Lecture 9: Interactions

Bob O'Hara

bob.ohara@ntnu.no

Before we start. . .

Categorical Variables, aka Factors

Start by looking again at 2 variables

Two categorical Variables

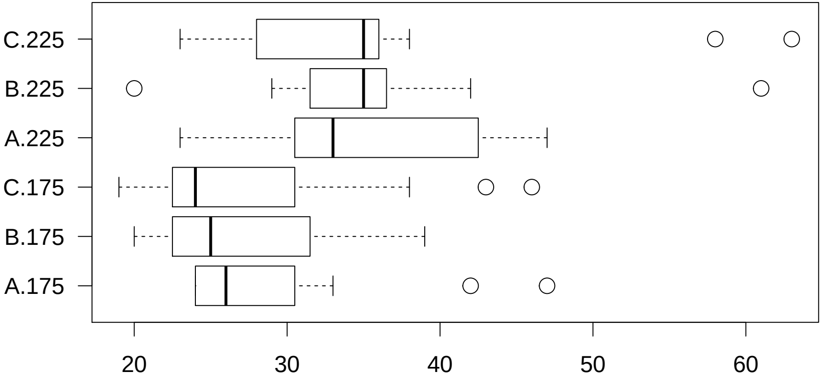
We do not have to only consider one variable: just like multiple regression, we can consider several.

e.g. with the cakes, temperature was also a factor that was controlled

```
 # use temperatures of 175C & 225C
Cakes2 <- cake[cake$temp%in%c("175", "225") ,]
```

Two categorical Variables

```r
par(mar=c(4.1,6,1,1), cex=1.5)
boxplot(Cakes2$angle~Cakes2$recipe + Cakes2$temp,
        horizontal=TRUE, las=1)
```



Two categorical Variables

The contrasts for the temperature can be written in the same way

```
Cakes2$tempF <- factor(Cakes2$temp)
contrasts(Cakes2$tempF)
```

```
    225
175   0
225   1
```

Two categorical Variables

The design matrix is now like this:

```
head(model.matrix(~recipe + tempF, data=Cakes2))
```

```
   (Intercept) recipeB recipeC tempF225
1            1       0       0        0
6            1       0       0        1
7            1       1       0        0
12           1       1       0        1
13           1       0       1        0
18           1       0       1        1
```

We now just have an extra column for temperature

Another way of writing the model

In matrix form (and back soon to something simpler)

$$\mathbf{Y} = X\beta + \varepsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{1,2} \\ \beta_{1,3} \\ \beta_{2,2} \\ \vdots \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \vdots \end{pmatrix}$$

Parameters

Where

- β_0 is the intercept
- $\beta_{1,2}$ and $\beta_{1,3}$ are the contrasts between the control and treatments 2 & 3 (hot chocolate & suger)
- $\beta_{2,2}$ is the cintrast between 225° and 175°

We could write this with $\beta_{1,1} = \beta_{2,1} = 0$ (the contrast of a treatment with itself)

Multiplying the matrices

Doing the matrix multiplication we get

$$y_1 = 1 \cdot \beta_0 + 0 \cdot \beta_{1,2} + 0 \cdot \beta_{1,3} + 0 \cdot \beta_{2,2} + \varepsilon_1 = \beta_0 + \varepsilon_1$$

for the first row, and for the rest...

$$\begin{aligned}
y_2 &= 1 \cdot \beta_0 + 0 \cdot \beta_{1,2} + 0 \cdot \beta_{1,3} + 1 \cdot \beta_{2,2} + \varepsilon_2 &&= \beta_0 + \beta_{2,2} + \varepsilon_2 \\
y_3 &= 1 \cdot \beta_0 + 1 \cdot \beta_{1,2} + 0 \cdot \beta_{1,3} + 0 \cdot \beta_{2,2} + \varepsilon_3 &&= \beta_0 + \beta_{1,2} + \varepsilon_3 \\
y_4 &= 1 \cdot \beta_0 + 1 \cdot \beta_{1,2} + 0 \cdot \beta_{1,3} + 1 \cdot \beta_{2,2} + \varepsilon_4 &&= \beta_0 + \beta_{1,2} + \beta_{2,2} + \varepsilon_4 \\
y_5 &= 1 \cdot \beta_0 + 0 \cdot \beta_{1,2} + 1 \cdot \beta_{1,3} + 0 \cdot \beta_{2,2} + \varepsilon_5 &&= \beta_0 + \beta_{1,3} + \varepsilon_5 \\
y_6 &= 1 \cdot \beta_0 + 0 \cdot \beta_{1,2} + 1 \cdot \beta_{1,3} + 1 \cdot \beta_{2,2} + \varepsilon_6 &&= \beta_0 + \beta_{1,3} + \beta_{2,2} + \varepsilon_6
\end{aligned}$$

