

Multiple regressions and interactions

April 5, 2018

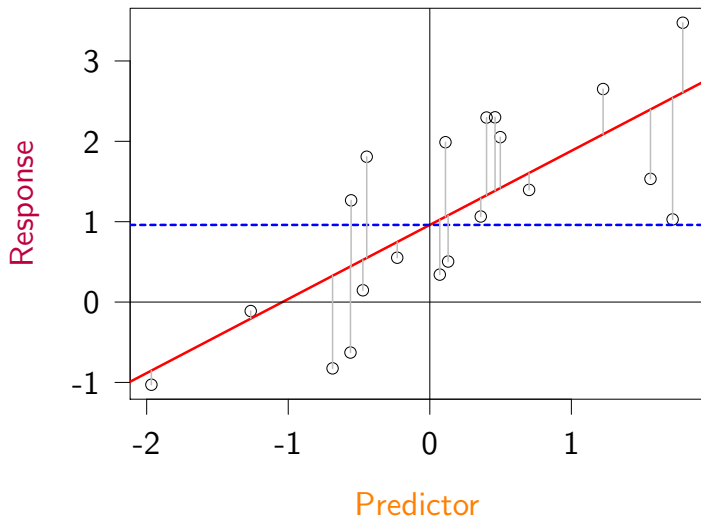
1 Linear model, reminder

2 Multiple regression

3 Interaction

A simple linear model

$$\text{Response} = \text{Intercept} + \text{Slope} \times \text{Predictor} + \text{Error}$$



A multiple linear model

$$\text{Response} = \text{Intercept} + \text{Slope1} \times \text{Predictor1} + \text{Slope2} \times \text{Predictor2} + \text{Error}$$

