Analysis of covariance (ANCOVA)

- Some terminology regarding ”types” of variables:

  1) **Continuous variables** – can assume any value (e.g. age measured in microseconds...)

  2) **Discrete variables** – have ”gaps” between possible values

     2.1. **Interval variables** – the absolute size of the ”gap” between two values represents absolute distance between observations (e.g. age measured in years).

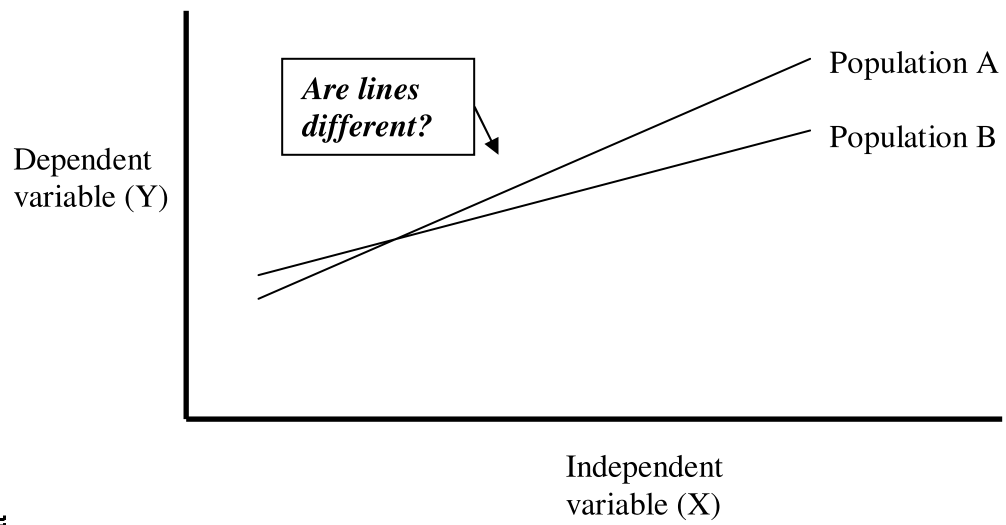
     2.2. **Ordinal variables** – categories of variable can be ordered in some clear fashion (e.g. age measured as either young or old).

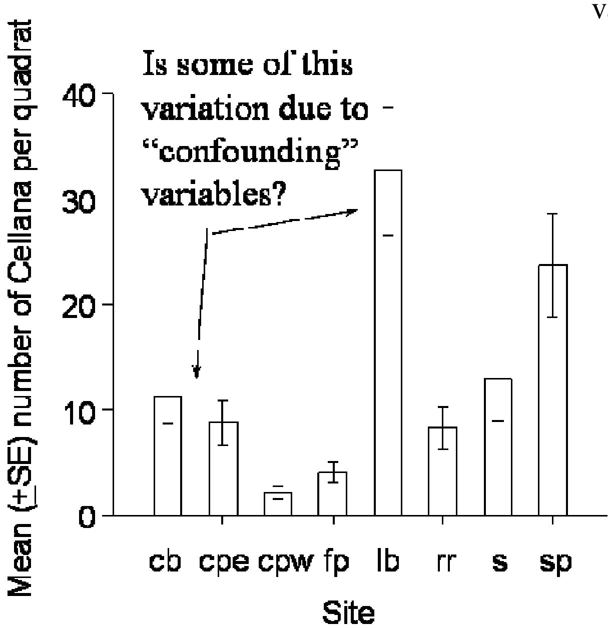
     2.3. **Nominal variables** – also called pure categorical variables; value reflects category but there is no intrinsic ordering to these categories (e.g. different individuals or cities).
- The distinction between types not always easy...
- So far, we have dealt with situations where we wish to relate a continuous response/dependent variable to either a set of continuous or interval variables (regression) or a set of ordinal or nominal variables (ANOVA).
- It is quite common that our set of ”explanatory” variables is a mix of continuous/interval and ordinal/nomial variables → **analysis of covariance.**

- **Two common uses of ancova (basic model is the same)**:

  1) When doing a **regression-type analysis**, you want to know whether the regression between X and Y differs (i.e., lines have different intercept [and/or slopes]) across different categories (e.g., populations, laboratories, seasons, etc).

