

# Linear Models: The General Linear Model

Dr Cameron Hurst  
cphurst@gmail.com

Faculties of Public Health and Medicine, Khon Kaen University

24<sup>th</sup> June 2013



## What we will cover....

- 1 Case study: The DMHT dataset
- 2 Naive methods for testing hypotheses about continuous outcomes
  - Correlation analysis
  - t-tests
- 3 A review of Linear Regression and ANOVA
  - Linear Regression
  - Analysis of Variance
- 4 ANCOVA: Mixing continuous and categorical predictors
  - ANCOVA
  - Limitations of ANCOVA and ANOVA
- 5 The General Linear Model
  - ANOVA as a linear regression model
  - General Linear Models with categorical and continuous predictors
  - Interpreting General Linear Model Coefficients: Advanced

- Before getting to the methods we will cover today, I will (re-)introduce you to the DMHT data
- We will use this data for the next few session (except survival analysis)
- I will also use this data to demonstrate the use of Stata for the methods I cover

## DMHT: Study background

- Collaborative clinical study supported by the Thailand National Health Security Office (NHSO) and the Thailand Medical Research Network (MedResNet)
- Official title: *An Assessment on Quality of Care among Patients Diagnosed with Type 2 Diabetes and Hypertension Visiting Ministry of Public Health and Bangkok Metropolitan Administration Hospitals in Thailand* (Thailand DM/HT)
- In short, main research objective is to assess quality of care of (Type 2) Diabetic and Hypertensive patients in Thailand
- At present, about 150,000 patients from about 600 across Thailand, sampled from 2553-5
- Three main groups: (1) Patients with T2DM (alone), (2) Those with HT (alone), and (3) Those with both (DMHT)

## (Our) research questions

For the purpose of our exercises today, our research questions are:

- 1 Among diabetics, does the hypertension comorbidity represent an additional burden for achieving clinical (quality of care) goals?
- 2 What other (diabetic) patient characteristics (e.g. demographic/lifestyle) might influence achievement of the clinical goals?

## Data we will consider

To keep things simple:

- We will consider a random sample of 5000 patients (sampled in 2554), and only the diabetics [T2DMs and DMHTs]
- Of the hundreds of variables, we will only consider a subset:

Outcomes	Study effect	Other covariates
a1cyn	ht	sex
bpyn		age
ldlcyn		religion
all3yn		duradm
any3yn		smoke
		bmigroup

*Detailed description of variables next few slides...*

## Variable description: Outcomes

The outcomes in our 'subset' of the dataset are the "ABC" clinical goals often used to assess the quality of diabetes care. These are:

- (a) `a1cyn` → Hemoglobin **A**1C: yes:  $< 7\%$ ; no:  $\geq 7\%$
- (b) `bpyyn` → **B**lood pressure: yes:  $< \frac{130}{80}$  mmHg; no:  $\geq \frac{130}{80}$  mmHg;
- (c) `ldlcyn` → Low density lipid-**C**holesterol: yes: LDL-C  $< 100$  mg/dL; no: LDL-C  $\geq 100$  mg/dL

We will also consider the collective quality performance outcomes:

- `all3yn`: **All** three (ABC) clinical goals are met: yes; no
- `any3yn`: **Any** of the three (ABC) goals are met: yes; no

**Question:** What type of measurement scale do all of these 5 variables have?

## Study effect

We are interested in whether there is any difference in the achievement of treatment goals between those with diabetes alone (T2DM) and those diabetics who ALSO have hypertension; Does a hypertension comorbidity represent an additional burden specifically in terms of diabetes care.

We will use the variable, `ht` to measure this: no (T2DM alone); yes (DM+HT)

### Note: Study Effects

Remember a **study effect** is the explanatory variable (X) that is of **primary interest** (in terms of our research hypothesis)

## Covariates

In observational studies (such as the DMHT study), many other explanatory (X) variables need to be considered. I will make a distinction between two different types of covariates here:

- 1 Independent risk factors: Other important **independent** explanatory variables of the outcome
- 2 Confounders: Variables that are associated with BOTH the outcome AND the study effect which, if ignored, can misleadingly enhance/diminish (bias) the true relationship between the outcome (Y) and the study effect (X)

### Example of a confounder: RCT

What would happen if randomisation (in an RCT) failed and we put the more sick people in the control group, and less sick people in the treatment group?

