Faculty of Health Sciences

Day 7: Multiple linear regression, confounding, interaction

Paul Blanche

Section of Biostatistics, University of Copenhagen

November 17, 2021

Outline

Case: vitamin D data

Data, n=412:

country	vitd	age	bmi	vitdintake
1	22.4	11.888	19.254	7.188
1	37.0	12.441	17.567	1.186
1	12.9	13.025	17.700	1.480
1	13.6	13.501	16.953	1.612
1	9.1	12.474	20.806	3.940
1	13.4	12.973	18.242	8.152

(also data on sun exposure: sunexp)

Outcome: vitamin D measured in morning blood samples, after an overnight fast (nmol/l).

Reference:
- Andersen and Skovgaard. *Regression with linear predictors*. Springer, 2010.
- Andersen et al., Eur. J. Clin. Nutr. (2005)

Note: the slides of today borrow many examples and explanations presented in more details in the above textbook reference.

Remarks on the case study and log-transformation

- It is common, and often sensible, to study the log of a concentration, instead of directly the concentration itself, when using linear regression. This is because:
  - concentration cannot be negative.
  - the variability between observations is often higher for higher concentrations.
- We will log-transform in our case study:

$$\text{outcome} = \log_{10}(\text{vitamin D concentration}) \quad .$$

Remarks on the case study and log-transformation

- It is common, and often sensible, to study the log of a concentration, instead of directly the concentration itself, when using linear regression. This is because:
  - concentration cannot be negative.
  - the variability between observations is often higher for higher concentrations.
- We will log-transform in our case study:

$$\text{outcome} = \log_{10}(\text{vitamin D concentration}) \quad .$$

- But, it is not always needed and important to log-transform! Do not systematically log-transform!
- Usually it is interesting to present and interpret the results on the original scale, using the back-transformation ($\exp$).
- We will not transform back in this lecture (although we could), just to keep everything as simple as possible and focus on different statistical considerations.

The multiple linear model

The i-th observation (e.g. from subject i) of the outcome Y is described as:

$$Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \cdots + \varepsilon_i$$

- x_i, z_i, ... are the predictor (i.e. explanatory) variables / covariates.
- the linear predictor $\alpha + \beta_1 x_i + \beta_2 z_i + \ldots$ is the mean outcome for any subject i having covariate values x_i, z_i, ...
- ε_i's are individual 'error' terms ("random/unexplained deviation from the mean") assumed normally distributed with zero mean and the same variance σ^2_ε regardless of the values x_i, z_i, ...

Model assumptions (1-2 important, 3 not always):

1. Individual observations are independent.
2. The variance of 'error' terms is the same for all groups (homogeneity).
3. 'Error' terms are normally distributed.

The multiple model generalizes simpler models

Many simple settings can be thought as a special case of the multiple linear model.

Which and why?

- t-test (Lecture 2)
  - one binary predictor variable
- univariate linear model (Lecture 3)
  - one quantitative predictor variable
- ANOVA (Lecture 4)
  - one categorical predictor variable (one-way ANOVA)
  - two categorical predictor variables (two-way ANOVA)
- ANCOVA (Today's Lecture)
  - one categorical and one quantitative predictor variable

Note: this holds when using t-test and ANOVA that assume the same standard deviation for all groups.

Case: one binary variable

- Research question: is the mean log vitamin D different between elderly women (> 69) having a "normal" weight and those being "overweight"?
- Predictor variable(s): body mass index "normal" (18.5-25) or "overweight" (>25).
- Data example: Irish women, $n = 42$ $(16 + 25)$.
- Linear model:

$$Y_i = \alpha + \beta z_i + \varepsilon_i$$

with

$$z_i = \begin{cases} 1 & \text{if } i \text{ is "overweight"} \\ 0 & \text{if } i \text{ has a "normal" weight} \end{cases}$$

  - α: mean for "normal" weight
  - $\alpha + \beta$: mean for "overweight"
  - β: difference in mean between "overweight" and "normal"

R code & default output

R code:

```
vitaminD$bmigroup <- factor(as.numeric(vitaminD$bmi > 25))
lm1 <- lm(log10(vitd)~bmigroup,data=irlwomen)
summary(lm1)
```

Output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.71987    0.04554  37.765   <2e-16 ***
bmigroup1   -0.12682    0.05832  -2.175   0.0358 *
```

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

R code & default output

R code:

```
vitaminD$bmigroup <- factor(as.numeric(vitaminD$bmi > 25))
lm1 <- lm(log10(vitd)~bmigroup,data=irlwomen)
summary(lm1)
```

Output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.71987    0.04554  37.765   <2e-16 ***
bmigroup1   -0.12682    0.05832  -2.175   0.0358 *
```

R code:

```
tapply(log10(irlwomen$vitd), irlwomen$bmigroup, mean)
diff(tapply(log10(irlwomen$vitd), irlwomen$bmigroup, mean))
```

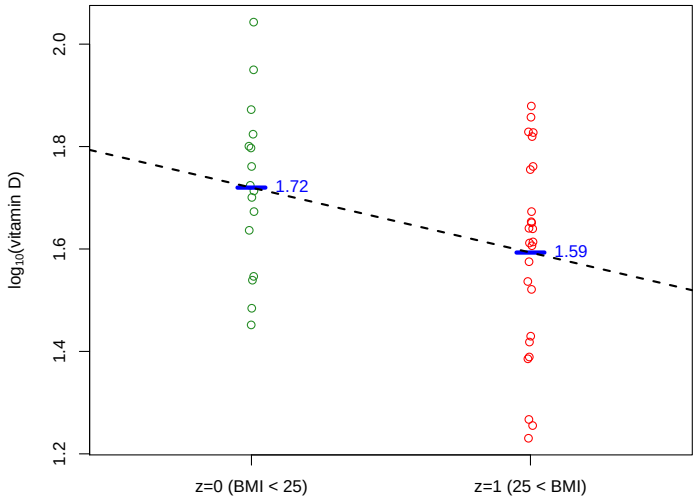
Output:

```
1.719873 1.593053
-0.1268206
```

8 / 59

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Visualizing the raw data & results



The regression line passes through the sample means, i.e. the two estimated means corresponds to the sample means in each group.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Formatted results & 95% CIs

R code:

```
publish(lm1)
```

Output:

```
    Variable Units Coefficient         CI.95  p-value
 (Intercept)              1.72   [1.63;1.81]   <1e-04
    bmigroup     0       Ref
                 1     -0.13 [-0.24;-0.01]   0.0358
```

R code:

```
t.test(log10(irlwomen$vitd) ~ irlwomen$bmigroup,var.equal=TRUE)
```

Output:

```
Two Sample t-test

data:  log10(irlwomen$vitd) by irlwomen$bmigroup
t = 2.1745, df = 39, p-value = 0.0358
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.008853898 0.244787253
sample estimates:
mean in group 0 mean in group 1
      1.719873        1.593053
```

Conclusions with only one binary variable

- Model estimates match the observed means in each group.
- The estimated regression coefficient (slope) is identical to the difference between the sample means.
- The p-values computed by the linear model and the t-test (assuming equal variances in the two groups) are identical.
- The confidence interval for the regression coefficient (slope) is identical to that computed along the t-test to complement the p-value (assuming equal variances in the two groups are identical).

