Factorial ANOVA: More than one categorical explanatory variable

STA312 Fall 2012

Factorial ANOVA

- Categorical explanatory variables are called **factors**
- More than one at a time
- Originally for true experiments, but also useful with observational data

- If there are observations at all combinations of explanatory variable values, it's called a *complete* factorial design (as opposed to a fractional factorial).

The potato study

- Cases are storage containers (of potatoes)
- Same number of potatoes in each container. Inoculate with bacteria, store for a fixed time period.
- Response variable is number of rotten potatoes.
- Two explanatory variables, randomly assigned
  - Bacteria Type (1, 2, 3)
  - Temperature (1=Cool, 2=Warm)

Two-factor design

	Bacteria Type		
Temp	1	2	3
1=Cool			
2=Warm			

Six treatment conditions

Factorial experiments

- Allow more than one factor to be investigated in the same study: Efficiency!
- Allow the scientist to see whether the effect of an explanatory variable *depends* on the value of another explanatory variable: Interactions
- Thank you again, Mr. Fisher.

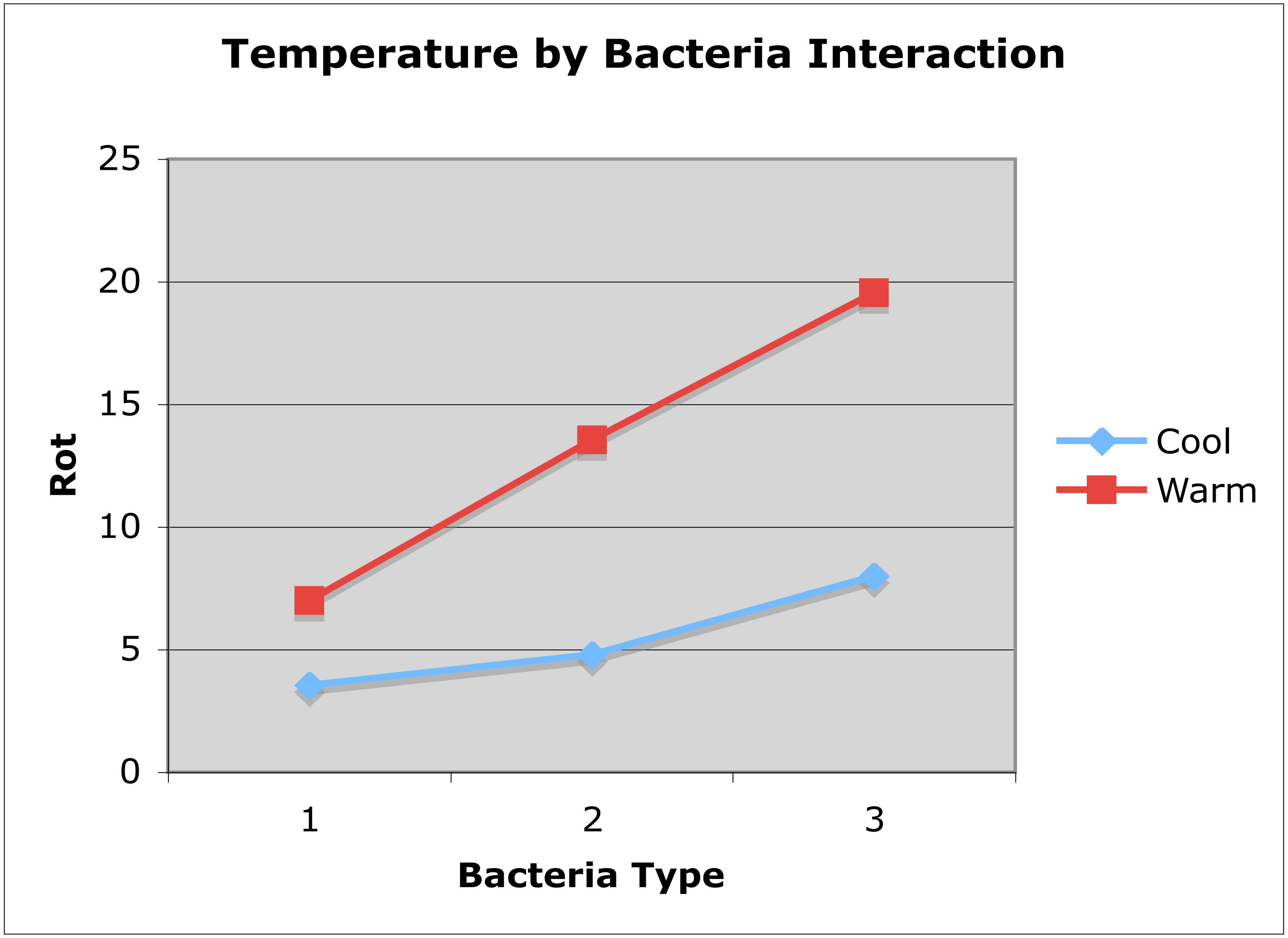
Normal with equal variance and conditional (cell) means $\mu_{i,j}$

	Bacteria Type			
Temp	1	2	3	
1=Cool	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$	$\frac{\mu_{1,1}+\mu_{1,2}+\mu_{1,3}}{3}$
2=Warm	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$	$\frac{\mu_{2,1}+\mu_{2,2}+\mu_{2,3}}{3}$
	$\frac{\mu_{1,1}+\mu_{2,1}}{2}$	$\frac{\mu_{1,2}+\mu_{2,2}}{2}$	$\frac{\mu_{1,3}+\mu_{2,3}}{2}$	μ