## Covariates

In our case study, I have included 6 covariates which can be split into two different groups:

- Demographic variables
  - sex: Patient gender (binary)
  - age: Age in years (continuous)
  - religion: Buddhist/Non-buddhist (binary)
- Lifestyle/patient history variables:
  - duradm: Duration of T2DM; How long (years) since patient was diagnosed with T2DM (continuous)
  - smoke: Smoking history: Current, Previous, Never, Unknown (Nominal/Ordinal)
  - bmgrou: Underweight, Normal, Overweight, Obese (Ordinal)

# Introduction

We will cover a lot of ground today. First I want to cover a few 'classical' methods (that you are probably already familiar with), that will not be appropriate for a large majority of studies.

However, they are still widely used (and misused). They are:

- 1 Correlation coefficients (Pearson's and Spearman's)
- 2 The test for two independent groups using the independent samples t-test
- 3 The test for paired data (Paired t-test)

From there we will go into statistical 'models'. A much more useful set of methods that are much more appropriate for the type of analysis you are likely to do.

## Which test should I use

- After a while, you will notice a pattern in which test should be used where
- The main factor that drives this decision is: **Which type (measurement scale) of OUTCOME variable do I have????**

		Explanatory	
		Numerical	Categorical
Dependant	Numerical	Linear regression, Correlation	ANOVA, t-tests
	Categorical	Logistic regression	$\chi^2$ test of independence, Logistic/Poisson regression

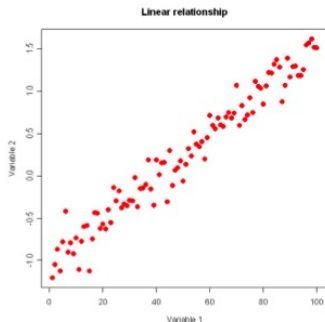
Note: This table only holds for cross-sectional data.

## Correlation analysis

- If we want to investigate the relationship between two continuous (quantitative) variables, we can use correlation analysis.
- If we believe the relationship is a linear (straight line) relationship we can use **Pearson's correlation coefficient**
- If we believe the relationship might depart from linearity, but is still monotonic (more later), we can use the nonparametric (distribution-free) **Spearman's correlation coefficient**

## Linear, Monotonic and Non-monotonic relationships

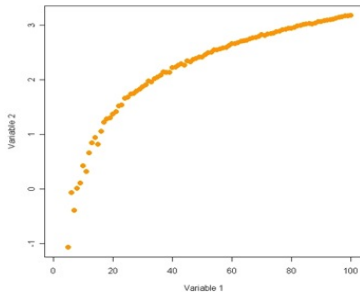
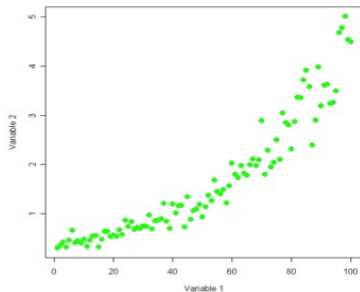
For **linear** relationships, use **Pearson's correlation coefficient**.



Defn of linear relationship: No matter where we are on the line, a single unit step in one variable (X), will always result in in the same level of change (increase/decrease) in the other variable (Y)

# Linear, Monotonic and Non-monotonic relationships

For **monotonic** relationships, use **Spearman's correlation coefficient**.

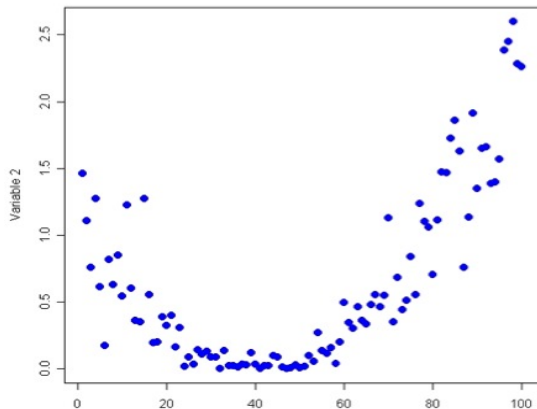


Both figures above are **Monotonically increasing**...uphill

Defn of monotonic relationship: If we move (e.g. walk) from left to right (on the X axis), we are always going up (or down) hill.  
However, the steepness of the slope doesn't need to be constant.

