size} \arrow[ur, "\beta_2\ (>0)"'] &
\end{tikzcd}$$

Second hypothesis:

$$\begin{tikzcd}
\text{Nutrition} \arrow[r, "\gamma\ (>0)"] \arrow[rr, bend right, "\gamma\cdot\beta_1\ (>0)"'] & \text{Body size} \arrow[r, "\beta_1\ (>0)"] & \text{Fitness}
\end{tikzcd}$$