Furthermore: similar remarks about identical results for the **ANOVA** case. That is why we were already using the `lm()` function of R in the ANOVA case (although R and other software have also specific function for ANOVA analyses).

Digression: median and back-transformation

R code:

```
rbind(tapply(log10(irlwomen$vitd), irlwomen$bmigroup, mean),
      tapply(log10(irlwomen$vitd), irlwomen$bmigroup, median))
```

Output:

```
           0        1
[1,] 1.719873 1.593053
[2,] 1.718883 1.613842
```

Here, because the "model" for the mean is a good model for the median (M) and because median$(\log(Y))$=$\log($median$(Y))$ we have:

$$\log_{10}(\widehat{M}_1) - \log_{10}(\widehat{M}_0) = \log_{10}\left(\frac{\widehat{M}_1}{\widehat{M}_0}\right) = -0.12682$$

and $\widehat{M}_1/\widehat{M}_0 = 10^{-0.12682} = 0.75$; hence we conclude that we estimate that overweight women have a 25% lower median vitamin D concentration compared to the normal weight women.

Outline

Why multiple regression?

- better adjusting / explaining (main focus in this course)
- better predict or gain power (more advanced topic, touch upon on day 4)

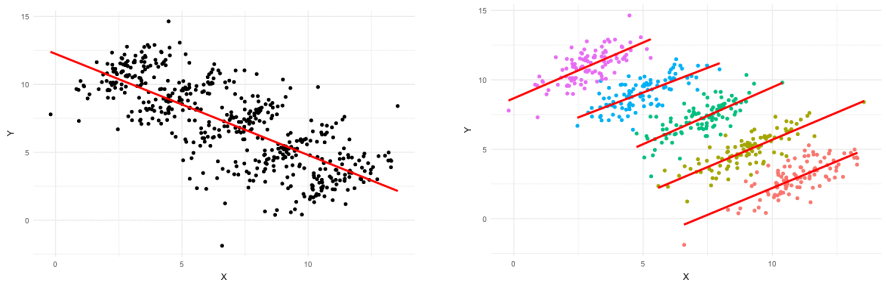
Same reasons as for why logistic regression can be more useful than simpler 2x2 tables analyses.

- Useful when we want to make comparisons with respect to one factor/variable (e.g. treatment or exposure) among individuals otherwise similar with respect to other variables that we adjust for (e.g. age, sex, comorbidity...).
- Multiple regression is a tool to deal with confounding and unbalanced designs.
- Multiple regression offers an alternative to stratification (i.e. subgroup analysis) when the data are not very large or/and we can assume that some differences are "similar" within different subgroups ($\rightarrow$).

Multiple regression to limit confounding

We often compare two groups with the aim to get a "tentative" causal interpretation of the statistical association that we can show. To do so, we adjust on some variables to make a comparison among subjects as similar as possible with respect to some relevant variables.

Extreme hypothetical example of confounding:



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Case: comparing countries

Initial research question: *"Is the average log-vitamin D different in the Irish and Polish population of elderly women?"*

Quick look at the collected data via a typical "Table I":

		Ireland (n=41)	Poland (n=65)
Age	median [iqr]	$72[70.8, 73.3]$	$71.7[70.4, 72.6]$
BMI	18.5-25	$16(39\%)$	$12(19\%)$
	> 25	$25(61\%)$	$53(81\%)$
Sun exposure	avoid	$16(39\%)$	$26(40\%)$
	sometimes	$21(51\%)$	$34(52\%)$
	prefer	$4(10\%)$	$5(8\%)$
Vitamin D intake	median [iqr]	$5.5[3.2, 12.1]$	$5.2[3.0, 11.9]$

Updated research question: *"Is there a difference in average log-vitamin D between Irish and Polish elderly women having the same BMI group?"*

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Case: comparing countries while "adjusting" for BMI group

- Research question: is there a difference in average log-vitamin D between Irish and Polish elderly women having the same BMI group?
- Predictor variable(s):
  - BMI "normal" (18.5-25) / "overweight" (>25).
  - Country Ireland / Poland
- Data example: Irish and Poland women, $n = 106$ $(41 + 65)$.
- Linear model: $Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$

$$x_i = \begin{cases} 1 & \text{if } i \text{ is Polish} \\ 0 & \text{if } i \text{ is Irish} \end{cases} \qquad z_i = \begin{cases} 1 & \text{if } i \text{ is "overweight"} \\ 0 & \text{if } i \text{ has a "normal" weight} \end{cases}$$

This is a two-way ANOVA model! (without interaction)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Parameters interpretation

According to the model, the means of log-vitamin D are:

BMI \ Country	Ireland	Poland
"Normal"	α	$\alpha + \beta_1$
"Overweight"	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2$

- α: mean outcome for Irish with "normal" BMI (reference group).
- β_1: difference in mean outcome between Irish and Polish among women of the same BMI group (whatever it is).
- β_2: difference in mean outcome between women with "overweight" and those having a "normal" BMI, among women of the same country (whatever it is).