- every data point has β_0 and 0 or 1 $\beta_{1,\cdot}$ and 0 or 1 $\beta_{2,\cdot}$

Two categorical Variables: Fitting the model

We can fit the model, just like last time!

```
mod.2way <- lm(angle~recipe + tempF, data=Cakes2)
print(summary(mod.2way)$coefficients, digits=2)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    28.41        1.7   16.48  2.4e-28
recipeB        -1.03        2.1   -0.49  6.3e-01
recipeC        -0.27        2.1   -0.13  9.0e-01
tempF225        7.38        1.7    4.28  4.9e-05
```

Little effect of recipe, big effect of temperature

- higher temperature means cake bends more before breaking

Why the intercept?

What happens if we remove the intercept?

```
mod.2way <- lm(angle~recipe + tempF-1, data=Cakes2)
print(summary(mod.2way)$coefficients, digits=2)
```

```
         Estimate Std. Error t value Pr(>|t|)
recipeA      28.4        1.7    16.5  2.4e-28
recipeB      27.4        1.7    15.9  2.7e-27
recipeC      28.1        1.7    16.3  4.5e-28
tempF225      7.4        1.7     4.3  4.9e-05
```

We get 3 levels of recipe, but still only one of tempF

Why the intercept?

Without the intercept, recipe has 3 parameters that need to be estimated: $\beta^*_{1,1}$, $\beta^*_{1,2}$, $\beta^*_{1,3}$

- these are the means for the 3
- with the intercept we have β_0, $\beta_{1,2}$, and $\beta_{1,3}$

For the one-way model, this is fine. With the two-way model and no intercept, we would have

- $\beta^*_{1,1}$, $\beta^*_{1,2}$, $\beta^*_{1,3}$ for recipe and
- $\beta^*_{2,1}$, $\beta^*_{2,2}$, for temperature.

Identifiability

The means for each combination are

$$
\begin{aligned}
\mu_{1,1} &= \beta^*_{1,1} + \beta^*_{2,1} \\
\mu_{1,2} &= \beta^*_{1,1} + \beta^*_{2,2} \\
\mu_{2,1} &= \beta^*_{1,2} + \beta^*_{2,1} \\
\mu_{2,2} &= \beta^*_{1,2} + \beta^*_{2,2} \\
\mu_{3,1} &= \beta^*_{1,3} + \beta^*_{2,1} \\
\mu_{2,2} &= \beta^*_{1,3} + \beta^*_{2,2}
\end{aligned}
$$

The problem now is that I can add C to every $\beta^*_{1,\cdot}$ and subtract C from every $\beta^*_{2,\cdot}$ and get the same μ

Thus we need at least one constraint, hence the second factor is set up as a contrast

Categorical and Continuous Data

The models for categorical data have been built by using the continuous data model, so thre is no reason why we cannot combine them.

For example

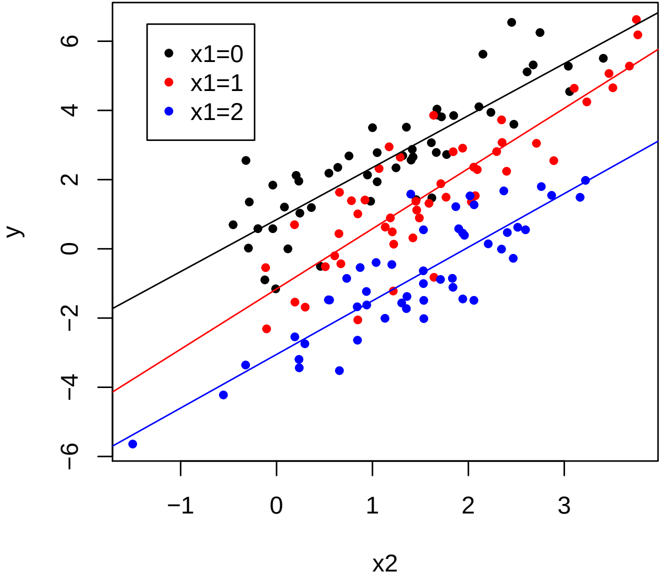
$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 2.4 \\ 1 & 0 & 0 & 8.3 \\ 1 & 1 & 0 & 9.7 \\ 1 & 1 & 0 & 4.6 \\ 1 & 0 & 1 & 2.3 \\ 1 & 0 & 1 & 1.0 \\ \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{1,2} \\ \beta_{1,3} \\ \beta_2 \\ \vdots \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \vdots \end{pmatrix}$$