  2) When doing a **ANOVA-type analysis**, you suspect that some residual variation in a continuous nuisance variable (e.g., temperature, moisture, age, size, etc) is introducing unwanted variance that you wish to "control for" in your analysis – more powerful analysis.





1. In regression type analyses

- To an ordinary regression of Y on X, add a factor by fitting the model

  $y_{ij} = \mu + \boldsymbol{\alpha_i} + \beta x_{ij} + \boldsymbol{(\alpha \times \beta x)_{ij}} + \varepsilon_{ij}$

  where μ is overall mean, α_i is the overall effect of a "factor" A on Y, βx_{ij} is the overall effect of X on Y, $(\alpha \times \beta x)_{ij}$ is the interaction between the two and ε_{ij} is unexplained residual variation (error).

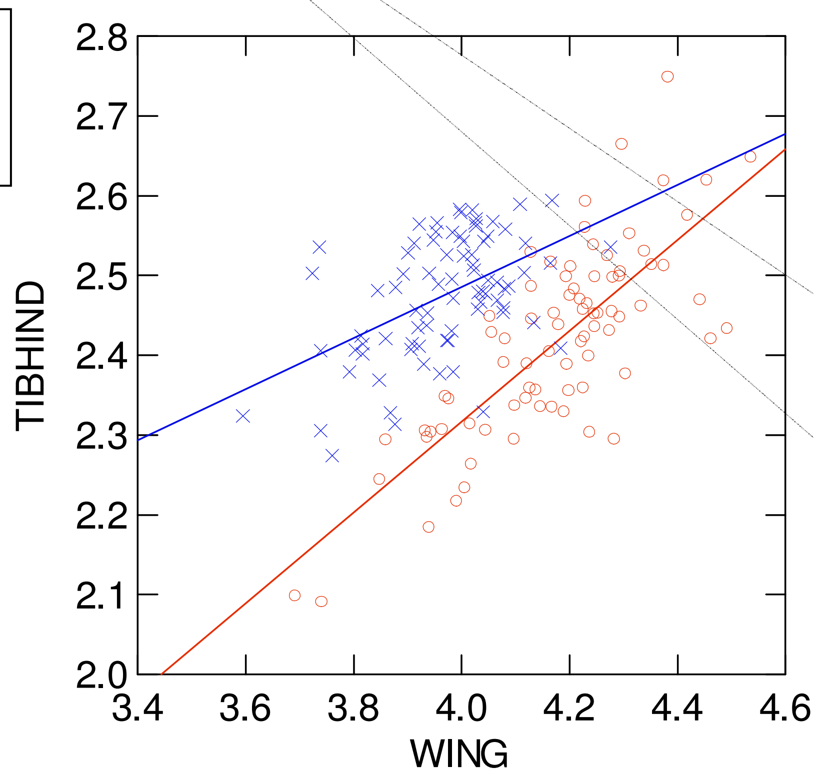
```
-------------------------------------------------------------------------------
Dep Var: TIBHIND   N: 160   Multiple R: 0.73346   Squared multiple R: 0.53796

Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P

SEX$                      0.06047     1      0.06047    11.85244     0.00074
WING                      0.57511     1      0.57511   112.71981     0.00000
SEX$*WING                 0.04521     1      0.04521     8.86178     0.00338

Error                     0.79593   156      0.00510
-------------------------------------------------------------------------------
```

Mis-interpreted in >50% of all published ANCOVAS!

Is the mean TIBHIND different for males and females, given that we take differences in wing length into account? Tests for **difference in intercept** of lines when interaction is included. Tests for **differences in elevation** (i.e., adjusted $\bar{Y}$ - $\hat{Y}$'s at overall $\bar{X}$!) **only when interaction is excluded** from the

Is there an overall effect of wing size on TIBHIND? Tests for **overall slope**.

Is the relationship between TIBHIND and wing different for males and females? Tests for **difference in slope**!



- Note that the ancova model contains one/more categorical predictor ("factor") and one/more continuous predictor ("**the covariate**").
- Focus here is on the relationship between one (or more) X and Y, and *whether and how* this relationship differs across factor levels.
- Much more complex "ancova's" than the above can be built on multiple regression models as well, with main effects of more than one factor and interactions with multiple X variables (**GLM's**).