In R:

```
lm(response ~ 1 + predictor1 + predictor2, data=data)
```

1 Linear model, reminder

2 Multiple regression

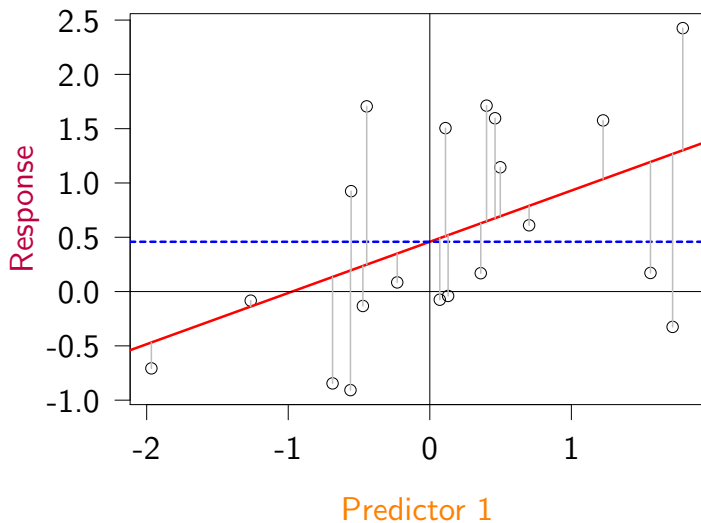
3 Interaction

Sequential regression

We want to explain a response by three predictors

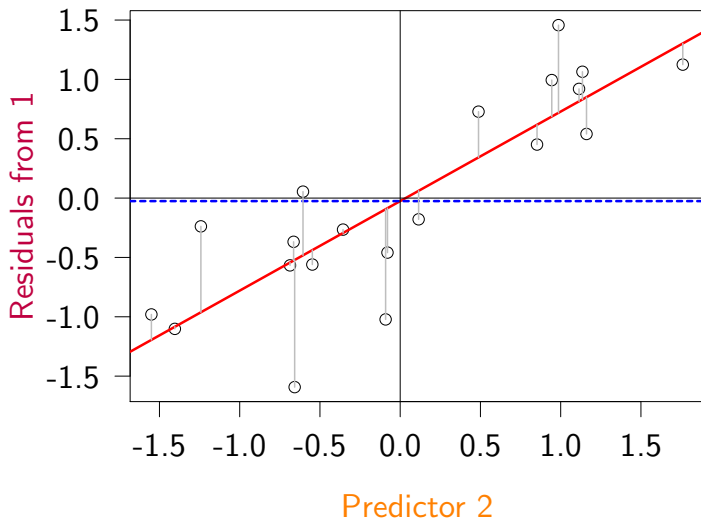
Sequential regression

We want to explain a response by three predictors



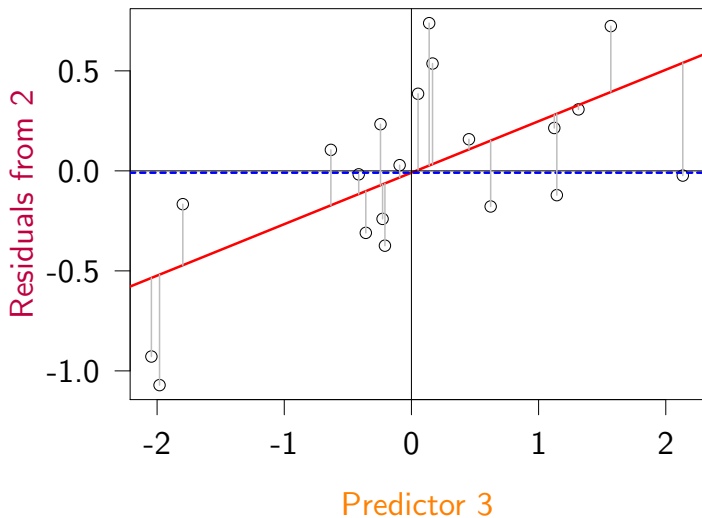
Sequential regression

We want to explain a response by three predictors



Sequential regression

We want to explain a response by three predictors



Sequential regression

```
m1 <- lm(y ~ x1)
m2 <- lm(m1$residuals ~ x2)
m3 <- lm(m2$residuals ~ x3)
```

Sequential regression

But estimates in

```
m1 <- lm(y ~ x1)
m2 <- lm(m1$residuals ~ x2)
m3 <- lm(m2$residuals ~ x3)

c(coefficients(m1)[2], coefficients(m2)[2], coefficients(m3)[2])
```

x1	x2	x3
0.4715738	0.7542078	0.2573059

are different from

```
m1 <- lm(y ~ x3)
m2 <- lm(m1$residuals ~ x2)
m3 <- lm(m2$residuals ~ x1)

c(coefficients(m1)[2], coefficients(m2)[2], coefficients(m3)[2])
```

x3	x2	x1
-0.1036939	0.9753419	-0.1019184

Sequential regression

Also what happens with classical ANOVA (aov in R)

```
summary(aov(y ~ x1 + x2 + x3))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	3.997	3.997	394.05	1.07e-12 ***
x2	1	13.998	13.998	1379.87	< 2e-16 ***
x3	1	0.120	0.120	11.82	0.00338 **
Residuals	16	0.162	0.010		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
summary(aov(y ~ x2 + x3 + x1))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x2	1	17.931	17.931	1767.562	< 2e-16 ***
x3	1	0.183	0.183	18.003	0.00062 ***
x1	1	0.002	0.002	0.176	0.68076
Residuals	16	0.162	0.010		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple regression

In contrast `lm()` optimizes relationships simultaneously
Order does **not** matter:

```
coefficients(lm(y ~ x1 + x2 + x3))
```

(Intercept)	x1	x2	x3
0.48948612	-0.01357404	1.03015700	0.08395938

```
coefficients(lm(y ~ x2 + x3 + x1))
```

(Intercept)	x2	x3	x1
0.48948612	1.03015700	0.08395938	-0.01357404

Multiple regression

BUT estimates may change with extra covariates

```
coefficients(lm(y ~ x1 + x2 ))
```

(Intercept)	x1	x2
0.50022999	-0.07029467	1.03858671

```
coefficients(lm(y ~ x1 + x2 + x3))
```

(Intercept)	x1	x2	x3
0.48948612	-0.01357404	1.03015700	0.08395938

Multiple regression

BUT estimates may change with extra covariates

```
coefficients(lm(y ~ x1 + x2 ))
```

(Intercept)	x1	x2
0.50022999	-0.07029467	1.03858671

```
coefficients(lm(y ~ x1 + x2 + x3))
```

(Intercept)	x1	x2	x3
0.48948612	-0.01357404	1.03015700	0.08395938

??

- That is a good thing
- Estimates are independent effects, conditional on the other parameters

Conditional estimation

Exercise

- 1 load jumpingdistance.csv
- 2 Use plots and `lm()` to test whether mass increases jumping distance

