General Linear Models, ANOVA and linear regression

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

20th August, 2558

What we will cover....

1. A brief review of Linear Regression
2. Analysis of Variance
   - Background
   - Test statistic: DIfferences between groups
   - ANCOVA: Mixing continuous and categorical predictors
   - Limitations of ANCOVA and ANOVA
3. The General Linear Model
   - ANOVA as a linear regression model
   - General Linear Models with categorical and continuous predictors
   - Interpreting General Linear Model Coefficients: Advanced models

The linear regression model

In the last session we saw that a multivariable linear regression is a linear model of the form:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

Linear regression model can be articulated using matrix algebra:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Where
y is vector of observations of our outcome variable;
X is a matrix containing a constant and at least one explanatory variable; and
β is a vector of parameters relating X to Y

Matrix formulation of linear regression

The matrix formulation:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Can be expanded out....

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k-1} \\ 1 & x_{2,1} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The first column of X is associated with the y-intercept (constant), β_0, and the rest of the columns (representing the individual covariates) are associated with the individual slopes, $\beta_1, \beta_2, \cdots, \beta_{k-1}$.

Interpretation of coefficients, β

β_0 **is y-intercept** (b_0 is the sample estimate)
Value of Y when X = 0

β_1 **is the slope associated with** x_1(b_1 is sample estimate)
The change in y for each unit change in x_1

All the remaining βs (through to $k-1$) are also slopes and can be interpreted in the same way as β_1.

Linear regression and least squares estimation

In the linear regression model we use the **Principle of Least Squares** to estimate β

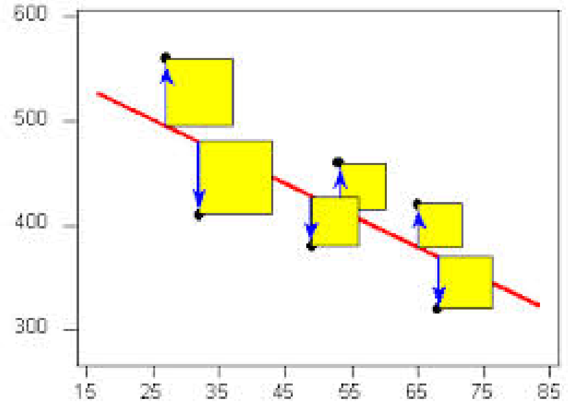
- Specifically, the model (e.g. line) is fit such that error sums of squares is minimized:

$$SSE = min\left(\Sigma_{i=1}^{n}\epsilon^2\right)$$

For this reason the the values of $\hat{\beta}$ (i.e. b) are called **least squares estimates**.

This use of *Least Square* estimates becomes important in the other models we will cover (next session onwards), which don't.

Geometric interpretation of least squares estimation



We want a line of best fit (**Red** line) that minimizes the sums of squares of the error (total of areas in yellow)

Hypothesis testing in Linear Regression

Overall model:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$

Interpretation: No covariates (Xs) explain variation in the outcome, y

H_A : At least one β_j differs from zero (for $j = 1, 2, \cdots, k-1$)

Interpretation: At least one of the covariates (Xs) explains y

Individual covariates:

If we reject the **global** H_0 above, then **for each covariate**:

$H_0 : \beta_j = 0$

Interpretation: The covariate, x_j, **does not** explain y

$H_A : \beta_j \neq 0$

Interpretation: The covariate, x_j, **does** explains y

Analysis of Variance

The expression, Analysis of Variance (ANOVA) can mean two things:

1. A statistical method (which uses a particular and outdated model); and
2. A type of hypothesis: Testing the equality of a continuous outcome across 2 or more groups

- Next we will cover Analysis of Variance (in the first sense)
- In other words, we will review the traditional ANOVA model only so I can show its redundancy (The General Linear Model is superior)
- We will also review ANOVA's hypothesis testing process (which carries through to the General Linear Models)

Analysis of Variance (ANOVA)

- One-way **An**alysis **o**f **Va**riance (ANOVA)
- Natural extension of the independent t-test to $>$2 groups
- Same assumptions as the independent t-test
  - Normally distributed dependent variable(within groups)
  - Equal variances
  - Independence between groups

ANOVA hypotheses

Null and alternative hypotheses are:

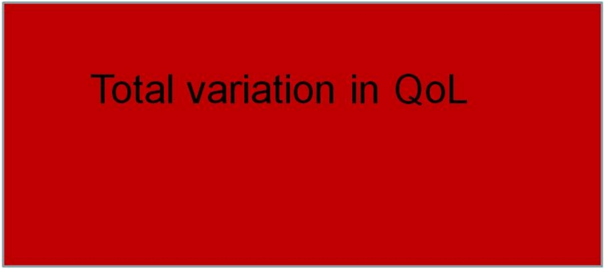
ANOVA hypotheses:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ (All group means are the same)
H_A: At least one group mean is different