So β_2 is the slope of the effect of the continuous covariate.

Simulate Some Data

```r
Cols=c(1,2,4)
x1 <- factor(rep(1:3, each=50))
x2 <- rnorm(length(x1), 1.5,1)
mu <- 1 - 2*(x1==2) -4*(x1==3) + (1.5)*x2
y <- rnorm(length(mu), mu, 1)
```

Plot Some Data



```
## [[1]]
## NULL
```

Fitting the model

```
mod.anc <- lm(y ~x1 + x2)
round(summary(mod.anc)$coefficients,3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.731      0.165   4.432        0
## x12           -1.659      0.196  -8.472        0
## x13           -3.855      0.194 -19.918        0
## x2             1.600      0.079  20.307        0
```

- x12 is the contrast between levels 1 & 2 of x1
- x2 is the slope of the relationship with x2

Interactions

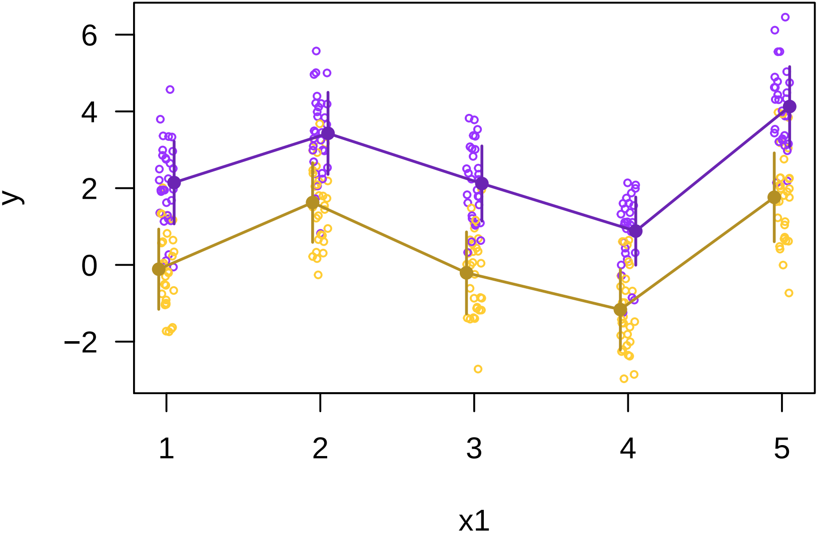
So far we have assumed that all effects act independently

e.g.

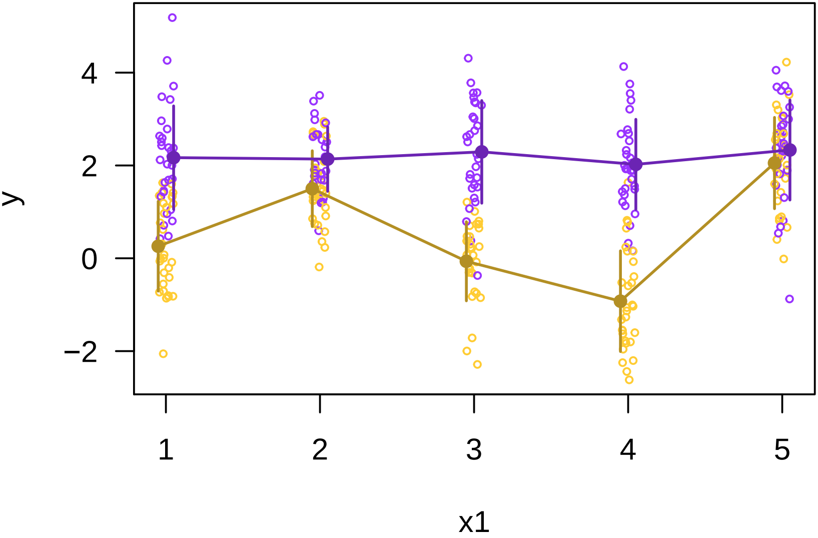
$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

But this makes some strict assumptions about the model

No Interactions



But sometimes. . . .