R code & default output

R code:

```
lm2 <- lm(log10(vitd) ~  Country + bmigroup, data = irlpolwomen)
summary(lm2)
```

Output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.72854    0.04016  43.040  < 2e-16 ***
CountryPoland -0.14164    0.03947  -3.589 0.000511 ***
bmigroup1     -0.14103    0.04360  -3.235 0.001638 **
```


Formatted results & 95% CIs

R code:

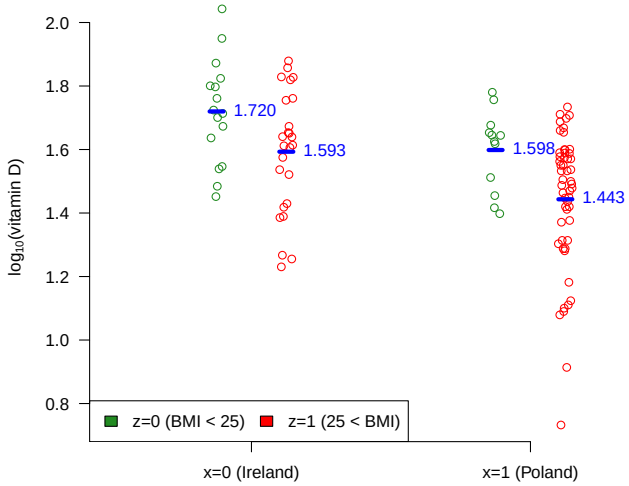
```
publish(lm2)
```

Output:

Variable	Units	Coefficient	CI.95	p-value
(Intercept)		1.73	[1.65;1.81]	< 1e-04
Country	Ireland	Ref		
	Poland	-0.14	[-0.22;-0.06]	0.0005108
bmigroup	0	Ref		
	1	-0.14	[-0.23;-0.06]	0.0016379

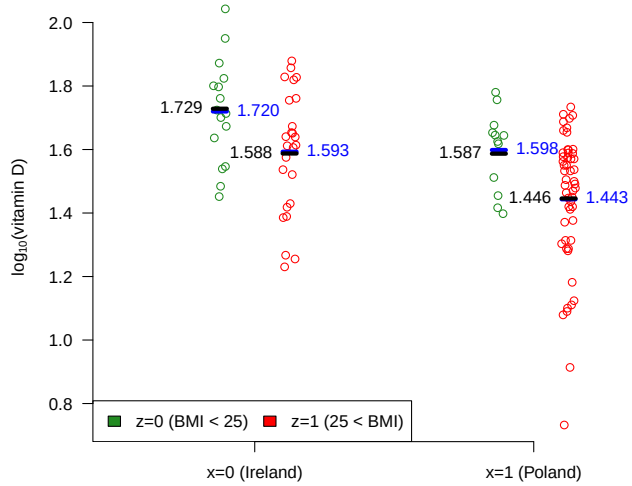
Conclusions?

Visualizing the raw data



- Observed (sample) means: blue

Visualizing the raw data & results



- Observed (sample) means: blue
- Estimated means (from the model): black

Outline

Case: comparing countries while "adjusting" for BMI

- Research question: is there a difference in mean log-vitamin D between Irish and Polish elderly women having the same BMI?
- Predictor variable(s):
  - BMI as a quantitative (continuous) variable
  - Country Ireland / Poland
- Data example: Irish and Poland women, $n = 106$ $(41 + 65)$.
- Linear model: $Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$

$$x_i = \begin{cases} 1 & \text{if } i \text{ is Polish} \\ 0 & \text{if } i \text{ is Irish} \end{cases} \qquad z_i = \text{BMI of subject } i.$$

This is a called an **ANCOVA** model (ANalysis of COVAriance).

- α: mean outcome for Irish (x=0) with BMI=0 (z=0) (meaningless!)
- β_1: difference in mean outcome between Polish and Irish among women having the same BMI (whatever it is).
- β_2: difference in mean outcome between two women, one having a BMI one unit higher than the other ($z + 1$ versus z), among women of the same country (whatever it is).
  Note: whatever the two BMI values being compared, as long as there is a one unit difference between the two.

R code:

```
irlpolwomen$bmi5 <- irlpolwomen$bmi/5
lm3 <- lm(log10(vitd) ~ bmi5 + Country, data = irlpolwomen)
summary(lm3)
```

Output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.04273    0.12291  16.620  < 2e-16 ***
bmi5         -0.07593    0.02262  -3.357  0.00110 **
CountryPoland -0.13135    0.04005  -3.280  0.00142 **
```

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

R code:

```
irlpolwomen$bmi5 <- irlpolwomen$bmi/5
lm3 <- lm(log10(vitd) ~ bmi5 + Country, data = irlpolwomen)
summary(lm3)
```

Output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.04273    0.12291  16.620  < 2e-16 ***
bmi5         -0.07593    0.02262  -3.357  0.00110 **
CountryPoland -0.13135    0.04005  -3.280  0.00142 **
```

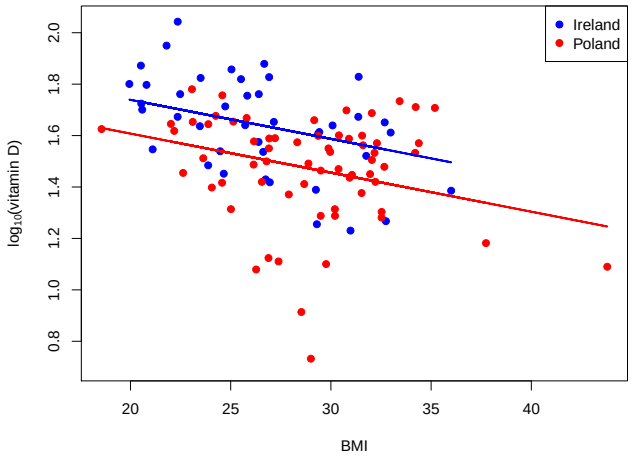
R code:

```
publish(lm3)
```

Output:

Variable	Units	Coefficient	CI.95	p-value
(Intercept)		2.04	[1.80;2.28]	< 1e-04
bmi5		-0.08	[-0.12;-0.03]	0.001103
Country	Ireland	Ref		
	Poland	-0.13	[-0.21;-0.05]	0.001418

Visualizing the raw data & results

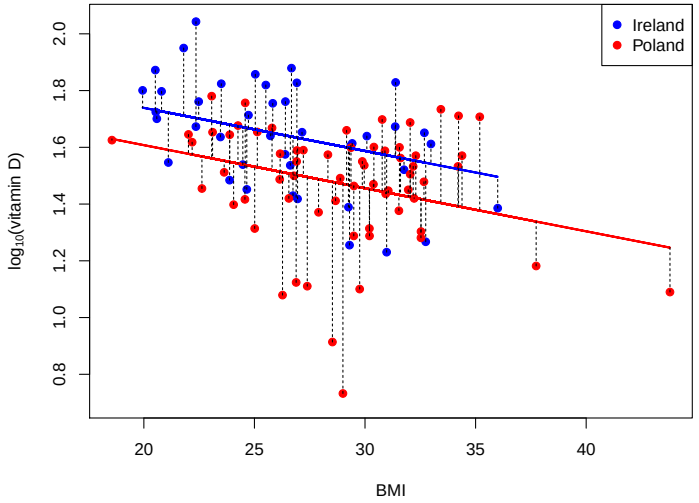


Note: an ANCOVA model is simply a regression model for parallel regression lines.

- β_2: is the common slope of the two lines.
- β_1: is the size of the vertical "shift" between the two lines.

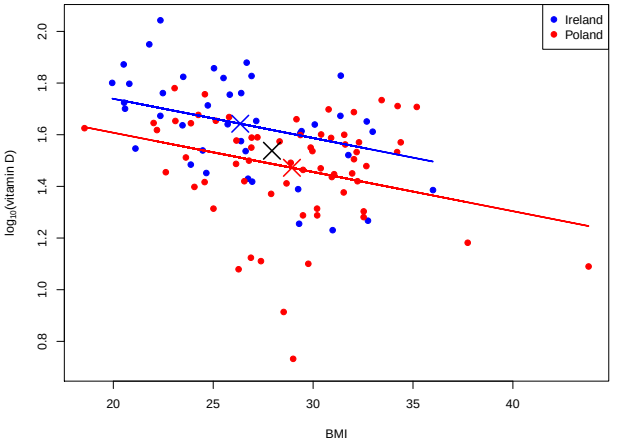
UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Estimate of σ_ε (standard deviation of the 'error' terms)



In the output of the `summary()`, "`Residual standard error: 0.1921`" is the estimate of σ_ε. It is computed "nearly" as the standard deviation of the residuals represented by the horizontal black dashed bars. It quantifies the vertical "spread" of the individual observations below/above the corresponding regression lines.

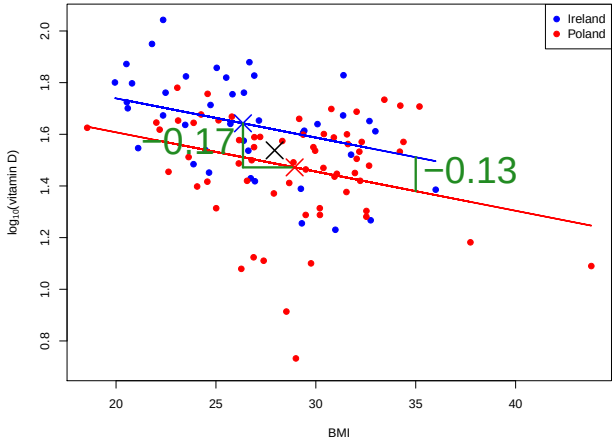
Comparing adjusted and unadjusted results



- crosses represent the mean of BMI (x-axis) and outcome (y-axis) of the entire sample (black) and for each country.

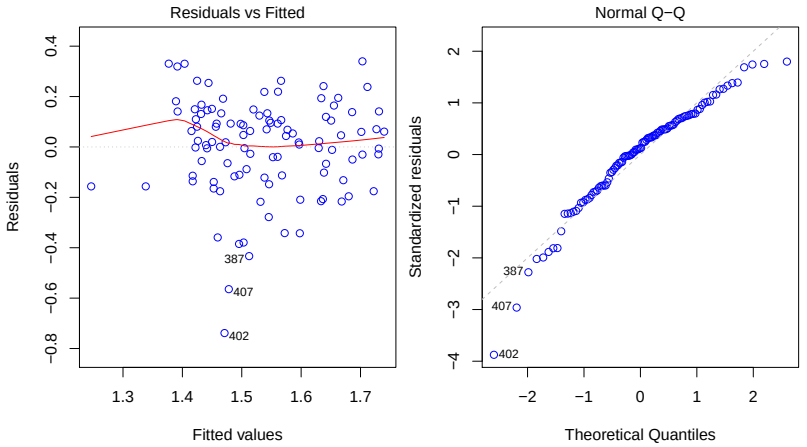
UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Comparing adjusted and unadjusted results



- crosses represent the mean of BMI (x-axis) and outcome (y-axis) of the entire sample (black) and for each country.
- Because the mean BMI is not the same in the two countries and because BMI is associated to the level of vitamin D, the adjusted and non-adjusted results are different.
- unadjusted difference between countries is -0.17, adjusted is -0.13.

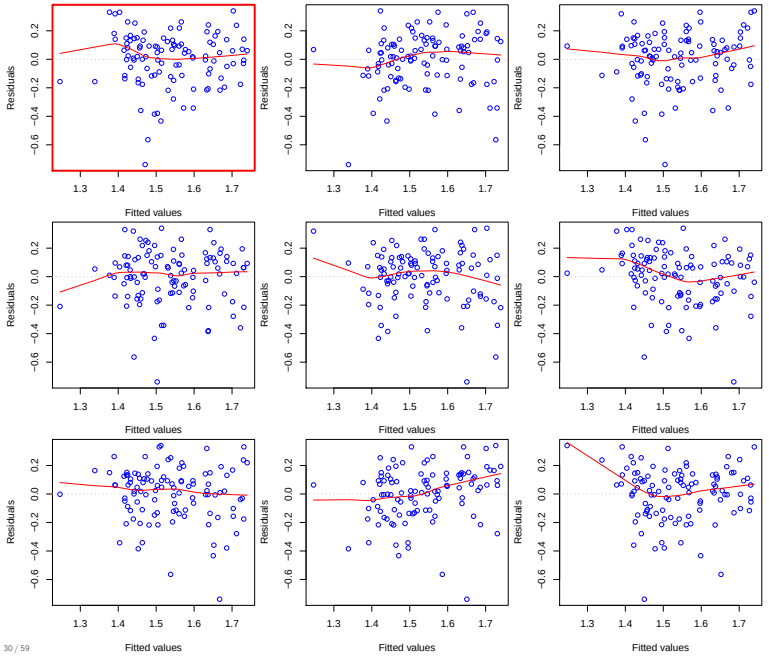
Model checking (default) plots



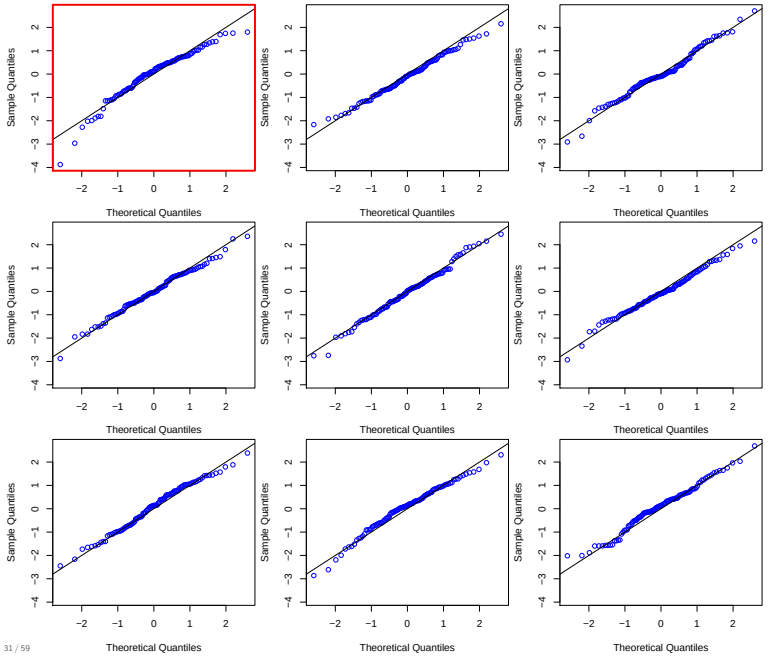
- Residual plot: always “important”.
- QQplot: mostly for small samples and when computing prediction intervals.
- Similar importance, for similar reasons, as in univariate linear regression (Lecture 3) and ANOVA model (Lecture 4).

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Wally residual plot



Wally QQplot



ANCOVA with more than two categories

- Research question: is there a difference in mean log-vitamin D between **Danish, Finnish, Irish and Polish** elderly women having the same BMI?
- Predictor variable(s):
  - BMI as a quantitative (continuous) variable
  - Country Denmark / Finland / Ireland / Poland
- Data example: all elderly women, $n = 213$ $(53 + 54 + 41 + 65)$.
- Linear model: $Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \beta_3 v_i + \beta_4 w_i + \varepsilon_i$

$x_i =$ BMI of subject i $\quad z_i = \begin{cases} 1 & \text{if } i \text{ is Finnish} \\ 0 & \text{otherwise} \end{cases}$

$v_i = \begin{cases} 1 & \text{if } i \text{ is Irish} \\ 0 & \text{otherwise} \end{cases}$ $\quad w_i = \begin{cases} 1 & \text{if } i \text{ is Polish} \\ 0 & \text{otherwise} \end{cases}$

Note: $z_i = v_i = w_i = 0$ for Danish women (reference group).

- β_2: difference in mean outcome between Finnish and Danish among women having the same BMI (whatever it is).
- β_3 & β_4: same but between Irish and Danish & Polish and Danish.

- β_2: difference in mean outcome between Finnish and Danish among women having the same BMI (whatever it is).
- β_3 & β_4: same but between Irish and Danish & Polish and Danish.

As in the simpler ANOVA context, we can test the global null hypothesis

"H_0: *there is no difference in mean log-vitamin level between women of the four countries, when comparing women of the same BMI*",

that is

$$\boxed{H_0 : \beta_2 = \beta_3 = \beta_4 = 0}$$

Similarly as in the (simpler) ANOVA context, we can use either:

- F-test
- min-P method

Pros and cons are similar to those in the ANOVA context (Lecture 4).

R code & default output

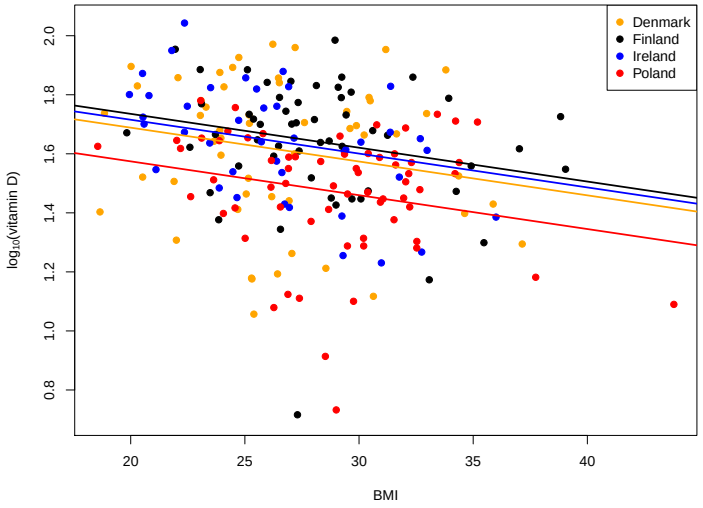
R code:

```
lm4 <- lm(log10(vitd) ~ bmi5 + Country, data = dwomen)
summary(lm4)
```

Output:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.91780    0.09710  19.750  < 2e-16 ***
bmi5          -0.05732    0.01746  -3.282  0.00121 **
CountryFinland 0.04674    0.04128   1.132  0.25891
CountryIreland 0.02683    0.04390   0.611  0.54170
CountryPoland -0.11415    0.03995  -2.857  0.00471 **
```


Visualizing the raw data & results



R code:

```
anova(lm(log10(vitd) ~ bmi5, data = dwomen),
      lm(log10(vitd) ~ bmi5 + Country, data = dwomen))
```

Output:

```
Analysis of Variance Table

Model 1: log10(vitd) ~ bmi5
Model 2: log10(vitd) ~ bmi5 + Country
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    211 10.1690
2    208  9.2635  3    0.9055 6.7773 0.0002212 ***
```

Comments:

- F-test p-value=0.0002212 is significant: there is a difference between countries, for the average outcome, when comparing women of the same BMI. But which differences?
- To avoid coding mistakes and misunderstings of R output do compare the two models: do not use "anova(lm4)".

Recommended analysis (see R-demo for code)

Statistical methods:
Comparisons between countries were made with a multiple linear model to adjust on BMI (ANCOVA). P-values and 95% confidence intervals were adjusted for multiple testing using the min-P method as implemented in the multcomp-package [ref.[1]] of the statistical software R [ref.[2]] and described in [ref.[3]].

Results (adjusted for multiple testing):

Comparison	Est. Diff	95% CI	p-value
Finland - Denmark	0.05	[-0.06; 0.15]	0.6695
Ireland - Denmark	0.03	[-0.09; 0.14]	0.9282
Poland - Denmark	-0.11	[-0.22;-0.01]	0.0239
Ireland - Finland	-0.02	[-0.13; 0.09]	0.9695
Poland - Finland	-0.16	[-0.26;-0.06]	0.0003
Poland - Ireland	-0.14	[-0.25;-0.03]	0.0069

Note:

- Significant association between countries and log vitamin D after adjusting on BMI, p-value= 0.0003 (i.e. the minimum)
- Similarly, we can use the method for the **"many-to-one"** setting (as in Lecture 4).
- This method works with any linear model, not just an ANCOVA (and does the F-test).