- We are comparing means, so why call it an analysis of variance?
- Because we are going to analyse (partition) the 'spread' of the data
- What follows is an illustration of separating this spread

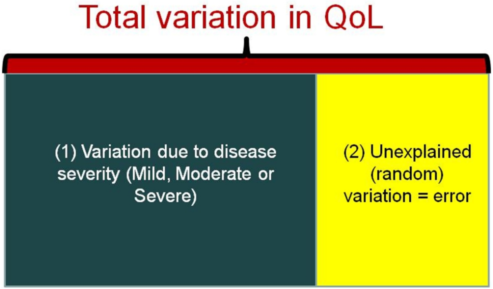
Analysis of Variance: ANOVA

Consider a illustration of the overall (total) variation in an outcome variable (Quality of Life $\sim$ QoL)



Analysis of Variance: ANOVA

Some variation in QoL might be explained by (1) Disease status (Mild, Moderate, Severe) and some by (2) the natural variation we would expect between patients.

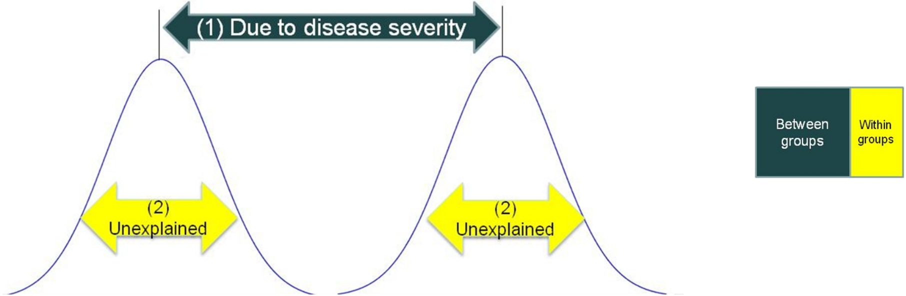

Total variation in QoL
(1) Variation due to disease severity (Mild, Moderate or Severe)
(2) Unexplained (random) variation = error

Testing: H_0: $\mu_{mild} = \mu_{mod} = \mu_{severe}$

We would expect to reject H_0 if between group differences (grey) was large relative to the natural variation (yellow).

Analysis of Variance: ANOVA

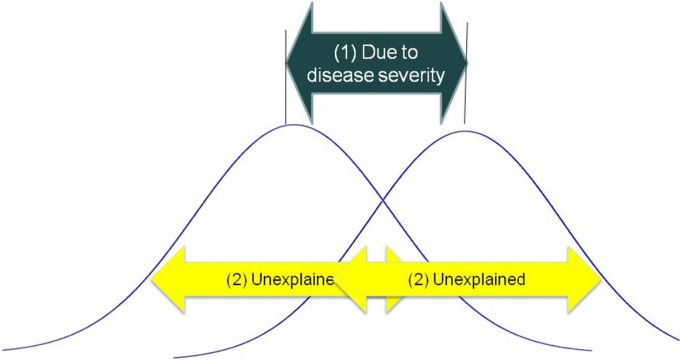
That is reject H_0: $\mu_{mild} = \mu_{mod} = \mu_{severe}$ if clear difference between (disease severity) groups:



Note: Two group case easier to illustrate.

Analysis of Variance: ANOVA

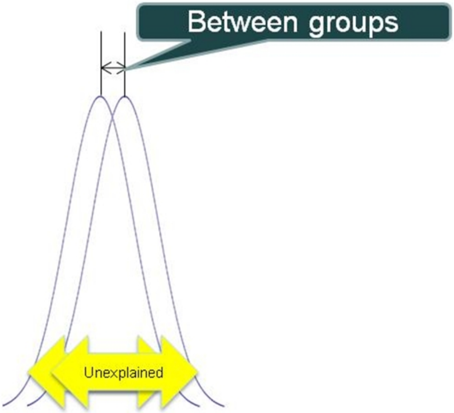
However, it would be much more difficult to reject H_0 (Groups are equal) if the difference between (disease severity) groups was not clear due to high within group variation:

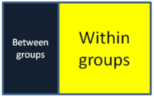


Between groups	Within groups

Analysis of Variance: ANOVA

Nor could we reject H_0:(Groups are equal) if there was little difference between (disease severity) groups:





Purpose of ANOVA

To recap, ANOVA is used to test for differences (in a continuous variable) between classes of categorical variables (**Factors**) and their interactions

For example, Systolic Blood Pressure between different racial groups:

$$H_0\colon\ \mu_{African} = \mu_{Asian} = \mu_{Caucasion}$$

ANOVA and Sums of Squares

- (Traditional) ANOVA is based on the Sums of Squares of certain differences (that relate to the variance formula)

$$SS_{total} = \sum_{i=1}^{N} [y_i - \bar{y}]^2 = SS_{between} + SS_{within}$$

where:

$SS_{between}$ represents variation **between** groups ('explained variation')

SS_{within} represents variation **within** groups ('error')

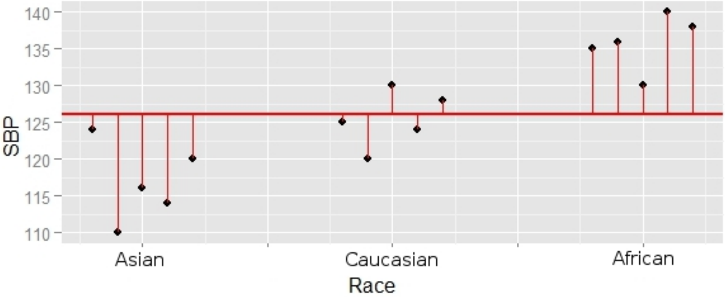
- So ANOVA is about partitioning the sums of squares ($\cong$ variation) hence, ANALYSIS OF VARIANCE

Sums of Squares formulae for ANOVA: SS_{total}

The total 'variation' (in Y) is represented by:

$$SS_{total} = \sum_{i=1}^{N} [y_i - \overline{\overline{y}}]^2$$

where $\overline{\overline{y}}$ is the grand (overall) mean.

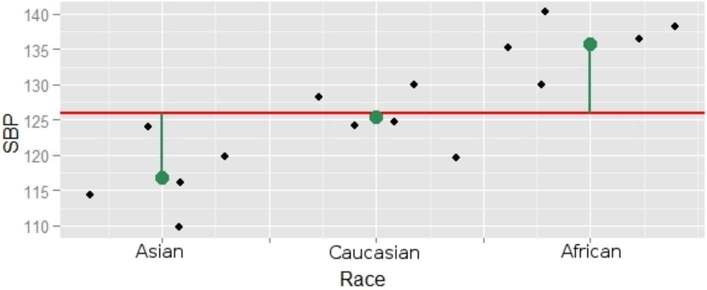


Sums of Squares formulae for ANOVA: $SS_{between}$

The between-groups variation SS will be:

$$SS_{between} = \sum_{i=1}^{k} [\overline{y}_i - \overline{\overline{y}}]^2$$

where $\overline{y}_i$ is the i^{th} group mean (for $i = 1, 2, \cdots, k$ groups)

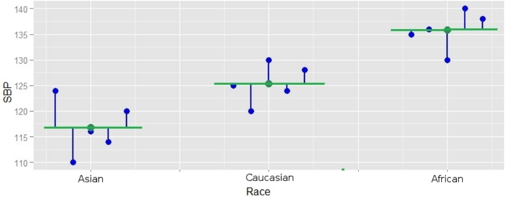


Sums of Squares formulae for ANOVA: SS_{within}

What is left over is the error SS:

$$SS_{error} = SS_{Total} - SS_{between} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij} - \overline{y}_i]^2$$

$$= \sum_{j=1}^{n_1}[y_{1,j} - \overline{y}_1]^2 + \sum_{j=1}^{n_2}[y_{2,j} - \overline{y}_2]^2 + \cdots + \sum_{j=1}^{n_k}[y_{k,j} - \overline{y}_k]^2$$



Sums of Squares $\rightarrow$ Mean squares

Now we have an idea of how much difference there is between groups (**which relates to our hypothesis**) relative to how much variation there is within groups (**the noise that can prevent us from demonstrating a difference**), can we now directly compare these these two quantities??

Answer: NO!!!!

If we look closely at the graphs on the previous slides, we should note that different numbers of values were used to calculate SS_{total}, $SS_{between}$ and SS_{within}. These three values are totals (the greater the number of values used to calculate them, the higher they will be).