## Linear, Monotic and Non-monotonic relationships

For **non-linear AND non-monotonic** relationships, we need to use more sophisticated methods (where we can assume the functional form of relationship). E.g. Non-linear regression



## Is there a correlation between Age and the Duration of diabetes

Let's use the DMHT data to examine whether there is (1) a linear relationship between Age and Duration of diabetes; and (2) a monotonic relationship between these two variables. These give us the following sets of hypotheses:

$H_0$ : *There is no LINEAR relationship between age and duration*

$H_A$ : *There is a LINEAR relationship between age and duration*

and:

$H_0$ : *There is no MONOTONIC r'ship between age and duration*

$H_A$ : *There is a MONOTONIC r'ship between age and duration*

## Stata syntax: Correlation analysis

Note: ALWAYS visualize the data before analysis

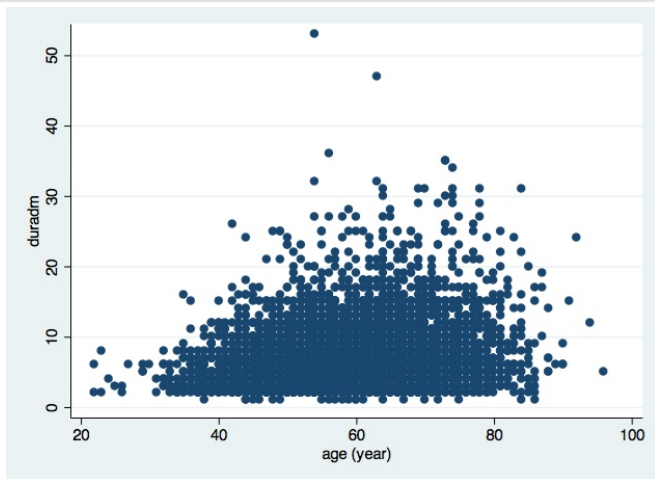
### Stata: Correlation analysis

```
* Scatter plot (note y before x)
twoway (scatter duradm age)

* Pearson's
pwcorr duradm age, sig

*Spearman's
spearman duradm age
```

# Visualizing relationship



Too much data to see what is going on (clearly)

# Stata output: Correlation analysis

```
. *Pearson's
. pwcorr duradm age, sig
```

	duradm	age
duradm	1.0000	
age	0.1989 0.0000	1.0000

```
.
.
. *Spearman's
. spearman duradm age

Number of obs = 4756
Spearman's rho = 0.1997

Test of H0: duradm and age are independent
Prob > |t| = 0.0000
.
.
end of do-file
```

## InterpretationCorrelation analysis

- We can see that both the Pearson's and Spearman's correlation coefficient are highly (statistically significant)
- BUT the relatively low value of the coefficients suggests that it may not be a clinically important association  
( $r_{pearsons} = 0.199$  and  $r_{spearsons} = 0.2$ )
- However, the similarity between the values of Pearson's and SPearman's coefficients suggests that the Pearson's coefficient is fine

Conclusion: Although the association between Age and Duration of DM can be shown to be statistically significant, this may be an artefact of the large sample size as neither the Pearson's or SPearman's correlation coefficient could be shown to be higher than 0.2

## Associations between continuous outcomes and categorical predictors

- Now let's move onto the situation where we want to see whether the level of a continuous outcome, differs between two groups
- For this we can use the independent samples t-test (aka Student's t-test)
- This test examines whether there is a difference (on average) between two groups. This is:

$H_0 : \mu_a = \mu_b$  (On average groups do not differ) vs

$H_A : \mu_a \neq \mu_b$  (Yes they do)

# t-tests

## Test statistic

The test statistic for the two sample t-test is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}^2}$$

Where

$$S_{\bar{x}_1 - \bar{x}_2}^2 = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

## DMHT data for t-test

Let's use the DMHT data to test the hypothesis that the duration of T2DM is the same between those with DM alone, and those with both DM and HT:

$H_0 : \mu_{T2DM} = \mu_{DMHT}$  (Duration, on average, the same)

$H_A : \mu_{T2DM} \neq \mu_{DMHT}$  (Duration, on average, differs)

### Stata: Independent t-test

```
*Independent t-test  
ttest duradm, by(ht)
```

## Results

### Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
DM alone	1552	7.051546	.1127257	4.440877	6.830436	7.272657
DM and H	3204	7.802434	.0865383	4.898404	7.632758	7.972111
combined	4756	7.557401	.069116	4.7665	7.421902	7.692901
diff		-.7508881	.1470232		-1.039122	-.4626545
diff = mean(DM alone) - mean(DM and H)				t =	-5.1073	
Ho: diff = 0				degrees of freedom =	4754	
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) = 0.0000		Pr( T  >  t ) = 0.0000		Pr(T > t) = 1.0000		

Although we can reject  $H_0$  and conclude a statistical difference ( $t = -5.11$ ,  $p < 0.001$ ), is 0.751 of a year, a clinically important difference????