```
jumping <- read.csv(file = "jumpingdistance.csv")
```


Conditional estimation

Total / marginal effects

height



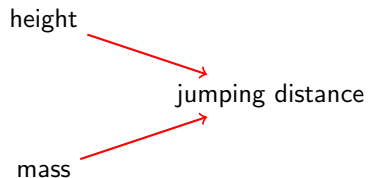
jumping distance

mass

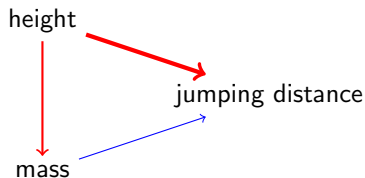


Conditional estimation

Total / marginal effects



Direct / conditional effects



- Marginal effects \approx raw correlations, sum of direct and indirect effects
- Multiple regression estimates direct effects (conditional on other predictors)
→ may reveal causal relationships

Conditional estimation

Exercise

- 1 Load babies.csv
- 2 What drives change in number of babies born?

Conditional estimation final warning: more is not always better

Are more innovative papers less rigorous?

Research question

Innovativeness

?

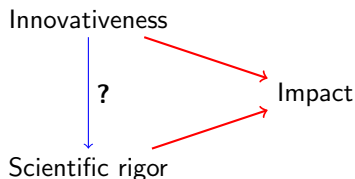
Scientific rigor

Should you correct for publication impact?

Conditional estimation final warning: more is not always better

Are more innovative papers less rigorous?

Research question



Should you correct for publication impact?

Conditional estimation final warning: more is not always better

Should you include publication impact?

```
summary(lm(rigor ~ innovativeness + impact))$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0301366	0.02188752	1.376885	1.688569e-01
innovativeness	-0.3150363	0.03051417	-10.324262	8.238502e-24
impact	0.5135830	0.01538756	33.376503	1.361378e-164

Apparent **negative** effect of innovativeness ?

```
summary(lm(rigor ~ innovativeness))$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.04104524	0.03182923	1.289545	1.975073e-01
innovativeness	0.38804729	0.03210760	12.085841	1.758144e-31

Apparent **positive** effect of innovativeness ?

Conditional estimation final warning: more is not always better

Should you include publication impact?

Conditional estimation final warning: more is not always better

Should you include publication impact?

Data simulated with positive effect of innovativeness on rigor (simulated slope 0.3)

Conditional estimation final warning: more is not always better

Should you include publication impact?

Data simulated with positive effect of innovativeness on rigor (simulated slope 0.3)

You should NOT correct for impact

Conditional estimation final warning: more is not always better

Should you include publication impact?

Data simulated with positive effect of innovativeness on rigor (simulated slope 0.3)

You should NOT correct for impact

Rule of Thumb: Do not correct for variables influenced by your predictor outside the causal path of interest

- 1 Linear model, reminder
- 2 Multiple regression
- 3 Interaction

Warnings

Vocabulary warning!

- **correlation:** linear association between two variables "*how well does x explain y ?*"

Warnings

Vocabulary warning!

- **correlation**: linear association between two variables "*how well does x explain y ?*"
- **interaction**: non-additive effect of two or more variables "*does the effect of x_1 on y change as a function of x_2 ?*". Adds a predictor (or several) to a model.

Warnings

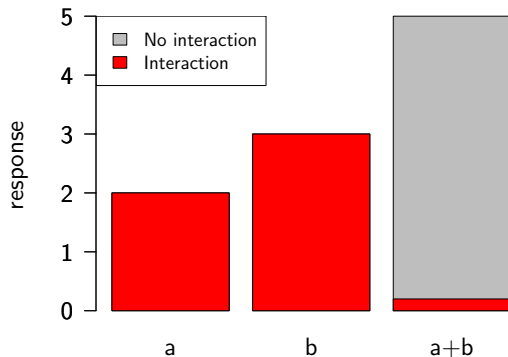
Vocabulary warning!

- **correlation**: linear association between two variables "*how well does x explain y ?*"
- **interaction**: non-additive effect of two or more variables "*does the effect of x_1 on y change as a function of x_2 ?*". Adds a predictor (or several) to a model.

Warnings

Vocabulary warning!

- **correlation**: linear association between two variables "*how well does x explain y ?*"
- **interaction**: non-additive effect of two or more variables "*does the effect of x_1 on y change as a function of x_2 ?*". Adds a predictor (or several) to a model.



Fitting an interaction