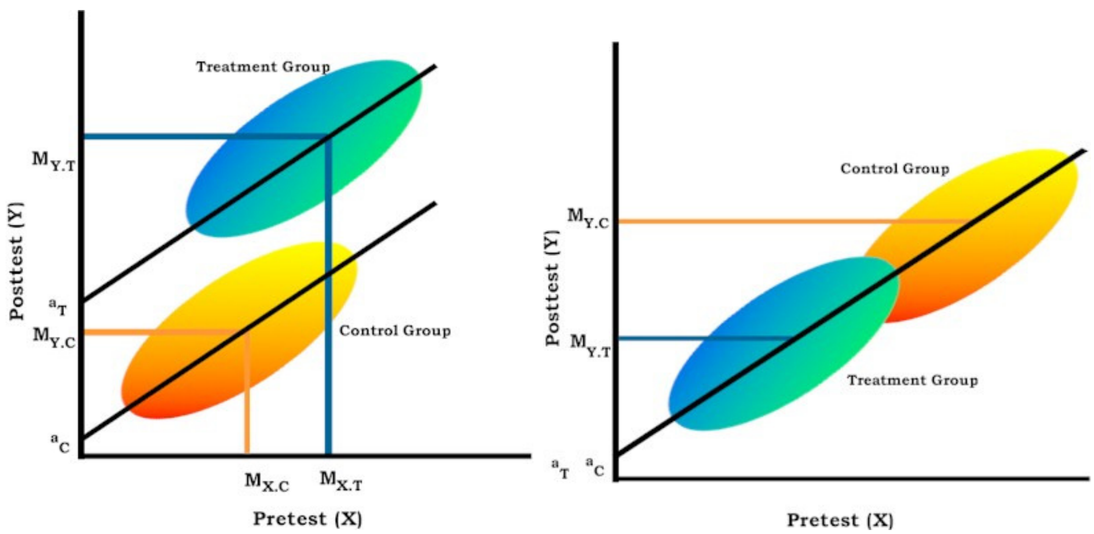
Example of GLM model of total offspring production in a beetle; a multipe regression model expanded to include two factors:

```
-------------------------------------------------------------------------------
Dep Var: TOTALOFFSPR   N: 353   Multiple R: 0.694   Squared multiple R: 0.482

Analysis of Variance
Source               Sum-of-Squares   df  Mean-Square     F-ratio       P

LIFESPAN   (X1)             271.817     1       271.817       0.954       0.329
NO_EGGS    (X2)            4997.186     1      4997.186      17.540       0.000
NO_EGGS*LIFESPAN(X1*X2)     263.593     1       263.593       0.925       0.337

POPULATION (Factor A)        77.967     1        77.967       0.274       0.601
TREATMENT  (Factor B)      5992.404     5      1198.481       4.207       0.001
POPULATION
*LIFESPAN                   152.571     1       152.571       0.536       0.465
POPULATION*NO_EGGS         2259.692     1      2259.692       7.931       0.005
TREATMENT*LIFESPAN        12289.474     5      2457.895       8.627       0.000
TREATMENT*NO_EGGS          2792.657     5       558.531       1.960       0.084

Error                     94303.214   331       284.904
-------------------------------------------------------------------------------
```

Contains two continuous variables (df=1; two parameters minus one), two factors (2 and 6 levels, respectively) and two-way interactions. ***Note that the results of complex Ancova's such as this one are not easily interpretable and have to rely on graphical plots and vizualization of patterns!*** E.g., the factor B may or may not have an effect on the level of the response variable, there is no way of telling from this Ancova table...

2. In ANOVA type analyses

- Same basic model as above; but focus here is on the effects of one (or more) factors on Y, given that the effects of one (or more) covariate is removed. Covariates are generally some sort of nuisance variable. ***Model for inference generally includes no interaction*** (see below).
- In the simple figures below, the treatment group has a different response (Y) in part because they tend to have a higher X – a simple one-way ANOVA would in part be misleading in left panel and entirely misleading in the right panel!