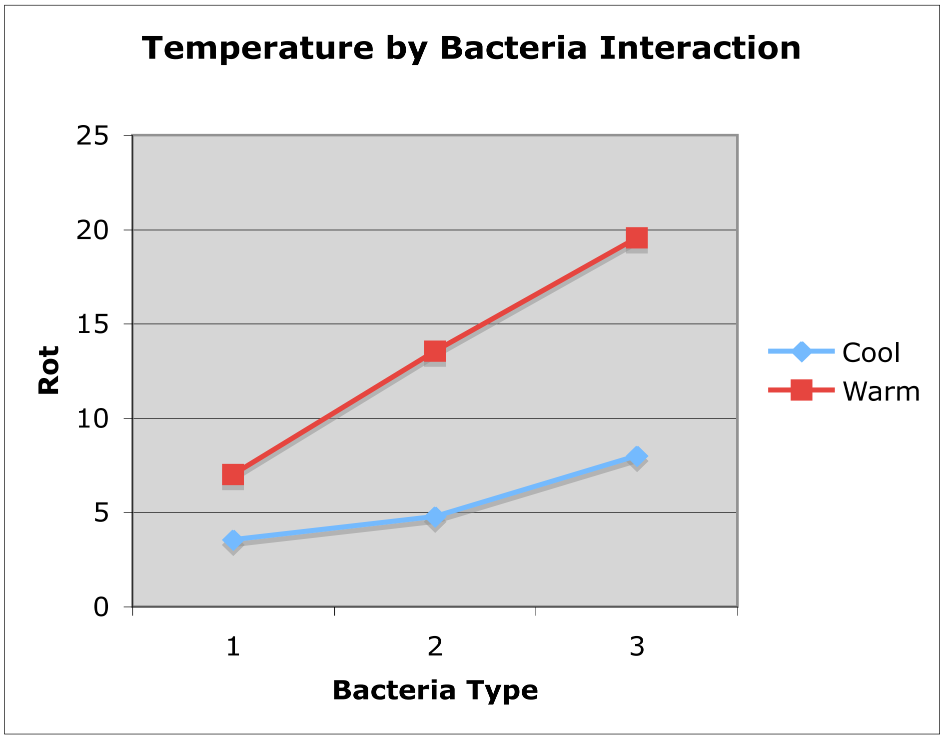
Tests

- Main effects: Differences among marginal means
- Interactions: Differences between differences (What is the effect of Factor A? **It depends** on Factor B.)

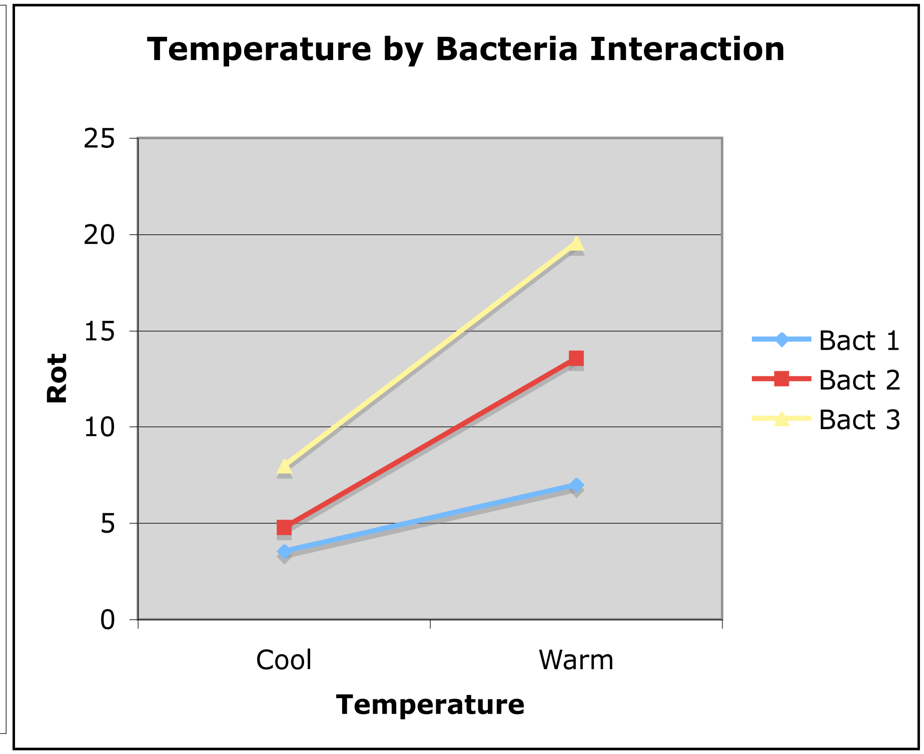
To understand the interaction, plot the means