Sums of Squares → Mean squares

- We need a way of offsetting this 'sample size' difference. This is where **Mean Squares** come in.
- Mean squares take into account the number of values used to calculate SS_{total}, $SS_{between}$ and SS_{within}
- For $SS_{between}$ and SS_{within} we calculate the corresponding mean squares $MS_{between}$ and MS_{within}
- Note we don't bother with MS_{total} as it is not used in the hypothesis test (see later), but if we did we would note that:

$$MS_{total} = \frac{SS_{total}}{N-1} = \frac{\sum_{i=1}^{N}[y_i - \overline{\overline{y}}]^2}{N-1} = S^2$$

where S^2 is the (overall) sample variance

$MS_{between}$ and MS_{within}

Now,

$$MS_{between} = \frac{SS_{between}}{k-1} = \frac{\sum_{i=1}^{k}[\overline{y}_i - \overline{\overline{y}}]^2}{k-1}$$

and,

$$MS_{error} = \frac{SS_{error}}{N-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij} - \overline{y}_i]^2}{N-k}$$

- Each SS is divided by it's corresponding **degrees of freedom**, which accounts for the number of values used to construct each sums of square.
- Now we have two standardized quantities that can tell us how different the groups are, relative to the variation within groups.

Note: k: number of groups; and N: overall sample size

The variance ratio

Now, how do we guage whether there is a **significant** difference between groups?

The variance ratio represents the ratio of the **between-group variance** (represented by $MS_{between}$) to the **within-group variance** (represented by MS_{within}). That is:

$$VR = F = \frac{MS_{between}}{MS_{within}} = \frac{\hat{\sigma}^2_{between}}{\hat{\sigma}^2_{within}} = \frac{S^2_{between}}{S^2_{within}}$$

The VR is the ratio between two variances, hence the name **Variance Ratio**

The variance ratio is also often represented by F. This is because (under H_0):

$$F \sim F_{df_1=k-1, df_2=N-k, \alpha}$$

The ANOVA table

With so many values floating around (SSs, MSs, degrees of freedom and the VR) it is more convenient to put them in a table, the **ANOVA table**:

Source	SS	df	MS	F
Groups	SS_{Groups}	$k-1$	$MS_{Groups} = \frac{SS_{Groups}}{k-1}$	$VR = \frac{MS_{Groups}}{MS_{Error}}$
Error	SS_{Error}	$N-k$	$MS_{Error} = \frac{SS_{Error}}{N-k}$	
Total	SS_{Total}	$N-1$		

ANOVA vs. Linear Regression

- Both Linear regression and ANOVA model quantitative outcome (response) variables
- Widely held view that **regression** for **covariates (i.e. continuous explanatory variables)** and **ANOVA** for **factors (i.e. categorical explanatory variables)**
- But traditional ANOVA can be extended to also incorporate covariates (quantitative explanatory variables)
- This extention is called Analysis of Covariance (ANCOVA)
- Won't cover in detail (General linear model a better approach)

Example of ANCOVA

We want to examine the effect of **Gender** on **Systolic Blood Pressure (SBP)** but we know that **Age** has a large effect on **SBP** (although we aren't interested in the age effect)

- For example, in an observational study we may find that there are more older females than males and we don't want the age effect CONFOUNDING the gender effect, so we need to adjust for age
- In this example, we can think of partialling out (adjusting for) age as effectively (and artificially) making every subject the same (average) age

ANCOVA example

So for our example....

$$SBP_{ij} = \mu + Gender_i + \beta Age_{ij} + \epsilon_{ij}$$

or (equivalently),

$$SBP_{ij} - \beta Age_{ij} = \mu + Gender_i + \epsilon_{ij}$$

Where,
SBP_{ij} is the j^{th} replicate of the i^{th} gender group
μ represents the grand (overall) mean
β is the slope for Age; and
Age_{ij} and ϵ_{ij} are the corresponding values of Age and the residual associated with SBP_{ij}
$Gender_i$ represents the difference (from the grand mean) due to being in the i^{th} gender group

Problems with ANCOVA and traditional ANOVA

- I will not even run through a formal example for ANOVA and ANCOVA
- Both of these methods have issues based on there use of sums of squares formulae
- In particular, these methods are susceptable to differences in group sample sizes (unbalanced designs)
- Where there are even moderate differences in sample sizes, the larger group is disproportionally weighted

Pitfall: What is in a name?