## Nonparametric test for two independent samples

If data are not normally distributed AND we have small samples ( $n_1, n_2 < 30$ ), we should use the nonparametric tests, Mann-Whitney (aka Wilcoxon SUM OF RANKS test). In Stata (and using our data):

### Stata: Mann-Whitney test

```
*Perform Mann-Whitney  
ranksum duradm, by(ht)
```

## Paired data

- If we have a natural basis for pairing in our data (e.g. Pre-post study), we need tests that account for this type of design
- The appropriate Parametric test is the **Paired t-test**
- The appropriate nonparametric test (small sample and non-normal) is the **Wilcoxon SIGNED RANK test** (Don't confuse with the Wilcoxon SUM OF RANK test)
- I won't go through an example for this, but just give you the stata syntax

## Stata: Paired t-test and Wilcoxon signed rank test

### Stata: Paired t-test and Wilcoxon signed rank test

\*Paired t-test

```
ttest preval == postval
```

\*Wilcoxon signed rank test

```
signrank preval = postval
```

Note: Your dataset will be set up differently here. The first set of values (e.g. Pre) need to sit side-by-side with the second set of values (e.g. Post). In other words your data has to be **unstacked**

## Classical (Naive) vs model based statistical analysis

- Now we will move away from these classical tests (which I hope you will never use), to a much more useful family of methods: The **Linear Models**
- Although we need a little bit more knowledge to use linear models, they have a MAJOR advantage over the classical 'bivariate' methods.
- They can consider the MULTIVARIABLE situation
- We are used to the idea of an **Outcome** and a **Study effect**, but in observational studies we also need to consider covariates (other X variables that may impact the analysis)

## The possible effect of covariates

There are three ways a covariate can impact the analysis. As an:

- 1 Independent risk factor: a risk factor that (INDEPENDANTLY) explains variation in our outcome
- 2 A Confounder: A covariable that will bias our study effect if we don't account for it (in the model)
- 3 An effect modifier: A covariate that MODIFIES the effect of the study effect on the outcome

People will often get confused about these (especially the differences between confounders and effect modifiers), but this is something I will go into a lot of detail about in the next few sessions.

Now, let's consider our first Multivariable model

## The linear regression model

In the past, you will have used Multivariable Linear Regression.  
The model for MLR takes the form:

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

Linear regression model can be articulated using matrix algebra:

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

Where

$\mathbf{Y}$  is vector of observation of our outcome variable;

$\mathbf{X}$  is a matrix containing a constant and at least one explanatory variable; and

$\boldsymbol{\beta}$  is a vector of parameters linearly relating  $\mathbf{X}$  to  $\mathbf{Y}$

## Matrix formulation of linear regression

The matrix formulation:

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

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,  $\beta_0$ , and the rest of the columns (representing the individual covariates) are associated with the individual slopes,  $\beta_1, \beta_2, \dots, \beta_{k-1}$ .

## Interpretation of coefficients, $\beta$

$\beta_0$  is **y-intercept** ( $b_0$  is the sample estimate)

**Value of Y when X = 0**

$\beta_1$  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  $\beta$ s (through to  $k - 1$ ) are also slopes and can be interpreted in the same way as  $\beta_1$ .

# Linear regression and least squares estimation

In the linear regression model we use the **Principle of Least Squares** to estimate  $\beta$

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

$$SSE = \min (\sum_{i=1}^n \epsilon^2)$$

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

# Hypothesis testing in Linear Regression

## Overall model:

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

Interpretation: No covariates ( $X$ s) explain variation in the outcome,  $y$

$$H_A : \text{At least one } \beta_j \text{ differs from zero (for } j = 1, 2, \dots, k-1)$$

Interpretation: At least one of the covariates ( $X$ s) 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  $y$