Either Way

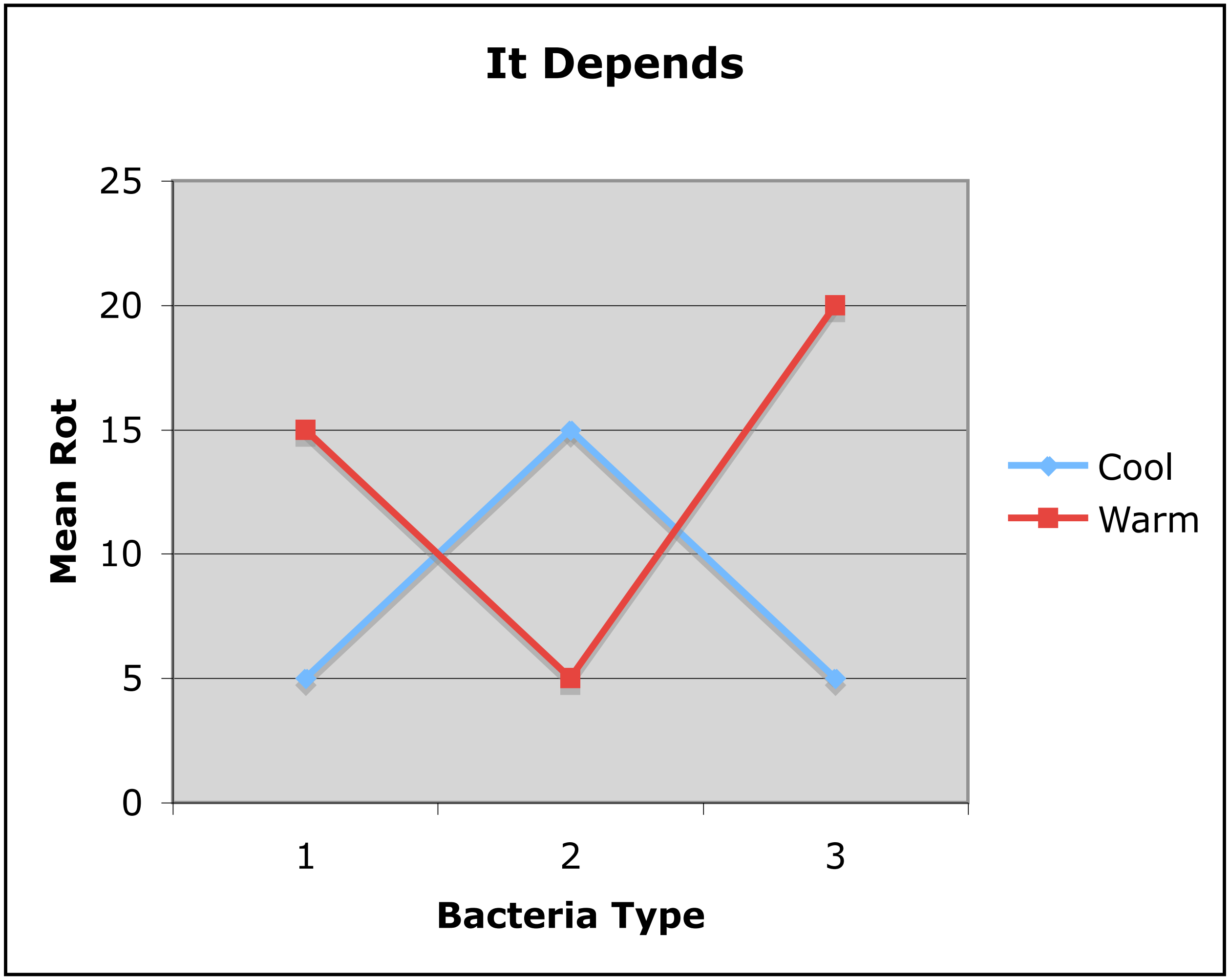


Temperature by Bacteria Interaction
25
20
15
10
5
0
Rot
1
2
3
Bacteria Type
Cool
Warm