- Here, an ANCOVA asks: ***if you hold the covariate (X) constant, what is the effect of the factor?***
- A simple model: $y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$ basically tests for effects of a factor A (α_i) on the residuals from the regression of Y on X.

ANCOVA example: the effects of a factor on egg production in beetles, given that we control for lifespan.

```
--------------------------------------------------------------------------------
Dep Var: NO_EGGS   N: 350   Multiple R: 0.327   Squared multiple R: 0.107

Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P

TREATMENT                 284.447     5       56.889       4.024       0.001
LIFESPAN                  289.939     1      289.939      20.510       0.000

Error                    4848.764   343       14.136
--------------------------------------------------------------------------------
```

ANOVA for comparison; the effects of the same factor on residual egg production from a regression of egg production on life span…*same MS*…

```
--------------------------------------------------------------------------------
Dep Var: RESIDUAL   N: 350   Multiple R: 0.235   Squared multiple R: 0.055

Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P

TREATMENT                 284.445     5       56.889       4.036       0.001

Error                    4848.767   344       14.095
--------------------------------------------------------------------------------
```

- In this sense, **an ANCOVA is an ANOVA of adjusted means** (adjusted for the regression of Y on the covariate/s).
- Easily extended for any type of ANOVA model (factorial, nested, partly nested, repeated measures designs, etc).

Example; a two-way factorial ANOVA with two covariates:

```
--------------------------------------------------------------------------------
Dep Var: TOTALOFFSPR   N: 347   Multiple R: 0.548   Squared multiple R: 0.300

Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P

POPULATION              17041.921     1    17041.921      47.794       0.000
TREATMENT               11283.372     5     2256.674       6.329       0.000
POPULATION
*TREATMENT               2061.313     5      412.263       1.156       0.331

LIFESPAN                  580.308     1      580.308       1.627       0.203
H_RATE                  13403.627     1    13403.627      37.590       0.000

Error                  118738.812   333      356.573
--------------------------------------------------------------------------------
```

- Note that factor/s are **assumed to be fixed** by default in ancovas - if random effects variables are involved, F-ratios need to be recalculated in most software packages.
- ***VERY IMPORTANT***: The models above are "equal slopes models" – they **force the same slope** between X's and Y across all treatment levels! This is because models include no interaction. This is called the **equality/homogeneity of slopes assumption** in ANCOVA.
- This assumption **must be tested**, by a (multiple) partial F-test comparing

  (1) Y = factor effects + covariate effects *[an equal slopes model]*

  with

  (2) Y = factor effects + covariate effects + (factor effects × covariate effects) *[a different slopes model]*

  → If the addition of the interaction term/s *does not* significantly improve the model fit to data, then the ANCOVA model (1) above is appropriate.

  → If the addition of the interaction term/s *does* significantly improve the model fit to data, then the ANCOVA model (1) above is **inappropriate** and the interpretation of factor effects is complicated – slopes differ across treatment levels (use model (2) and interpret plots of the results).

Example: ANCOVA above no doubt **inappropriate** because of strong interaction between factor and covariate!

```
-------------------------------------------------------------------------------
Dep Var: NO_EGGS   N: 353   Multiple R: 0.411   Squared multiple R: 0.169

Analysis of Variance
Source              Sum-of-Squares   df  Mean-Square     F-ratio       P

TREATMENT                  379.868    5       75.974       5.126       0.000
LIFESPAN                   141.522    1      141.522       9.548       0.002
TREATMENT*LIFESPAN         417.115    5       83.423       5.628       0.000

Error                     5054.393  341       14.822
-------------------------------------------------------------------------------
```

- ANCOVA is superior to an ANOVA of residuals because it allows tests of interactions – cf. the equality/homogeneity of slopes assumption.
- ANCOVA also only makes sense if the relationship between covariate/s and Y is approximately linear – **check** with graphical inspections of data
- In sum, including one or more covariates can much improve your ANOVA by

  1) Accounting for small random differences between treatment levels in mean value of the covariate (X)

  2) Getting more precise/better measures of the reponse by "holding the covariate constant"
- But, only makes sense if

  1) Relationship between your covariate and Y is linear

  2) Slopes are homogenous – ***must*** be tested!

  3) There is a significant effect of the covariate (otherwise ANCOVA model collapses to an ANOVA).