# The central role of linear regression in biostatistics

- Linear regression underpins the General (Normal) Linear Model (later this session)
- It is also fundamental to the more advanced methods:
  - **Generalized Linear Models:** Analysis of other outcomes (e.g. Binary) arising from cross sectional studies (Next session)
  - Cox Proportional Hazards regression (Survival analysis)
  - **Linear Mixed Models:** Analysis of continuous outcomes from longitudinal studies
  - **Generalized Linear Mixed Models and Generalized Estimating Equations:** Analysis of other outcomes (e.g. Binary) arising from longitudinal studies
- In other words, Linear regression is the most important and fundamental method for a strong understanding of biostatistical modelling

# Analysis of Variance

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

- ❶ A statistical method (which uses a particular and outdated model); and
  - ❷ 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 **A**nalysis **o**f **V**ariance (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:

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$  (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 an illustration of the overall (total) variation in an outcome variable (e.g. Quality of Life  $\sim$  QoL)

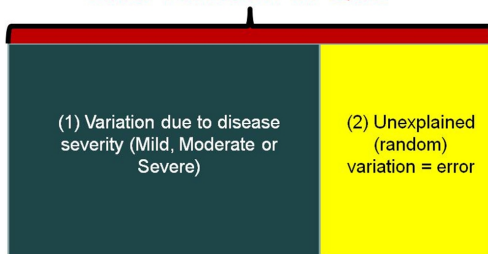


Total variation in QoL

## Analysis of Variance: ANOVA

Some of the variation in QoL might be explained by (1) Disease status (Mild, Moderate, Severe) and some is the (2) natural variation we would expect between individuals in this population

### Total variation in QoL

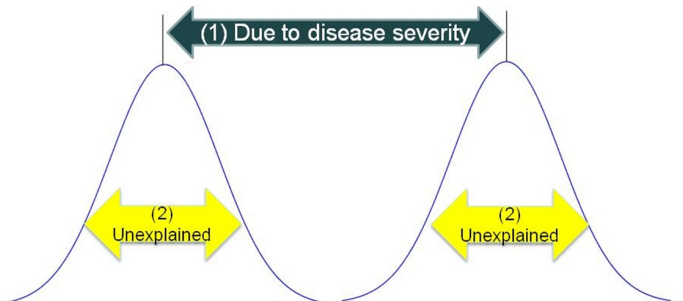


Testing:  $H_0: \mu_{mild} = \mu_{mod} = \mu_{severe}$

We would expect to reject  $H_0$  if between group differences (grey)

## Analysis of Variance: ANOVA

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

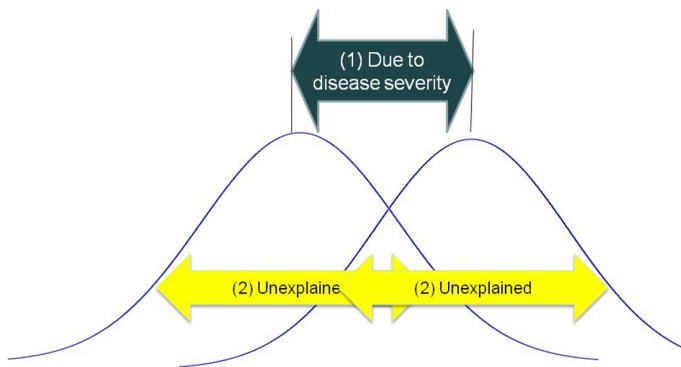


Between groups	Within groups
----------------	---------------

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:

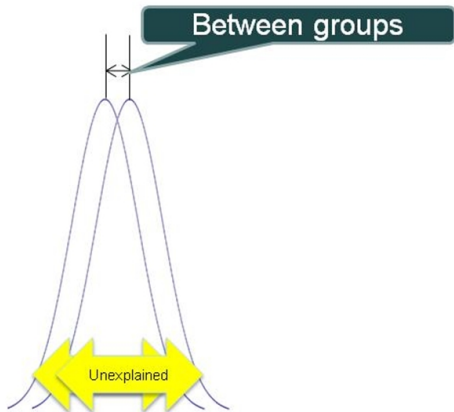


Between  
groups

Within  
groups

## Analysis of Variance: ANOVA

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



Between groups	Within groups
----------------	---------------

## Purpose of ANOVA

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

For example, Systolic Blood Pressure between different racial groups:

$$H_0: \mu_{\text{African}} = \mu_{\text{Asian}} = \mu_{\text{Caucasian}}$$

## ANOVA and Sums of Squares

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

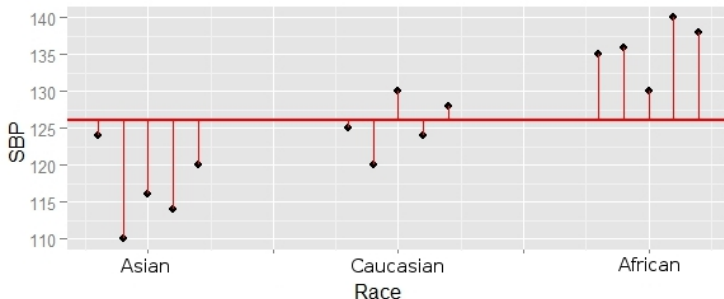
- So ANOVA is about partitioning the sums of squares ( $\cong$  variation) hence, ANALYSIS OF VARIANCE