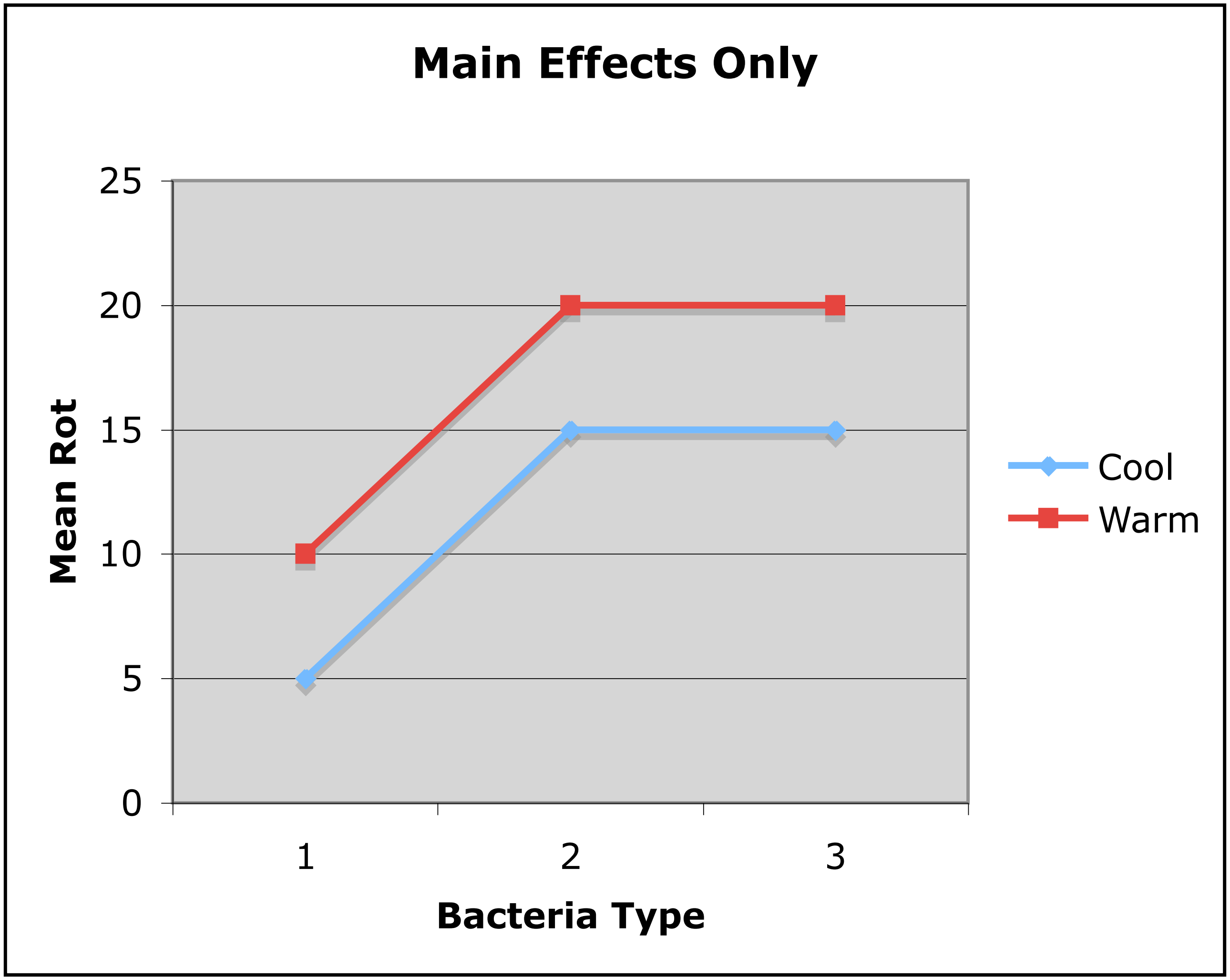


Temperature by Bacteria Interaction
25
20
15
10
5
0
Rot
Cool
Warm
Temperature
Bact 1
Bact 2
Bact 3