[1] Hothorn, Bretz & Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346–363.
[2] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[3] Bretz, Hothorn, & Westfall (2016). Multiple comparisons using R. CRC Press.

Outline

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Digression: How to adjust?

There is usually no unique "best" way to choose the variables to adjust on, but several interesting options, all with pros and cons. But, the choice should be supported by:

- Research question
  - Which groups do we want to compare? In which population? Among subjects similar with respect to what?
- Background knowledge
  - Why these groups? Why these population and comparisons among these "similar" subjects?
- Available data (variables & sample size)
  - What is the best compromise between what we ideally want to do and what we can do?

Note: several models may be needed when there are several research questions.[4]

[4] See e.g. Westreich & Greenland, Am. J. Epidemiol, 177.4 (2013): 292-298.

Digression: Table 1

- Using a simple descriptive "Table 1" to compare the distribution of all variables (except the outcome) between the groups that we want to compare often helps to choose how we should adjust.
- It is often useful to adjust on age, gender, baseline comorbidities etc or any variable which is not equally distributed between the groups[5].
- This is fine and not "cheating" (i.e. not "data snooping" or "p-hacking") as long we do not look at any association between the outcome and any variable before we make the choice on how to adjust.
- The aim of this descriptive "Table 1" is only to describe the population of each group, hence it is usually recommended that it does not include p-values.[6]

[5] and which is not a consequence of the treatments or exposures being studied

[6] See e.g. STROBE or CONSORT statements endorsed by most medical journals.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Digression: efficient "Table 1" making in R ("Publish" R package)

R code:

```
Tab1ex <- univariateTable(Country~age+Q(bmi)+Q(vitdintake)+sunexp,
                          data=IrFinPo,
                          compare.groups = FALSE,
                          show.totals = FALSE)
Tab1ex
```

Output:

Variable	Level	Denmark (n=78)	Finland (n=88)	Ireland (n=54)
age	mean (sd)	52.7 (27.7)	49.1 (29)	57.8 (25.9)
bmi	median [iqr]	24.9 [20.9, 27.5]	25.5 [20.8, 28.8]	24.9 [22.4, 28.9]
vitdintake	median [iqr]	6.1 [2.7, 11.6]	7.9 [5.0, 15.2]	5.3 [2.9, 10.5]
sunexp	avoid	14 (17.9)	15 (17.0)	18 (33.3)
	sometimes	43 (55.1)	42 (47.7)	25 (46.3)
	prefer	21 (26.9)	31 (35.2)	11 (20.4)

Digression: statistical analysis plan (SAP)

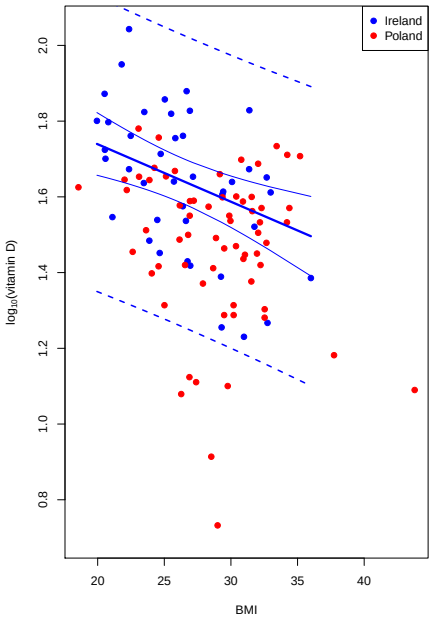
- It is **strongly** recommended to make a statistical analysis plan (SAP) before starting any analysis on the outcome data. This is a must for confirmatory research (e.g. randomized clinical trials).
- It consists of a list of research questions and corresponding analyses, ideally with a few comments to explain their rationale.

It helps to:

- better discuss with your collaborators and supervisors.
- anticipate challenges.
- rigorously prespecify your analyses and therefore increase the trust that you can have in your results.

It is completely fine to make revisions to the statistical plan and perform post hoc analyses as long as conclusions based on these additional analyses are suitably calibrated.

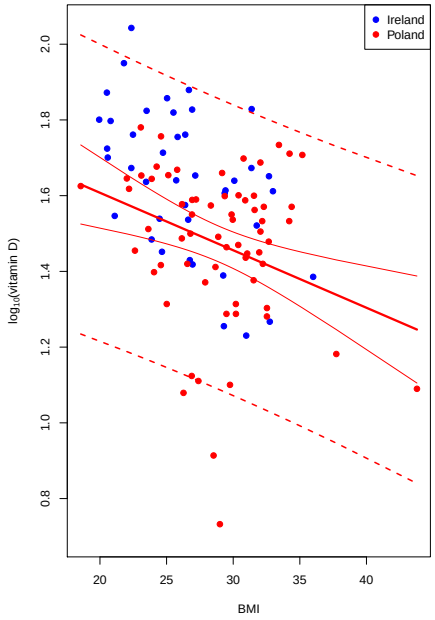
Prediction interval vs confidence intervals



- **Confidence interval** (of the estimated mean value): it quantifies the uncertainty in the estimation of the population mean. It tells us where we are "confident" that the population mean is (plain lines).
- **Prediction interval**: it tells us the range of values that include most (95%) of the observations in the entire population (dashed lines). Its width essentially depends on the estimated standard error of the "error term" σ_ε. It relies strongly on the normal distribution assumption of the "error term".

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Prediction interval vs confidence intervals



- **Confidence interval** (of the estimated mean value): it quantifies the uncertainty in the estimation of the population mean. It tells us where we are "confident" that the population mean is (plain lines).
- **Prediction interval**: it tells us the range of values that include most (95%) of the observations in the entire population (dashed lines). Its width essentially depends on the estimated standard error of the "error term" σ_ε. It relies strongly on the normal distribution assumption of the "error term".

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Outline

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Case: a (first) model with an interaction

- Research questions:
  - Is BMI associated with log-vitamin D in both Irish and Polish elderly women?
  - Is there a different association between log-vitamin D and BMI in Irish and Polish elderly women?
- Predictor variable(s):
  - BMI as a quantitative (continuous) variable
  - Country Ireland / Poland
- Data example: Irish and Poland women, $n = 106$ $(41 + 65)$.
- Linear model: $Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i \cdot z_i + \varepsilon_i$

$$x_i = \begin{cases} 1 & \text{if } i \text{ is Polish} \\ 0 & \text{if } i \text{ is Irish} \end{cases} \qquad z_i = \text{BMI of subject } i.$$

The term $\beta_3 x_i \cdot z_i$ models an interaction between x and z.