I will avoid the names ANOVA and ANCOVA because the confusion between the hypotheses they test (which are fine), and the mathematical approach they use (which is outdated and can lead to problems)

Solution: Linear Regression (Life, the universe and everything)

- Fortunately, everything that is done using an ANOVA or an ANCOVA can be formulated using a plain old garden variety linear regression model
- Regression models of this type deal with unequal sample sizes (i.e. every observation is appropriately weighted)
- The only cost: AT FIRST, seem trickier to interpret coefficients

Linear Regression and the General Linear Model

Linear Regression models used to perform analysis of variance type analyses or even those including categorical predictor variables somewhere in the model are given the name:

General (or Normal) Linear Models

Do NOT confuse with **Generalized** Linear Models (e.g. Logistic and Poisson regression)

A general linear model example(over-parametrized)

Consider the model:

$$SBP_i = \beta_0 + \beta_{age} Age_i + \beta_M M_i + \beta_F F_i + \epsilon_i$$

Noting:

- Unlike ANOVA/ANCOVA models, the indicies have been restricted to i, for $i = 1, 2, \cdots, N$
- Instead of **Gender**, we have **M** and **F** where:
  - If the patient is male $M = 1$ (and $F = 0$); and
  - If the patient is male $F = 1$ (and $M = 0$)

This model is incorrect:
CAN YOU SEE THE PROBLEM????

A General Linear Model Example

- M and F are co-linear (in fact, perfectly)
- If M=1 then F=0 and Vice Versa (perfectly correlated)
- To be a little bit mathematical about it the $\mathbf{X}$ matrix (matrix containing covariates) is not of full rank (there is redundancy).
- We only need **one** variable to indicate **two** states of GENDER (and get an estimate of difference)

A General Linear Model Example (Correct)

$$SBP_i = \beta_0 + \beta Age_i + \beta_{Gender} Gender_i + \epsilon_i$$

- Where Gender = 0 for males and Gender = 1 for females.
- In this case we are using Males as the **referent**, and β_{Gender} tells us the difference (in SBP) from males if you are female (It is not a 'slope' in the traditional sense).

Important point: Referents

This will be the first time you have been given the formal (mathematical) definition of the **referent**. An important concept that carries through to almost all statistical modeling

Let's look at this from the point of view of the X matrix

A General Linear Model Example

The additional (redundant) variable is removed to give us a matrix of full rank (no redundancy)

$$\begin{bmatrix} 1 & 23 & 1 & 0 \\ 1 & 46 & 1 & 0 \\ 1 & 38 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 64 & 0 & 1 \\ 1 & 36 & 0 & 1 \\ 1 & 27 & 0 & 1 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 23 & 0 \\ 1 & 46 & 0 \\ 1 & 38 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 64 & 1 \\ 1 & 36 & 1 \\ 1 & 27 & 1 \end{bmatrix}$$

Note the columns of ones on the left hand side is associated with the constant, β_0

A fictitious example

$\beta_0 = 100$ (p$<$0.05); $\beta_{Age} = 0.333$ (p$<$0.05); and $\beta_{Gender} = -25$ (p$<$0.05)

Interpretation:

- $\beta_0 = 100$: Expected SBP of someone who is 0 years old and gender of 0 (Males). p$<$0.05 $\Rightarrow$ Baby male SBP (on average) is significantly different from 0 (So what???)
- $\beta_{Age} = 0.333$: As we age one year, on average our SBP should rise by 0.333. p$<$0.05 $\Rightarrow$ Age does explain SBP
- $\beta_{Gender} = -25$: When Gender is 1(i.e. female) we expect our SBP to be lower (relative to males) by 25 units. p$<$0.05 $\Rightarrow$ On average Female SBP differs from Males

β for categorical variables

Since:

- p<0.05 there is significant difference between males and females (i.e. $\beta_{Gender} \neq 0$).
- Males are the referent, $\beta_{Gender} = -25$ represents the difference due to being a female (i.e. The mean SBP for females is 25 units less than males).

We can use this information to work out the **estimated marginal means**:

$\hat{\mu}_{Males} = 100$ i.e. $100 + \beta_{Gender}(0)$
$\hat{\mu}_{Females} = 75$ i.e. $100 + \beta_{Gender}(1)$

Actually $\hat{\mu}_{Males}$ and $\hat{\mu}_{Females}$ not yet right, I still need to account for age (see next slide)

Estimated marginal means

Technically,
$\hat{\mu}_{Males} = 100$ i.e. $100 - \beta_{Gender}(0)$
$\hat{\mu}_{Females} = 75$ i.e. $100 - \beta_{Gender}(1)$
gives us the the SBP of new born babies (i.e. Age=0). This is clearly not appropriate in this population (which is an adult population)
We should use the average age in our calculations.
Remembering $\beta_{Age} = 0.33$, if the average age ($\overline{age}$) in our sample is 30 years old, then

$\hat{\mu}_{Males} = 110$ i.e. $100 + \beta_{Gender}(0) + \beta_{Age}(30)$
$\hat{\mu}_{Females} = 85$ i.e. $100 + \beta_{Gender}(1) + \beta_{Age}(30)$

Estimated marginal means Vs. sample means, $\bar{x}$

- This seems like a very complicated process to go through to get an estimate of the group means!!!
- Why didn't we just calculate the sample means ($\bar{x}$s)

Has anybody got an answer????