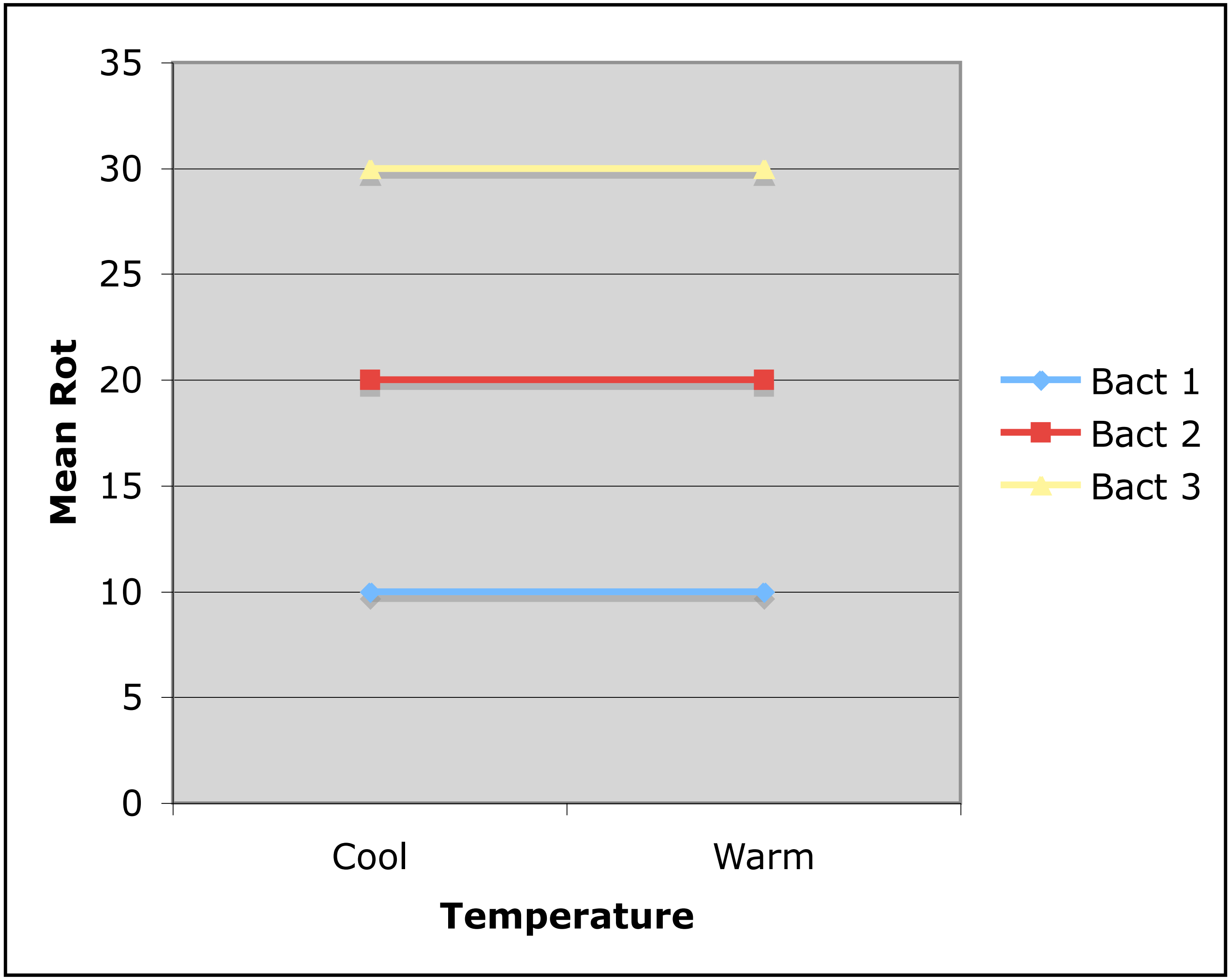
Non-parallel profiles = Interaction



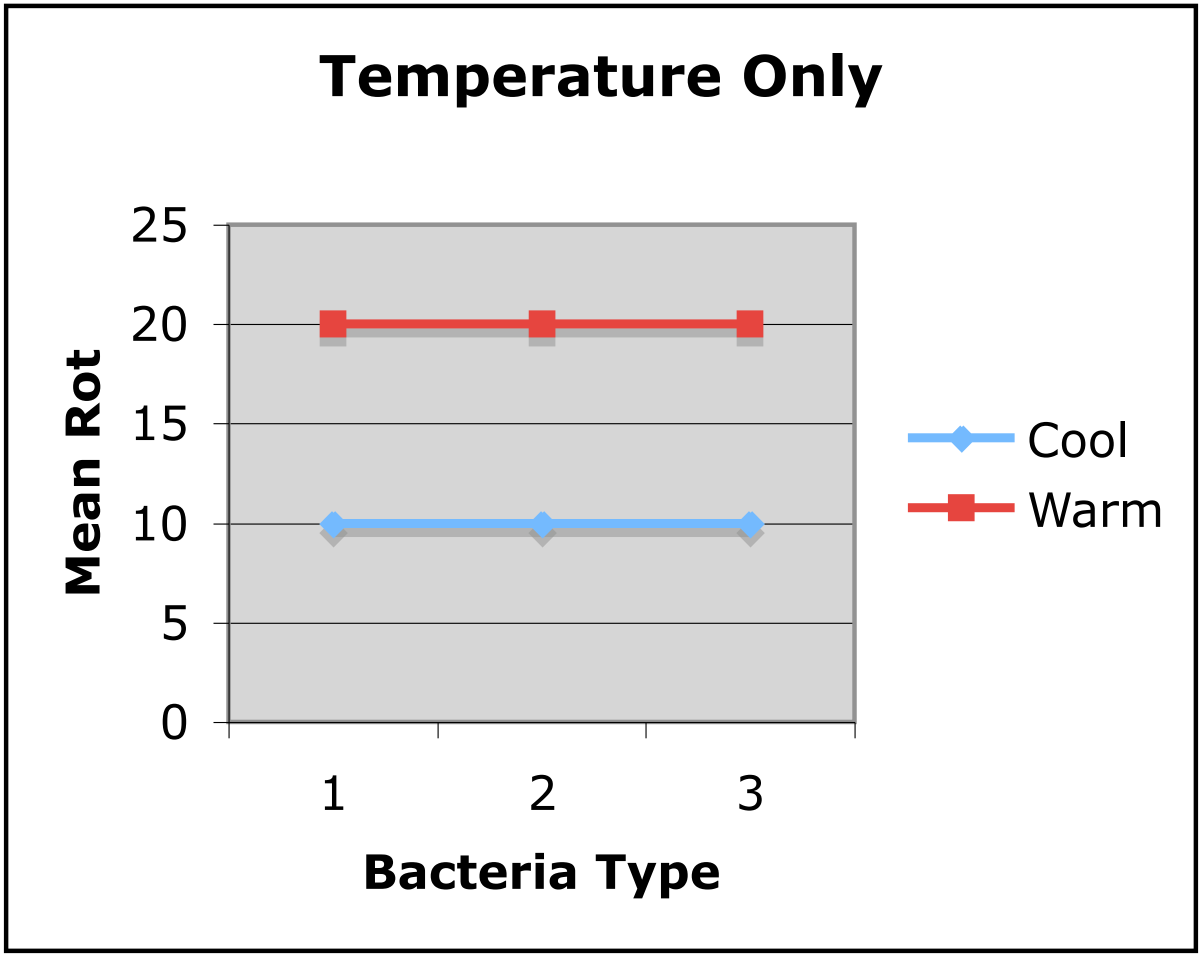
Main effects for both variables, no interaction