The effect of X1 depends on X2

Interactions

These are called *interactions*

- this is a “first order interaction”
- can get more complicated

Interactions in R

These are (too) easy to code. If we want the interaction between X1 and X2, we use

```r
lm(Y ~ X1*X2)
```

we can also have interactions between 3 variables:

```r
lm(Y ~ X1*X2*X3)
```

. . . or even more

Estimates

```r
mod.2way <- lm(y ~ x$x1*x$x2)
summary(mod.2way)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|
## (Intercept)  0.2604997  0.1781632  1.462141 1.447855e-0
## x$x12        1.2391995  0.2519609  4.918222 1.465173e-0
## x$x13       -0.3249573  0.2519609 -1.289713 1.981778e-0
## x$x14       -1.1845008  0.2519609 -4.701130 4.004891e-0
## x$x15        1.7915995  0.2519609  7.110626 9.033426e-1
## x$x21        1.9081034  0.2519609  7.573015 4.904665e-1
## x$x12:x$x21 -1.2707679  0.3563265 -3.566302 4.233650e-0
## x$x13:x$x21  0.4491638  0.3563265  1.260540 2.084879e-0
## x$x14:x$x21  1.0389884  0.3563265  2.915833 3.824253e-0
## x$x15:x$x21 -1.6269455  0.3563265 -4.565885 7.358989e-0
```

Interpretation

```r
round(summary(mod.2way)$coefficients[c(1:3, 6:8),], 2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.26       0.18    1.46     0.14
## x$x12           1.24       0.25    4.92     0.00
## x$x13          -0.32       0.25   -1.29     0.20
## x$x21           1.91       0.25    7.57     0.00
## x$x12:x$x21    -1.27       0.36   -3.57     0.00
## x$x13:x$x21     0.45       0.36    1.26     0.21
```

- x1=0, x2=0: (Intercept)
- x1=1, x2=0: (Intercept) + x1
- x1=0, x2=1: (Intercept) . . . . . + x2
- x1=1, x2=1: (Intercept) + x1 + x2 + x1:x2

Why the colon?

When we write A*B, R interprets it as the effect of A + the effect of B + the interaction

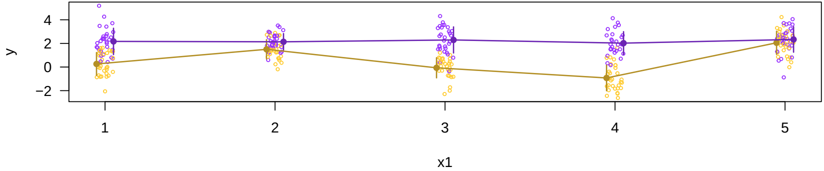
i.e. as A + B + A:B

So A:B is (just) the interaction

If we fit a model with just A:B, we assume that the average effect of A (and B) is exactly 0. This is almost never true (I have seen it once, but I had to prove it)

Plot

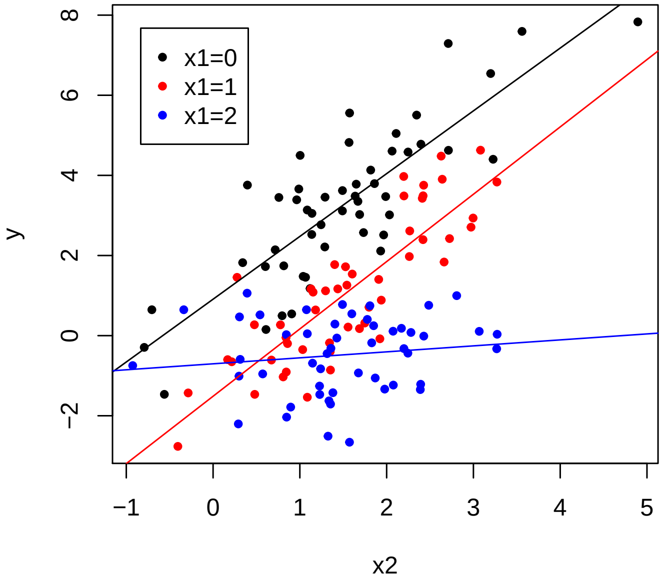
```
plot(jitter(as.numeric(x$x1), amount=0.05), y,
     col=c("#ffcc33", "#9933ff")[1+(x$x2==1)], pch=1, cex=
points(Mean$at.x, Mean$x, col=c("#b38f24", "#6b24b3")[1+(x
segments(Mean$at.x, Mean$x+SD$x, Mean$at.x, Mean$x-SD$x, c
lines(Mean$at.x[Mean$x2==0], Mean$x[Mean$x2==0], col="#b38f
lines(Mean$at.x[Mean$x2==1], Mean$x[Mean$x2==1], col="#6b24
```