## Sums of Squares formulae for ANOVA: $SS_{total}$

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

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

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

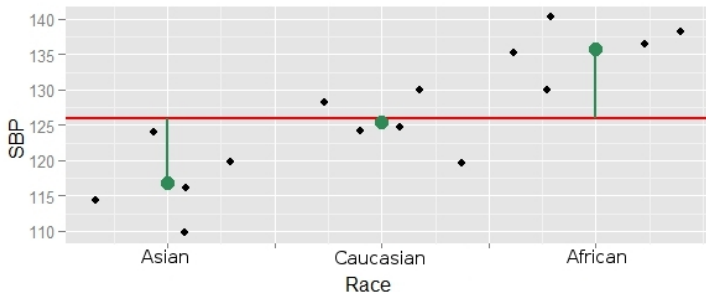


## Sums of Squares formulae for ANOVA: $SS_{between}$

The between-groups variation SS will be:

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

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

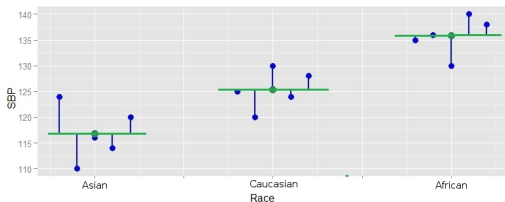


## Sums of Squares formulae for ANOVA: $SS_{within}$

What is left over is the error SS:

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

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



## Sums of Squares → 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 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:

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

i.e. The Sample Variance,  $S^2$

## $MS_{between}$ and $MS_{within}$

Now,

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

and,

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

- 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.

## The variance ratio

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:

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

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 **linear regression** is for **quantitative explanatory variables** and **ANOVA** is for **categorical explanatory variables**
- But traditional ANOVA can be extended to also incorporate covariates (quantitative explanatory variables)
- This extension is called Analysis of covariance (ANCOVA)

## 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 age as effectively making every subject the same (average) age

## ANCOVA example

So for our example....

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

or,

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

Where,

$SBP_{ij}$  is the  $j^{th}$  replicate of the  $i^{th}$  gender group

$\mu$  represents the grand (overall) mean

$\beta$  is the slope for Age; and

$Age_{ij}$  and  $\epsilon_{ij}$  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 their use of sums of squares formulae
- In particular, these methods are susceptible to differences in group sample sizes (unbalanced designs)
- Where there are even moderate differences in sample sizes, the larger of the two samples is generally disproportionately weighted

## 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, trickier to interpret the 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 (next session)

## 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 indices have been restricted to  $i$ , for  $i = 1, 2, \dots, N$
- Instead of **Gender**, we have **M** and **F** where:
  - If the patient is male  $M = 1$  (and  $F = 0$ ); and
  - If the patient is female  $F = 1$  (and  $M = 0$ )

**CAN YOU SEE ANY PROBLEMS????**

## 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 **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

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

- Where Gender = 0 for males and Gender = 1 for females.
- In this case we are using Males as the **REFERENT**, and  $\beta_{Gender}$  tells us the difference (in SBP) from males if you are female.
- Let's look at this from the point of view of the X matrix

Aside:

This system of dummy coding by assigning the referent all zeros is called **Zero cell referencing**

## 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,  $\beta_0$

## A fictitious example

Let's say we run a model (in R or Stata) and get the following parameter estimates:

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

### Interpretation:

- $\beta_0 = 100$ : The expected SBP of someone who is zero years old and has a gender of zero (Males).  $p < 0.05 \Rightarrow$  Baby male SBP (on average) is significantly different from zero (So what???)
- $\beta_{Age} = 0.333$ : As we get one year older, on average our SBP should rise by 0.333.  $p < 0.05 \Rightarrow$  Age does explain SBP
- $\beta_{Gender} = -25$ : When Gender is one (i.e. we are female) we would expect our SBP to decrease (relative to males) by 25 units.  $p < 0.05 \Rightarrow$  On average Females' SBP differs from males.