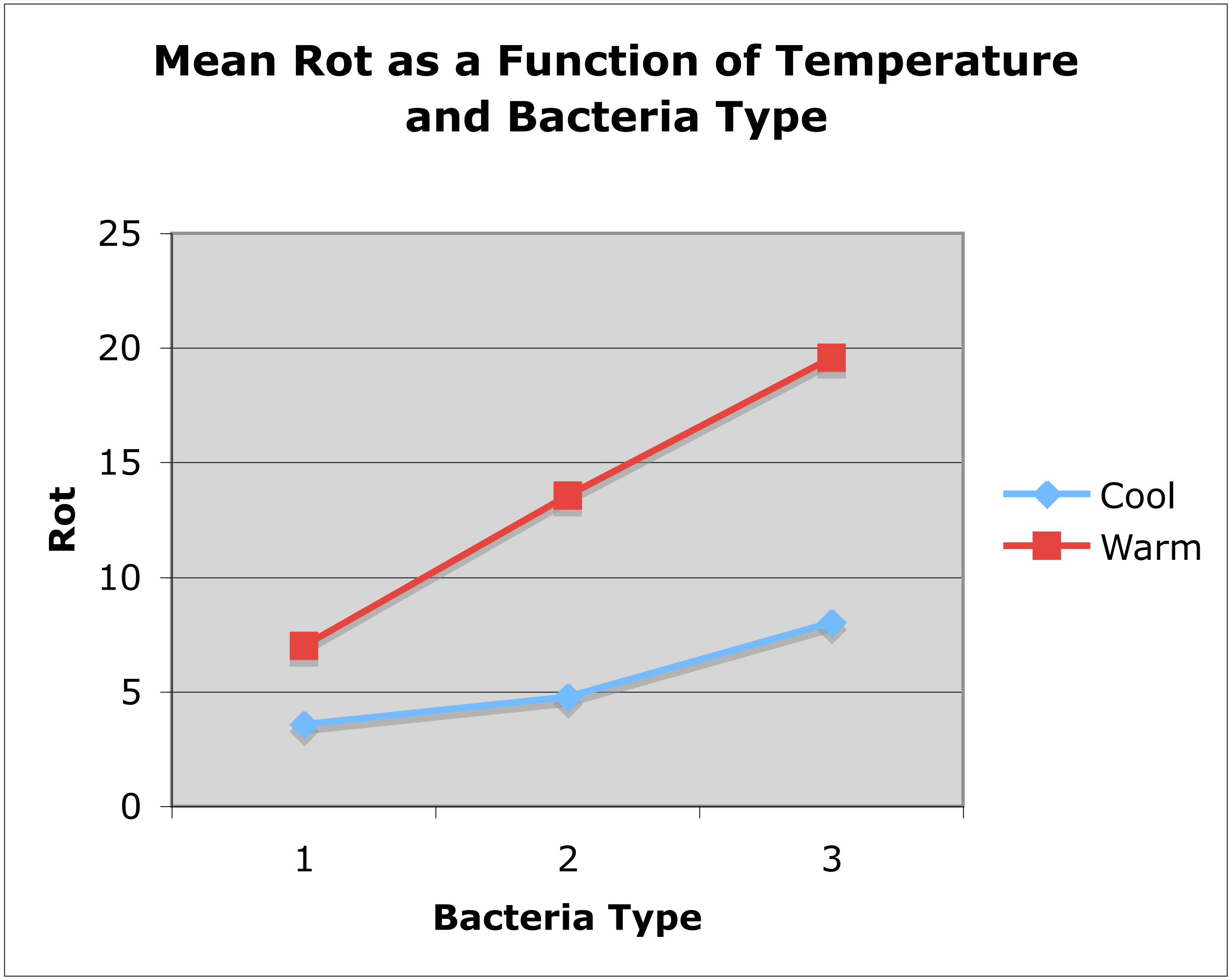
Main effect for Bacteria only



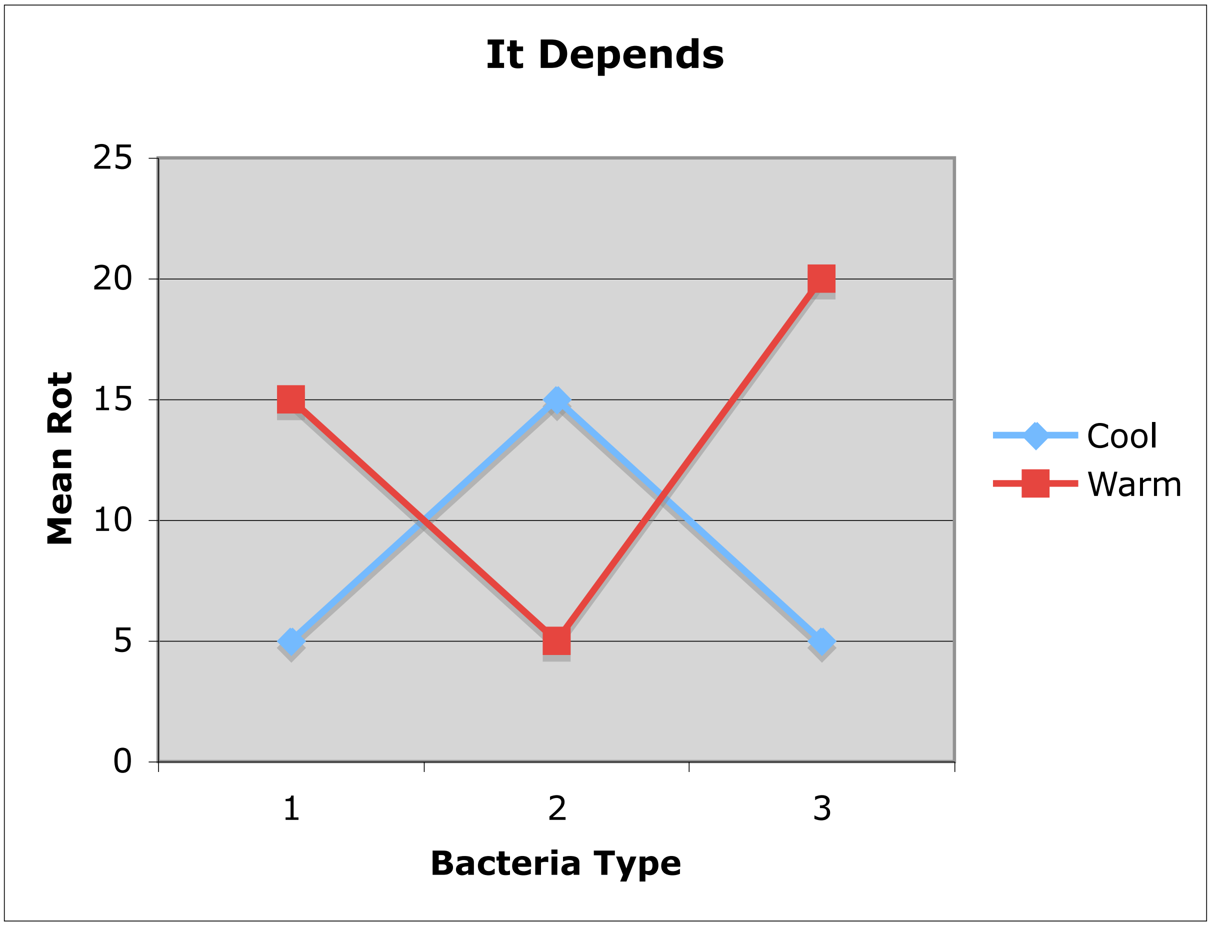
Main Effect for Temperature Only



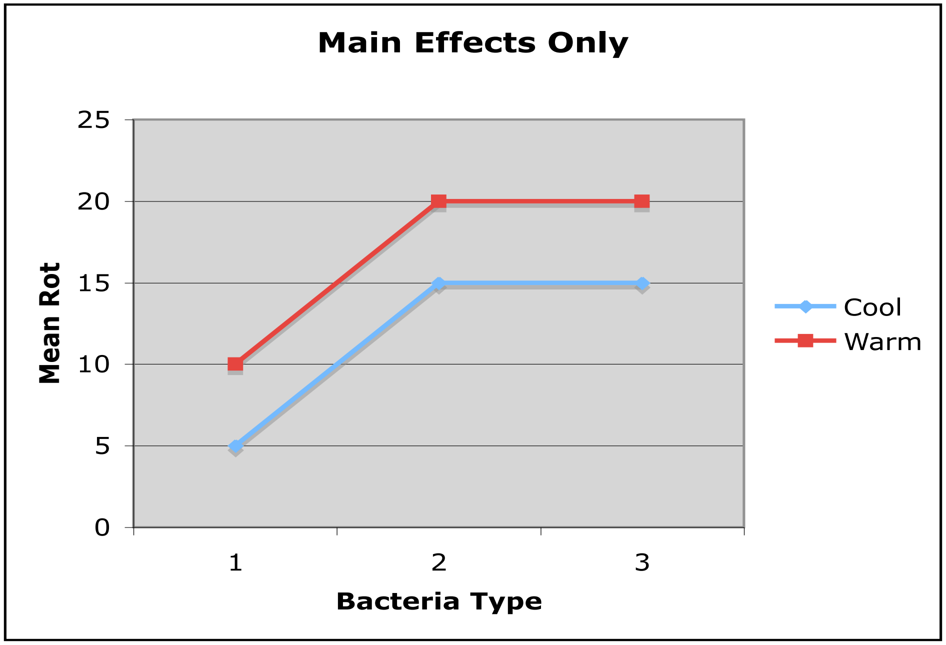
Both Main Effects, and the Interaction



Should you interpret the main effects?



Testing for Interactions



- $H_0 : \mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3}$

- $H_0 : \mu_{1,2} - \mu_{1,1} = \mu_{2,2} - \mu_{2,1}$ and

  $\mu_{1,3} - \mu_{1,2} = \mu_{2,3} - \mu_{2,2}$

Equivalent statements

- The effect of A depends upon B
- The effect of B depends on A

$$H_0 : \mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3}$$

$$H_0 : \mu_{1,2} - \mu_{1,1} = \mu_{2,2} - \mu_{2,1} \text{ and}$$
$$\mu_{1,3} - \mu_{1,2} = \mu_{2,3} - \mu_{2,2}$$

Three factors: A, B and C

- There are three (sets of) main effects: One each for A, B, C
- There are three two-factor interactions
  - A by B (Averaging over C)
  - A by C (Averaging over B)
  - B by C (Averaging over A)
- There is one three-factor interaction: AxBxC

Meaning of the 3-factor interaction

- The form of the A x B interaction depends on the value of C
- The form of the A x C interaction depends on the value of B
- The form of the B x C interaction depends on the value of A
- These statements are equivalent. Use the one that is easiest to understand.

To graph a three-factor interaction

- Make a separate mean plot (showing a 2-factor interaction) for each value of the third variable.

- In the potato study, a graph for each type of potato

Four-factor design

- Four sets of main effects
- Six two-factor interactions
- Four three-factor interactions
- One four-factor interaction: The nature of the three-factor interaction depends on the value of the 4th factor
- There is an F test for each one
- And so on …

As the number of factors increases

- The higher-way interactions get harder and harder to understand