Interactions with continuous variables

```
Cols=c(1,2,4)
x1 <- factor(rep(1:3, each=50))
x2 <- rnorm(length(x1), 1.5,1)
mu <- 1 - 2*(x1==2) -2*(x1==3) + (1.5 - 1.3*(x1==3))*x2
y <- rnorm(length(mu), mu, 1)
```

Plot Some Data



Fitting the model

```r
mod.anc <- lm(y ~x1*x2)
round(summary(mod.anc)$coefficients,3)
```

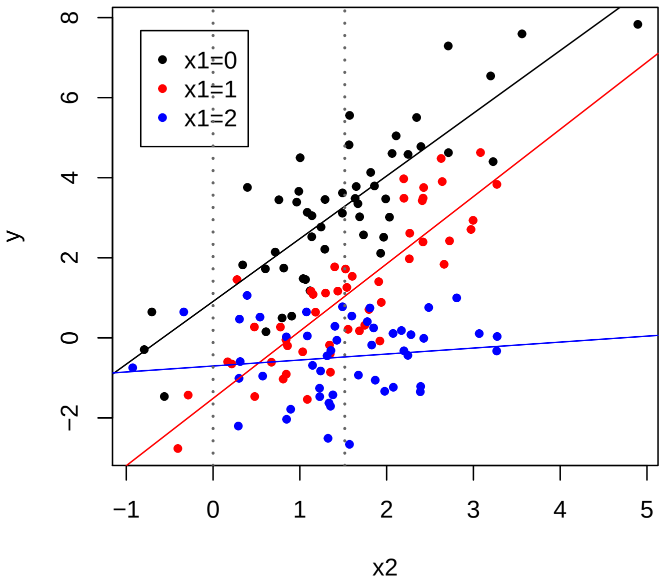
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.909      0.255   3.571    0.000
## x12           -2.423      0.386  -6.270    0.000
## x13           -1.613      0.382  -4.222    0.000
## x2             1.567      0.141  11.131    0.000
## x12:x2         0.114      0.214   0.530    0.597
## x13:x2        -1.418      0.217  -6.534    0.000
```

A problem of Interpretation

```r
mod.anc <- lm(y ~x1*x2)
mod.ancS <- lm(y ~x1*scale(x2, scale=FALSE))
round(cbind(Orig=coef(mod.anc), Centred=coef(mod.ancS)),3)
```

```
##                Orig Centred
## (Intercept)  0.909   3.288
## x12         -2.423  -2.250
## x13         -1.613  -3.764
## x2           1.567   1.567
## x12:x2       0.114   0.114
## x13:x2      -1.418  -1.418
```

Plot Again



Nested Effects

Another Complication!

An interaction means that any observation can take any value of all levels

- all recipes can be cooked at all temperatures

But some data is nested

e.g. regions within countries

- Trøndelag is nested within Norway,
- Hesse is nested within Germany
- Germany and Norway are nested within Europe

Nested Effects leading to nests

Linnean classification is another example:

Animalia - Chordata - Aves - Psittaciformes - Psittacidae - Eclectus - E. roratus - E. r. solomonensis - Eric

There is a grey parrot called Eric, but he is not the same

Nestedness

Take a model with A/B

Each level of B can only be in one level of A

Fitting Nested Models

```
##                                        Estimate Std. Erro
## (Intercept)                                0.86       0.2
## OrderGaviiformes                           1.43       0.4
## OrderGruiformes                           -0.08       0.3
## OrderGaviiformes:FamilyGruidae             0.58       0.6
## OrderGalliformes:FamilyOdontophoridae     -0.71       0.5
```

No OrderGalliformes:FamilyGruidae effect!

Next Week

Model Selection!