Hint: Confounding

Important point: Adjusted estimates

Adjusting estimates is one of the main advantages (among others) of using a multi-variable 'modelling' approach over the classical bivariate tests (i.e. t-test, χ^2 tests etc)

Estimated marginal means Vs. sample means, $\bar{x}$

Confounding:

Let's say our study is observational and we note that (on average) females are older than our males.

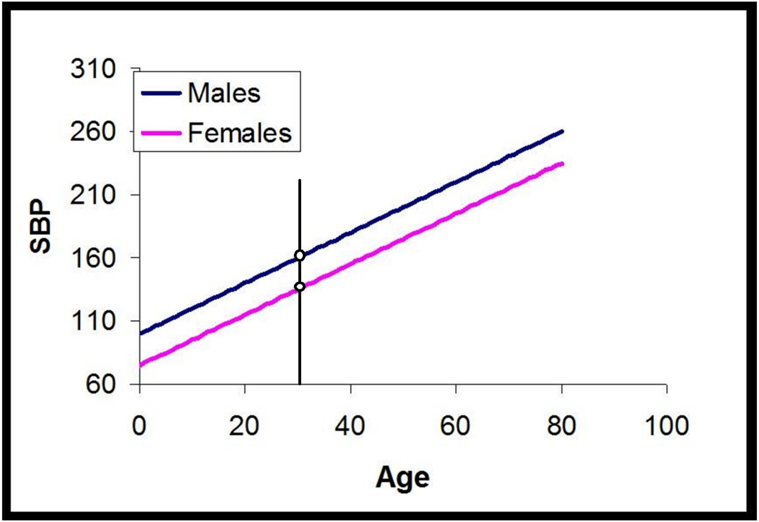
In this case the difference between our sample means ($\bar{x}_{Males}$ and $\bar{x}_{Females}$) represents:

- *The difference in SBP between males and females*; AND
- *the differences in SBP among males and females due purely to the age differences between these groups*

In other words, the Age effect is **CONFOUNDING** the Gender effect.

As Estimated marginal means adjust for confounders (in the model), the resulting means represent the difference purely due to the effect of interest(study effect)

A graphical interpretation



Variables with more than two categories (Polytomous categoricals)

- Interpretation becomes a little trickier when there are more than two groups.
- Let's try a three group problem.
- Same problem as before (SBP), but what about ethnicity (African, Asian, Caucasian)?

What about the multicollinearity problem?

- As with gender, we can't use as many variables as there are classes (collinearity)
- We only need two indicator(dummy) variables to represent the three states (ethnicity).
  1. D1 = 1, D2 = 0 implies African
  2. D1 = 0, D2 = 1 implies Asian
  3. D1 = 0, D2 = 0 implies Caucasian
- There are a number of ways of doing this (varies from package to package) which changes interpretation
- In this case, which is the referent??

ANS:????

Number of classes vs. number of dummy variables

- For two classes (e.g. Gender) we only needed one indicator variable.
- For three classes (e.g. Race), two indicator variables.
- **Generalizing: for k classes we need k-1 indicator variables**

WHERE HAVE YOU SEEN THIS BEFORE?

ANS: Degrees of freedom

Fictitious Example 2:SBP by race

$$SBP_i = \beta_0 + \beta_1 D1_i + \beta_2 D2_i + \epsilon_i$$

$\beta_0 = 100$ (p<0.05); $\beta_1 = 23$ (p=0.12); $\beta_2 = -35$ (p<0.05) $\Rightarrow$
$\hat{\mu}_{Caucasian} = \beta_0 + \beta_1(0) + \beta_2(0) = 100$
$\hat{\mu}_{African} = \beta_0 + \beta_1(1) + \beta_2(0) = 123$
$\hat{\mu}_{Asian} = \beta_0 + \beta_1(0) + \beta_2(1) = 65$

- No sign. diff between African and Caucasians (p=0.12)
- Asians lower SBP than Caucasian (p<0.05)
- Incidently, the diff between Africans and Asians is: $\beta_1 - \beta_2 = 58$

Aside:

Comparisons not involving the referent (e.g. Asian vs. African) need to be tested *post-hoc*