- All the tests are still tests of differences between differences of differences …
- But it gets complicated
- Effect coding to the rescue

Effect coding

Bact	B_1	B_2
1	1	0
2	0	1
3	-1	-1

Temperature	T
1=Cool	1
2=Warm	-1

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

Interaction effects are products of dummy variables

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

- The A x B interaction: Multiply each dummy variable for A by each dummy variable for B
- Use these products as additional explanatory variables in the multiple regression
- The A x B x C interaction: Multiply each dummy variable for C by each product term from the A x B interaction
- Test the sets of product terms simultaneously

Make a table

$$E(Y|\mathbf{X}=\mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

Bact	Temp	B_1	B_2	T	B_1T	B_2T	$E(Y\|\mathbf{X}=\mathbf{x})$
1	1	1	0	1	1	0	$\beta_0 + \beta_1 + \beta_3 + \beta_4$
1	2	1	0	-1	-1	0	$\beta_0 + \beta_1 - \beta_3 - \beta_4$
2	1	0	1	1	0	1	$\beta_0 + \beta_2 + \beta_3 + \beta_5$
2	2	0	1	-1	0	-1	$\beta_0 + \beta_2 - \beta_3 - \beta_5$
3	1	-1	-1	1	-1	-1	$\beta_0 - \beta_1 - \beta_2 + \beta_3 - \beta_4 - \beta_5$
3	2	-1	-1	-1	1	1	$\beta_0 - \beta_1 - \beta_2 - \beta_3 + \beta_4 + \beta_5$

Cell and Marginal Means

	Bacteria Type			
Tmp	1	2	3	
1=C	$\beta_0 + \beta_1 + \beta_3 + \beta_4$	$\beta_0 + \beta_2 + \beta_3 + \beta_5$	$\beta_0 - \beta_1 - \beta_2$ $+\beta_3 - \beta_4 - \beta_5$	β_0 $+\beta_3$
2=W	$\beta_0 + \beta_1 - \beta_3 - \beta_4$	$\beta_0 + \beta_2 - \beta_3 - \beta_5$	$\beta_0 - \beta_1 - \beta_2$ $-\beta_3 + \beta_4 + \beta_5$	β_0 $-\beta_3$
	$\beta_0 + \beta_1$	$\beta_0 + \beta_2$	$\beta_0 - \beta_1 - \beta_2$	β_0

We see

- Intercept is the grand mean
- Regression coefficients for the dummy variables are deviations of the marginal means from the grand mean
- What about the interactions?