```
lm(y ~ 1 + x1 * x2)
```

```
lm(y ~ 1 + x1 + x2 + x1:x2)
```


Fitting an interaction

```
lm(y ~ 1 + x1 * x2)
lm(y ~ 1 + x1 + x2 + x1:x2)
```

```
summary(lm(y~ 1 + x1*x2))
```

Call:

```
lm(formula = y ~ 1 + x1 * x2)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.8719	-0.6777	-0.1086	0.5897	2.3166

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.14098	0.09578	11.913	< 2e-16	***
x1	-0.49281	0.10834	-4.549	1.58e-05	***
x2	0.53434	0.09881	5.408	4.67e-07	***
x1:x2	0.35911	0.11449	3.137	0.00227	**

Fitting an interaction

Why the multiplication sign?

Fitting an interaction

Why the multiplication sign?

```
x1Xx2 <- x1*x2
```

Fitting an interaction

Why the multiplication sign?

```
x1Xx2 <- x1*x2
```

```
summary(lm(y ~ 1 + x1 + x2 + x1Xx2))
```

Call:

```
lm(formula = y ~ 1 + x1 + x2 + x1Xx2)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.8719	-0.6777	-0.1086	0.5897	2.3166

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.14098	0.09578	11.913	< 2e-16	***
x1	-0.49281	0.10834	-4.549	1.58e-05	***
x2	0.53434	0.09881	5.408	4.67e-07	***
x1Xx2	0.35911	0.11449	3.137	0.00227	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Warnings

Modeling warning!

- ~~DO NOT COMPARE P-VALUES OF TWO MODELS TO TEST FOR AN INTERACTION~~

Exercise

- 1 Load the data `massex.csv`
- 2 Fit a simple regression explaining movement by mass for each sex separately. Is the relationship different between sexes?
- 3 Fit the multiple regression explaining movement by mass, sex, and `mass:sex`, using the full dataset. Is the relationship different between sexes?
- 4 Try to understand the discrepancy by plotting the data

Warnings

1.

```
masssex <- read.csv(file="masssex.csv")
```

Warnings

1.

```
massex <- read.csv(file="massex.csv")
```

2.

```
summary(lm(movement ~ mass, data=massex[massex$sex==0,]))  
summary(lm(movement ~ mass, data=massex[massex$sex==1,]))
```

Warnings

1.

```
massex <- read.csv(file="massex.csv")
```

2.

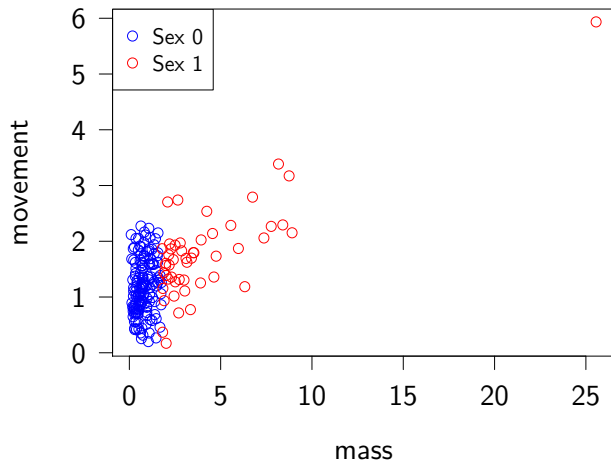
```
summary(lm(movement ~ mass, data=massex[massex$sex==0,]))  
summary(lm(movement ~ mass, data=massex[massex$sex==1,]))
```

3.