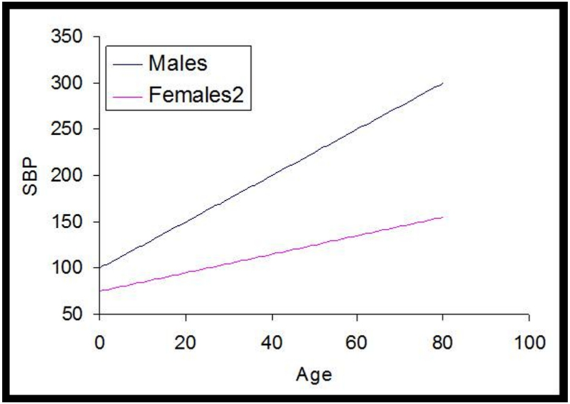
Interactions: Effect modification

- So far we have considered the case where we can interpret risk factors independently of each other
- For instance in our model: $SBP_i = \beta_0 + \beta Age_i + \beta_{Gender} Gender_i + \epsilon_i$
  - Gender effect (on SBP) was interpreted independently of age; and
  - Age effect (on SBP) was interpreted without reference to gender.

However, what if gender modifies the effect of Age on SBP??

Effect modification

What about if we came across the situation:



Here the effect of Age (on SBP) is more profound for males, than for females

Interactions in GLMs: Interpretation

From previous slide:

- It appears that the nature of the SBP-Age relationship is **modified by** (interacts with) the gender effect
- Gender (in this case) represents an **effect modifier**

Now how would we represent (and test for) effect modification in a General Linear Model??

Interaction term in General Linear Models

- Effect modification is often called an **Interaction** and it represent a multiplicative effect of two factors (or covariates, or some combination thereof) in the model
- Where A (e.g. Age) and B (e.g. Gender) are **main** effects, A*B is an **interaction** effect
- Now, to implement these in the General Linear Model....

Example: General Linear Model with interaction effect

$$SBP_i = \beta_0 + \beta_{age} Age_i + \beta_{Gend} Gend_i + \beta_{Age \times Gend}(Age_i \times Gend_i) + \epsilon_i$$

Males (Referent)

$$SBP_i = \beta_0 + \beta_{age} Age_i + \beta_{Gend}(0) + \beta_{Age \times Gend}(Age_i \times (0)) + \epsilon_i$$

$$SBP_i = \beta_0 + \beta_{age} Age_i + \epsilon_i$$

Females (A little messier)

$$SBP_i = \beta_0 + \beta_{age} Age_i + \beta_{Gend}(1) + \beta_{Age \times Gend}(Age_i \times (1)) + \epsilon_i$$

with a little manipulation

$$SBP_i = (\beta_0 + \beta_{Gend}) + (\beta_{age} + \beta_{Age \times Gend}) Age_i + \epsilon_i$$

Interaction effects in General Linear Models

Now we have a 'model' for **Males** and one for **Females**

Males (Referent)

$$SBP_i = \beta_0 + \beta_{age} Age_i + \epsilon_i$$

and

$$\hat{\mu}_{Males} = \beta_0 + \beta_{age} \bar{Age}$$

Females

$$SBP_i = (\beta_0 + \beta_{Gend}) + (\beta_{age} + \beta_{Age \times Gend}) Age_i + \epsilon_i$$

$$\hat{\mu}_{Females} = (\beta_0 + \beta_{Gend}) + (\beta_{age} + \beta_{Age \times Gend}) \bar{Age}$$

Interpreting the coefficients β_{Gend} and $\beta_{Age \times Gend}$...

β_{Gend} is the **difference in average** SBP due to being female

$\beta_{Age \times Gend}$ is the **difference in age effect** due to being female

Effect modification for a multilevel factor

Now let's see how smart you are.....
Considering Ethnicity: Caucasian, African and Asian groups

$$\begin{aligned} SBP_i = {} & \beta_0 + \beta_{Age} Age_i + \beta_{D1} D1_i + \beta_{D2} D2_i \\ & + \beta_{Age \times D1}(Age_i \times D1_i) + \beta_{Age \times D2}(Age_i \times D2_i) + \epsilon_i \end{aligned} \tag{1}$$

1. Using Model (1) find a 'model' for each group
2. From step 1, provide estimated marginal means
3. Interpret all of the coefficients