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

A bit of algebra shows

$$\mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} \text{ is equivalent to } \beta_4 = \beta_5$$

$$\mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3} \text{ is equivalent to } \beta_4 = -\beta_5$$

$$\text{So } \beta_4 = \beta_5 = 0$$

What are “effects?”

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

- **There are 3 main effects for Bacteria** Type: beta1, beta2 and -beta1 -beta2.
- They are deviations of the marginal means from the grand mean.
- **There are 2 main effects for Temperature**: beta3 and - beta3
- They are deviations of the marginal means from the grand mean.
- **There are 6 interaction effects**.
- They are deviations of the cell mean from the grand mean plus the main effects.
- They add to zero across rows and across columns.
- The non-redundant ones are beta4 and beta5.

- This is regression notation. There are ANOVA notations as well.

Factorial ANOVA with effect coding is pretty automatic

- You don't have to make a table unless asked
- It always works as you expect it will
- Significance tests are the same as testing sets of contrasts
- Covariates present no problem. Main effects and interactions have their usual meanings, "controlling" for the covariates.
- Could plot the least squares means

Again

- Intercept is the grand mean
- Regression coefficients for the dummy variables are deviations of the marginal means from the grand mean
- Test of main effect(s) is test of the dummy variables for a factor.
- Interaction effects are products of dummy variables.

Balanced vs. Unbalanced Experimental Designs

- Balanced design: Cell sample sizes are proportional (maybe equal)
- Explanatory variables have zero relationship to one another
- Numerator SS in ANOVA are independent
- Everything is nice and simple
- Most experimental studies are designed this way.
- As soon as somebody drops a test tube, it's no longer true

Analysis of unbalanced data

- When explanatory variables are related, there is potential ambiguity.
- A is related to Y, B is related to Y, and A is related to B.
- Who gets credit for the portion of variation in Y that could be explained by either A or B?
- With a regression approach, whether you use contrasts or dummy variables (equivalent), the answer is **nobody**.
- Think of full, reduced models.
- Equivalently, general linear test

Some software is designed for balanced data

- The special purpose formulas are much simpler.
- Very useful *in the past*.
- Since most data are at least a little unbalanced, a recipe for trouble.
- Most textbook data are balanced, so they cannot tell you what your software is really doing.
- R's anova and aov functions are designed for balanced data, though anova applied to lm objects can give you what you want if you use it with care.
- SAS proc glm is much more convenient. SAS proc anova is for balanced data.

Rotten potatoes with R

```
> spuds = read.table("http://www.utstat.toronto.edu/~brunner/312f12
                      /code_n_data/potato2.data")
> attach(spuds)
> bact = factor(Bact); temp = factor(Temp)
> # Table of means
> meanz = tapply(Rot,INDEX=list(temp,bact),FUN=mean); meanz

         1         2        3
1 3.555556  4.777778  8.00000
2 7.000000 13.555556 19.55556
```

```
> # Make it prettier
> Labels = NULL # Make an empty list for row, col labels
> Labels$Temp = c("Low","High")
> Labels$Bacteria = c("1","2","3")
> dimnames(meanz) = Labels
> # Could use rownames, colnames instead
> meanz = addmargins(meanz,FUN=mean) # Add marginal means
> meanz = round(meanz,2) # Round to 2 decimal places
> meanz
```

```
      Bacteria
Temp      1     2     3  mean
  Low  3.56  4.78  8.00  5.44
  High 7.00 13.56 19.56 13.37
  mean 5.28  9.17 13.78  9.41
```

Two-factor ANOVA

```
> # Two-factor ANOVA
> table(temp,bact)


    bact
temp 1 2 3
   1 9 9 9
   2 9 9 9


> # Balanced design. aov is safe
> summary(aov(Rot ~ temp + bact + temp:bact))


            Df Sum Sq Mean Sq F value   Pr(>F)
temp         1  848.1   848.1  38.614 1.18e-07 ***
bact         2  651.8   325.9  14.839 9.61e-06 ***
temp:bact    2  152.9    76.5   3.481   0.0387 *
Residuals   48 1054.2    22.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Get same results with  Rot ~ temp*bact
```

One more comment about the potatoes

Note that aov is smart enough to produce the right tests even with indicator dummy variables. If you wanted to reproduce the tests for main effects with regression you'd use effect coding.

More about Interactions

- Interaction between independent variables means “It depends.”
- Relationship between one explanatory variable and the response variable *depends* on the value of another explanatory variable.
- Can have
  - Quantitative by quantitative
  - Quantitative by categorical
  - Categorical by categorical

Quantitative by Quantitative

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\ E(Y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \end{aligned}$$

For fixed x_2

$$E(Y|\mathbf{x}) = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

Both slope and intercept depend on value of x_2

And for fixed x_1, slope and intercept relating x_2 to E(Y) depend on the value of x_1

Quantitative by Categorical

- Interaction means slopes are not equal
- Form a product of quantitative variable by each dummy variable for the categorical variable
- For example, three treatments and one covariate: x_1 is the covariate and x_2, x_3 are dummy variables

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
&\quad + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon
\end{aligned}
$$

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

What null hypothesis would you test for

- Equal slopes
- Compare slopes for group one vs three
- Compare slopes for group one vs two
- Equal regressions
- Interaction between group and x_1

What to do if H_0: $\beta_4 = \beta_5 = 0$ is rejected

- How do you test Group "controlling" for x_1?
- A good choice is to set x_1 to its sample mean, and compare treatments at that point.

- How about setting x_1 to sample mean of the group (3 different values)?
- With random assignment to Group, all three means just estimate $E(X_1)$, and the mean of all the x_1 values is a better estimate.

Copyright Information