R code & default output

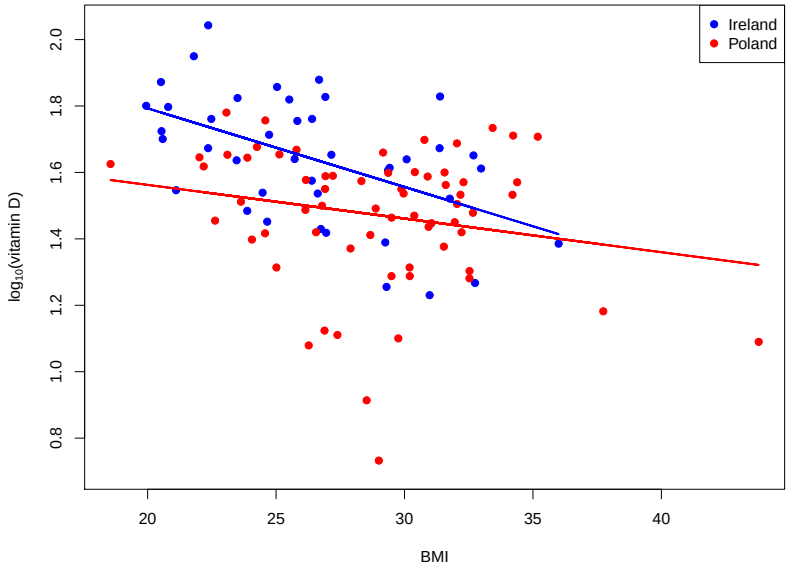
R code:

```
lm5 <- lm(log10(vitd) ~  Country * bmi5, data = irlpolwomen)
summary(irlpolwomen)
```

Output:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.26626    0.19630  11.545  < 2e-16 ***
CountryPoland      -0.50113    0.25719  -1.948  0.05410 .
bmi5               -0.11834    0.03681  -3.215  0.00175 **
CountryPoland:bmi5  0.06768    0.04650   1.455  0.14865
```

Conclusions?



- Blue slope (Ireland) $\hat{\beta}_2$= -0.11834/5 $\approx$ -0.12/5
- Red slope (Poland) $\hat{\beta}_2 + \hat{\beta}_3$= (-0.11834 + 0.06768)/5 $\approx$ -0.05/5
- Difference in slope (Poland - Ireland) $\hat{\beta}_3$= 0.06768/5 $\approx$ 0.07/5

Formatted results & 95% CIs

R code:

```
publish(lm5)
```

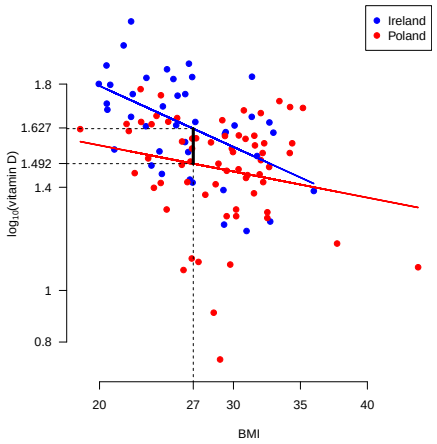
Output:

```
                Variable Units Coefficient          CI.95     p-value
             (Intercept)              2.27   [1.88;2.65]     < 1e-04
 bmi5: Country(Ireland)             -0.12 [-0.19;-0.05]    0.001749
  bmi5: Country(Poland)             -0.05  [-0.11;0.01]    0.077570
```

Note: the effect of BMI on log-vitamin D is not significant among Polish elderly women (p-value=0.078). This could not be read from the default R output, unless you refit the model after changing the reference level for the country, which is a nice trick.

Trick: re-parametrization a nicer interpretation of all estim

We refit the same model after substracting 27 to the BMI variable.



R code:

```
irlpolwomen$bmi5b <- (irlpolwomen$bmi-27)/5
lm5b <- lm(log10(vitd) ~  Country * bmi5b,
           data = irlpolwomen)
summary(lm5b)
```

Output:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.62722    0.03021  53.862  < 2e-16 ***
CountryPoland      -0.13567    0.03995  -3.396 0.000975 ***
bmi5b              -0.11834    0.03681  -3.215 0.001749 **
CountryPoland:bmi5b 0.06768    0.04650   1.455 0.148649
```

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

- `(Intercept) 1.62722`: estimated mean log vitamin D for an elderly Irish woman having a BMI of 27 (i.e. the "reference" woman here). This is an estimated mean.
- `CountryPoland -0.13567`: we estimate that, in average, the log vitamin D of a Polish elderly woman having a BMI of 27 is 0.13567 lower than that of an Irish elderly woman also having a BMI of 27. This is an estimated difference in means.
- `bmi5b -0.11834`: we estimate that, in average, the log vitamin D of any elderly Irish woman is 0.11834 lower than that of another elderly Irish woman having a 5-unit lower BMI. This is an estimated difference in means.
- `CountryPoland:bmi5b 0.06768`: we estimate that, in average, the difference in log vitamin D between two elderly Irish women, one having a 5-unit lower BMI than the other, is 0.06768 larger than the same difference among Polish elderly women. This is an estimated difference in differences in means.
  We further estimate that, in average, the log vitamin D of any elderly Polish woman is 0.11834-0.06768$\approx$ 0.05 lower than that of an elderly Polish woman having a 5-unit lower BMI.

Note: be careful when writing conclusion sentences: are you comparing "A to B" or "B to A" in the sentence? Is it the same in the output of the software?

Two-way ANOVA with interaction

- Research question: is there a difference in average log-vitamin D between Irish and Polish elderly women having the same BMI group?
- Predictor variable(s):
  - BMI "normal" (18.5-25) / "overweight" (>25).
  - Country Ireland / Poland
- Data example: Irish and Poland women, $n = 106$ $(41 + 65)$.
- Linear model: $Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i \cdot z_i + \varepsilon_i$

$$x_i = \begin{cases} 1 & \text{if } i \text{ is Polish} \\ 0 & \text{if } i \text{ is Irish} \end{cases} \qquad z_i = \begin{cases} 1 & \text{if } i \text{ is "overweight"} \\ 0 & \text{if } i \text{ has a "normal" weight} \end{cases}$$

This is a two-way ANOVA model with interaction.