## $\beta$ 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**:

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

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

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

## Estimated marginal means

Technically,

$$\hat{\mu}_{Males} = 100 \text{ i.e. } 100 - \beta_{Gender}(0)$$

$$\hat{\mu}_{Females} = 75 \text{ i.e. } 100 - \beta_{Gender}(1)$$

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

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

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

## 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

Key point:

THIS IS VERY IMPORTANT YOU UNDERSTAND THIS!!!!!!!!!!

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

### Confounding:

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

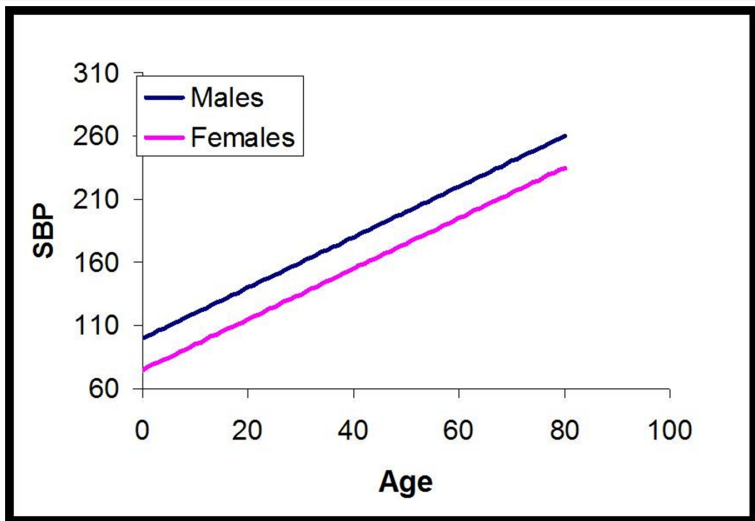
In this case the difference between our sample means ( $\bar{x}_{Males}$  and  $\bar{x}_{Females}$ ) represents:

- *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 categorical)

- 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:Caucasian**

## 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

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

$$\beta_0 = 100 \text{ (p<0.05)}; \beta_1 = 23 \text{ (p=0.12)}; \beta_2 = -35 \text{ (p<0.05)} \Rightarrow$$

$$\hat{\mu}_{Caucasian} = \beta_0 + \beta_1(0) + \beta_2(0) = 100$$

$$\hat{\mu}_{African} = \beta_0 + \beta_1(1) + \beta_2(0) = 123$$

$$\hat{\mu}_{Asian} = \beta_0 + \beta_1(0) + \beta_2(1) = 65$$

Interpretation:

- No significant difference between African and Caucasians (p=0.12)
- Asians lower SBP than Caucasian (p<0.05)
- Difference between Africans and Asians is:  $\beta_1 - \beta_2 = 58$

Aside:

Important to note that comparisons not involving the referent (e.g. African Vs Asian) need to be performed *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:

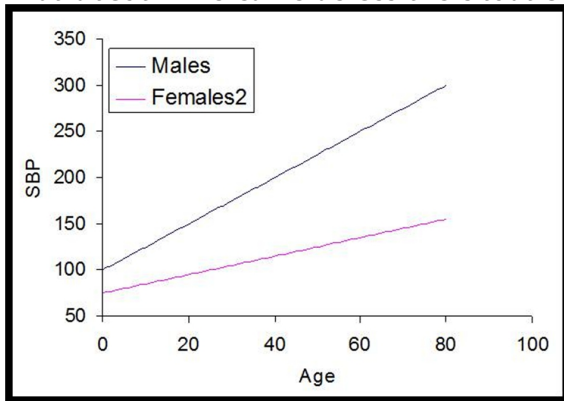
$$SBP_i = \beta_0 + \beta Age_i + \beta_{Gender} Gender_i + \epsilon_i$$

- 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

- Interactions represent a multiplicative effect of two factors (or covariates, or some combination thereof)
- 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

$$\hat{\mu}_{Males} = \beta_0 + \beta_{age}A_{ge}^{-}$$

**Females**

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

$$\hat{\mu}_{Females} = (\beta_0 + \beta_{Gend}) + (\beta_{age} + \beta_{Age \times Gend})A_{ge}^{-}$$