```
summary(lm(movement ~ mass*sex, data=massex))
```


Warnings

4.



Exercise

- 1 Load `plantsize.csv` and plot the data
- 2 Fit an additive model explaining plant size by `x` and `y` coordinates

```
plantsize <- read.csv("plantsize.csv")  
m0 <- lm(plantsize ~ x_location + y_location, data=plantsize)
```

Prediction

Exercise

- 1 Load `plantsize.csv` and plot the data
- 2 Fit an additive model explaining plant size by `x` and `y` coordinates
- 3 Create a prediction for plant size as a function of `x` for two values of `y`

```
plantsize <- read.csv("plantsize.csv")  
m0 <- lm(plantsize ~ x_location + y_location, data=plantsize)
```

Prediction

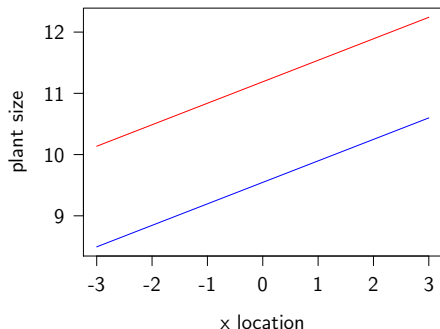
3.1. Predict

```
newdata <- data.frame(x_location = rep(seq(-3,3, length.out = 100),2),  
                      y_location = c(rep(-3, 100), rep(4,100)))  
newdata$prediction <- predict(m0, newdata = newdata)
```

Prediction

3.2 Visualize

```
setPar()  
plot(newdata$x_location[newdata$y_location==3],  
     newdata$prediction[newdata$y_location==3],  
     xlab="x location", ylab="plant size", type="l",  
     ylim = range(newdata$prediction), col="blue")  
lines(newdata$x_location[newdata$y_location==4],  
      newdata$prediction[newdata$y_location==4], col="red")
```

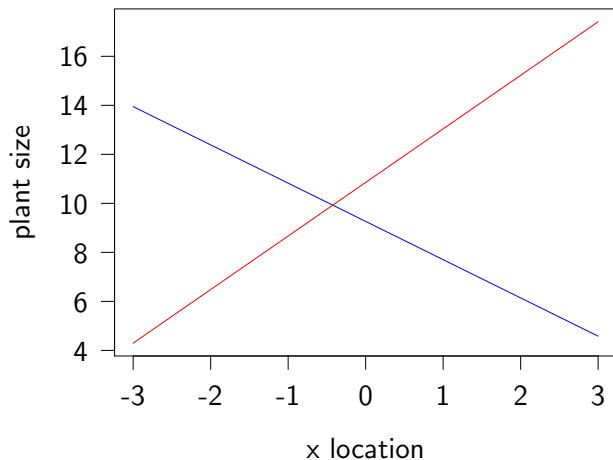


Prediction with interaction

Exercise

- 1 Load `plantsize.csv` and plot the data
- 2 Fit an additive model explaining plant size by `x` and `y` coordinates
- 3 Create a prediction for plant size as a function of `x` for two values of `y` and plot it
- 4 Fit an interaction between `x` and `y` coordinates
- 5 Create a new prediction with interaction, and plot it

Prediction with interaction

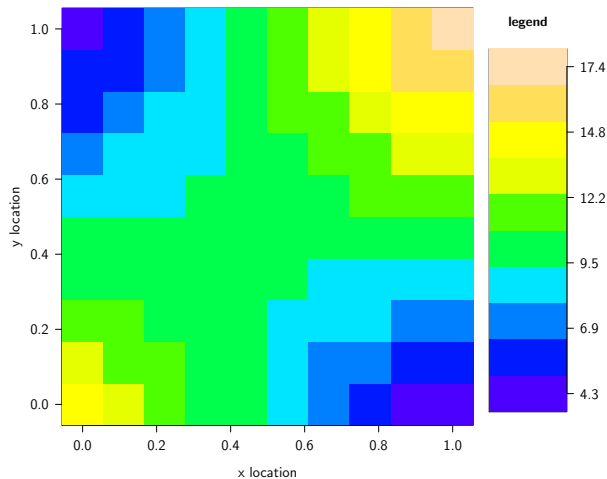


Prediction with interaction

Exercise

- 1 Load `plantsize.csv` and plot the data
- 2 Fit an additive model explaining plant size by x and y coordinates
- 3 Create a prediction for plant size as a function of x for two values of y and plot it
- 4 Fit an interaction between x and y coordinates
- 5 Create a new prediction with interaction, and plot it
- 6 Compare estimates and p -values across models. Do you think x location has an effect or not?

Prediction with interaction



Next times

- April 20th Kevin on ggplot
- May 4th Nina on Structural Equation Modeling
- then, mixed models and GLM
- **Other requests?**