Parameters interpretation

According to the model, the means of log-vitamin D are:

BMI \ Country	Ireland	Poland
"Normal"	α	$\alpha + \beta_1$
"Overweight"	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2 + \beta_3$

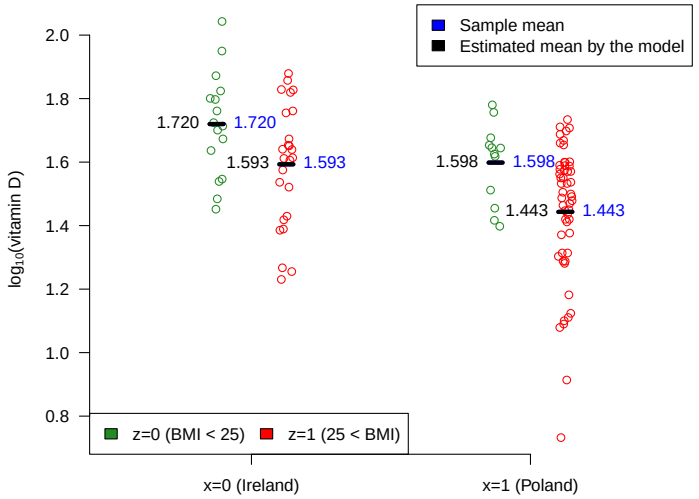
- α: mean outcome for Irish with "normal" BMI (reference group).
- β_1: difference in mean outcome between Irish and Polish among women with "normal" BMI.
- β_2: difference in mean outcome between women with "overweight" and those having a "normal" BMI, among Irish women.
- $\beta_1 + \beta_3$: difference in mean outcome between Irish and Polish among women with "overweight".
- $\beta_2 + \beta_3$: difference in mean outcome between women with "overweight" and those having a "normal" BMI, among Polish women.
- β_3: difference in differences in means....

R code:

```
lm6 <- lm(log10(vitd) ~  Country * bmigroup, data = irlpolwomen)
summary(lm6)
```

Output:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.71987    0.04840  35.538   <2e-16 ***
CountryPoland             -0.12142    0.07393  -1.643   0.1036
bmigroup1                 -0.12682    0.06198  -2.046   0.0433 *
CountryPoland:bmigroup1   -0.02838    0.08758  -0.324   0.7466
```

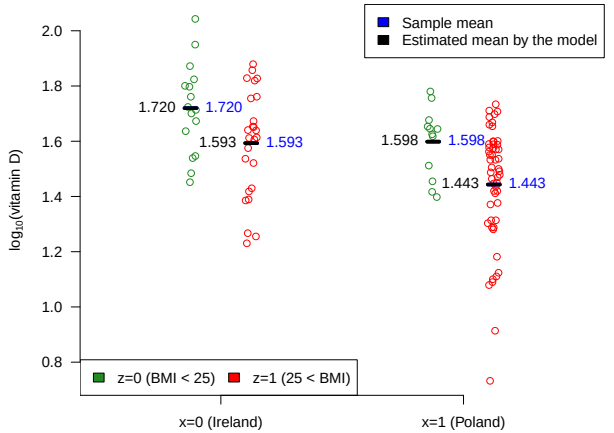


- Here the **estimated means** are equal to the sample means. We say that the model for the mean is saturated (because we have 4 parameters to estimate 4 means).

- 1.593 - 1.720 = -0.127 ≈ -0.13
- 1.598 - 1.443 = 0.155 ≈ 0.16
- 1.598 - 1.720 = -0.122 ≈ -0.12
- 1.443 - 1.593 = -0.15

R code:

```
publish(lm6)
```



Output:

Variable	Units	Coefficient	CI.95	p-value
(Intercept)		1.72	[1.63;1.81]	< 1e-04
Country(Ireland): bmigroup(1 vs 0)		-0.13	[-0.25;-0.00]	0.043303
Country(Poland): bmigroup(1 vs 0)		-0.16	[-0.28;-0.03]	0.013723
bmigroup(0): Country(Poland vs Ireland)		-0.12	[-0.27;0.03]	0.103562
bmigroup(1): Country(Poland vs Ireland)		-0.15	[-0.24;-0.06]	0.00188

Interaction versus subgroup analysis

- In the two previous examples, the only difference in the model assumptions between using a model with an interaction and performing a subgroup analysis (one per country) is the way we model the standard deviation of the error term σ_ε: we would model two different values with the subgroup analysis, whereas only one with the interaction model.

- If we had adjusted on more variables, then the difference would be more important, because the subgroup analysis would implicitly also model interactions with all these other variables.

Case: stratifying vs adjusting with interaction

Comparing estimated parameters:

	Adjust + inter		Subgroup	
	Poland	Ireland	Poland	Ireland
BMI (by 5)	-0.126	-0.050	-0.103	-0.047
Sun: sometimes vs avoid	0.020	0.020	-0.068	0.073
Sun: prefer vs avoid	0.054	0.054	-0.117	0.159

From the three models:

1. `lm(log10(vitd) ~ Country * bmi5b + sunexp, data = irlpolwomen)`
2. `lm(log10(vitd) ~ bmi5b + sunexp, data = poland)`
3. `lm(log10(vitd) ~ bmi5b + sunexp, data = ireland)`

Note: in model 1 (adjust + interaction), we assume that the "effect" of sun exposure is similar in Poland and Ireland, which is not the case with the subgroup analysis.

Final words on modeling

Many topics discussed today and on day 6 are important beyond the linear and logistic model.

Most of the reasoning about modeling choices, including:

- which variables to include?
- how? (with or without interaction, categorized version or not...)
- why does it matter?

applies for more complicated model that you may encounter/need during your research career.

When, why, where to seek statistical help?

When? if you are not sure about how to...

- plan your experiment or clinical trial
- analyze your data
- answer reviewers or collaborators concerns

It is often more complicated than initially thought...

Why?

- you might get quick help and simple advice that make a big difference.
- to minimize the risk of "wasting" your precious research time and work by inappropriately analysis your data.
- why not? it can sometimes be free of charge :-)

Where?

- Section of Biostatistics (`https://publichealth.ku.dk/about-the-department/biostat/`)
  - by phone (quick & simple questions): free
  - short meeting (20 mins): free
  - new collaborations: sometimes free, but usually not.
- private companies also exist.