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

$\beta_{Gend}$  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

$$SBP_i = \beta_0 + \beta_{Age}Age_i + \beta_{D1}D1_i + \beta_{D2}D2_i \\ + \beta_{Age \times D1}(Age_i \times D1_i) + \beta_{Age \times D2}(Age_i \times D2_i) + \epsilon_i \quad (1)$$

- 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

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

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

### 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

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

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

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

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  $\beta$ s should be tested against zero** (except  $\beta_0$ , this hypothesis test would be meaningless)

## Concluding remarks

- ▶ At first glance the General Linear Model looks nasty 
- ▶ But an understanding of them is vital to truly understand the common extensions to the General Linear Model used in biostatistics including:
  - ❶ Generalized Linear Models (e.g. Logistic regression, Poisson regression etc..)
  - ❷ Survival analysis methods e.g. Cox proportional hazards regression
  - ❸ Linear Mixed Models (continuous outcomes in longitudinal studies)
  - ❹ Generalized Estimating Equations and Generalized Linear Mixed Models (categorical outcomes from longitudinal 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!!!!

## The General Linear Model in Stata

We will use the dmht5000 dataset (provided in a CSV file in classroom). We will investigate the effect of ht (the study effect: no/yes) and the covariate age (continuous) on DM duration. We will be testing the three following (sets of) hypotheses:

$H_0 : \mu_{noht} = \mu_{ht}$  (Hypertension not assoc with duration)

$H_A : \mu_{noht} \neq \mu_{ht}$

$H_0 : \beta_{age} = 0$  (Age does not explain hypertension)

$H_A : \beta_{age} \neq 0$

$H_0 : \beta_{ht \times age} = 0$  (There is no effect modification)

$H_A : \beta_{ht \times age} \neq 0$

## The General Linear Model in Stata

### Stata: General Linear Model: main effects model

```
*Read in data from CSV text file
insheet using "c:\mydata\dmht5000.csv", comma clear

*General linear model (Stata V11+)
* Note the '#' is used for interaction effects
* Need to tell stata that we want:
  *ht as a dummy variable (i.)
regress duradm i.ht age

*Get average age to calc estimated marginal means
summarize age

*For Stata V10 (or less) need to use xi: directive
xi: regress duradm i.ht age
```

# Output: Main effects model

Source	SS	df	MS	Number of obs = 4797		
Model	5258.71753	2	2629.35877	F( 2, 4794) = 130.34		
Residual	96710.3586	4794	20.1732079	Prob > F = 0.0000		
Total	101969.076	4796	21.2612752	R-squared = 0.0516		
				Adj R-squared = 0.0512		
				Root MSE = 4.4915		
duradm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.ht	.5528473	.1428762	3.87	0.000	.2727444	.8329502
age	.0887002	.0062046	14.30	0.000	.0765362	.1008641
_cons	1.861553	.3654389	5.09	0.000	1.145125	2.577981

## Interpreting results

Now:

- 1 Obtain a model for those without (referent) and those with hypertension
- 2 Using the average age (below) calculate the estimated marginal mean for each group

Variable	Obs	Mean	Std. Dev.	Min	Max
age	5000	59.761	10.72983	25	96

## Stata: General LM with interaction effect

### Stata: General Linear Model

- \*General linear model (Stata V11+)
- \* Note the '#' is used for interaction effects
- \* Tell stata age as a continuous variable (c.)
  - \*If you forget c., it will give you
  - \*inteaction for seperate ages

```
regress duradm i.ht c.age i.ht#c.age
```

# Results: Factorial model

Source	SS	df	MS	Number of obs = 4797		
Model	5300.37355	3	1766.79118	F( 3, 4793) = 87.60		
Residual	96668.7025	4793	20.1687258	Prob > F = 0.0000		
				R-squared = 0.0520		
				Adj R-squared = 0.0514		
Total	101969.076	4796	21.2612752	Root MSE = 4.491		

duradm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.ht	-.533916	.7695739	-0.69	0.488	-2.042634	.9748021
age	.0763399	.0106047	7.20	0.000	.0555498	.0971299
ht#c.age						
1	.0187917	.0130758	1.44	0.151	-.0068428	.0444262
_cons	2.553091	.6042018	4.23	0.000	1.368578	3.737604

## Interpretation

Now:

- 1 Write of the model for 1. no ht and ht groups.
- 2 What was the difference between the age slopes for each group (was this a significant difference in slope?)
- 3 Given the results of the interaction hypothesis, is it appropriate to interpret the main effects?