I have provided the answer at the end (No peeking) ☺

Space for working: Caucasians

-

Space for working: African

-

Space for working: Asians

-

Race group models

Caucasian ('Referent model')

$$SBP_i = \beta_0 + \beta_{Age} Age_i + \beta_{D1} 0 + \beta_{D2} 0 + \beta_{Age \times D1} (Age_i \times 0) + \beta_{Age \times D2} (Age_i \times 0) + \epsilon_i \quad (2)$$

$$SBP_i = \beta_0 + \beta_{Age} Age_i + \epsilon_i$$

African

$$SBP_i = \beta_0 + \beta_{Age} Age_i + \beta_{D1} (1) + \beta_{D2} 0 + \beta_{Age \times D1} (Age_i \times 1) + \beta_{Age \times D2} (Age_i \times 0) + \epsilon_i \quad (3)$$

$$SBP_i = \beta_0 + \beta_{Age} Age_i + \beta_{D1} + \beta_{Age \times D1} Age_i + \epsilon_i$$

$$SBP_i = (\beta_0 + \beta_{D1}) + (\beta_{Age} + \beta_{Age \times D1}) Age_i + \epsilon_i$$

Race group models

Asian

$$\begin{aligned} SBP_i = \beta_0 + \beta_{Age} Age_i + \beta_{D1}(0) + \beta_{D2} 1 \\ + \beta_{Age \times D1}(Age_i \times 0) + \beta_{Age \times D2}(Age_i \times 1) + \epsilon_i \end{aligned} \quad (4)$$

$$SBP_i = \beta_0 + \beta_{Age} Age_i + \beta_{D2} + \beta_{Age \times D2} Age_i + \epsilon_i$$

$$SBP_i = (\beta_0 + \beta_{D2}) + (\beta_{Age} + \beta_{Age \times D2}) Age_i + \epsilon_i$$

Now for the estimated marginal means

Step 2: Estimated Marginal Means

Caucasian: from (2)

$$\hat{\mu}_{Caucasians} = \beta_0 + \beta_{Age}\bar{Age}$$

African: from (3)

$$\hat{\mu}_{Africans} = (\beta_0 + \beta_{D1}) + (\beta_{Age} + \beta_{Age \times D1})\bar{Age}$$

Asian: from (4)

$$\hat{\mu}_{Africans} = (\beta_0 + \beta_{D2}) + (\beta_{Age} + \beta_{Age \times D2})\bar{Age}$$

Interpretation of coefficients

$\beta_0 \rightarrow$ SBP for (new born) Caucasians(referent)
$\beta_{age} \rightarrow$ Age slope for caucasians (expected change in SBP for every year older)
$\beta_{D1} \rightarrow$ Difference in SBP (from referent) due to being African (Note: only appropriate if no interaction i.e. $\beta_{Age \times D1} = 0$)
$\beta_{D2} \rightarrow$ Difference in SBP (from referent) due to being Asian (Note: only appropriate if no interaction i.e. $\beta_{Age \times D2} = 0$)
$\beta_{Age \times D1} \rightarrow$ Age effect modification: Difference in Age slope (from referent) due to being African
$\beta_{Age \times D2} \rightarrow$ Age effect modification: Difference in Age slope (from referent) due to being Asian
All of these βs should be tested against zero (except β_0, this hypothesis test is generally meaningless)

Linear Regression and General Linear Models in R

As Linear regression and General Linear Linear models are really the same model, R uses the `lm()` method for both.

Assuming the data has already been read into R...

Gender only

```
my.model1<-lm(sbp∼as.factor(gender), data=mydata.df)
```

Note: we have to tell R that *Gender* is categorical

Main effects only: Gender and Age

```
my.model1<-lm(sbp∼as.factor(gender) + age, data=mydata.df)
```

Now with interaction effect: Gender and Age

```
my.model1<-lm(sbp∼as.factor(gender )+ age +

as.factor(gender):age , data=mydata.df)
```

It is **very important** to note the form of these models, because the same basic convention is used for all R modelling (e.g. Generalized linear models)

Concluding remarks

- At first glance, the General Linear Model looks nasty Biohazard
- But an understanding of them is vital to truly understand the common extentions to the General Linear Model used in biostatistics including:

1. Generalized Linear Models (e.g. Logistic regression, Poisson regression etc..)
2. Linear Mixed Models (continuous outcomes in longitudinal and other 'correlated data' studies)
3. Survival analysis methods e.g. Cox proportional hazards regression
4. Generalized Estimating Equations and Generalized Linear Mixed Models (categorical outcomes from longitudinal and other 'correlated data' studies); a 'theoretical' mix of 1 and 2

Final word

Biostatistics is not 'Book-learnt'

- Rarely do we understand biostatistical methods as we are introduced to them (I didn't)
- My advice is to review these notes when you get home tonight...slowly
- Put them aside for a few days and review them again (believe me it works)

Thank You!!!!