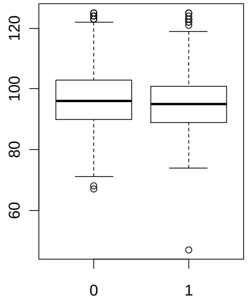
STK4900/9900 - Lecture 4

Program

1. Counterfactuals and causal effects
2. Confounding
3. Interaction
4. More on ANOVA

- Sections 4.1, 4.4, 4.6
- Supplementary material on ANOVA

Example (cf. practical exercise 10)



How does exercise affect blood glucose level?

Use the HERS data, disregarding women with diabetes

Simple linear regression:

```
              Estimate     Std. Error     t value     Pr(>|t|)
(Intercept)    97.36       0.282          345.8      < 2e-16
exercise       -1.693      0.438          -3.87       0.00011

Residual standard error: 9.715 on 2030 degrees of freedom
Multiple R-squared: 0.0073,   Adjusted R-squared: 0.0068
F-statistic: 14.97 on 1 and 2030 DF,  p-value: 0.00011
```

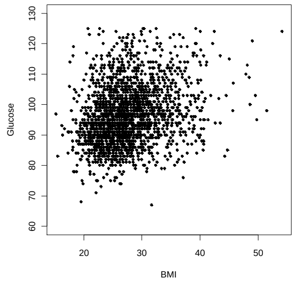
Can we conclude that exercise on average decreases the blood glucose level with 1.7 mg/dL ?

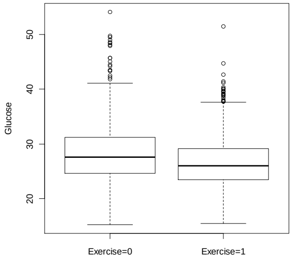
Problem:

The women who exercise are not a random sample of all women in the cohort (as they would have been in a clinical trial), but differ from the women who don't exercise, e.g. with respect to age, alcohol use, and body mass index (BMI)

Further age, alcohol use, and BMI may influence the glucose level

Illustration for BMI:





Considering this problem, can anything be said about the "causal effect" of exercise on blood glucose level?

Counterfactuals and causal effects

For the general discussion we consider some outcome (e.g. glucose level) and we want to see how this is affected by a binary predictor, or "exposure", X_1 (e.g. exercise) with X_1 =1 corresponding to "exposed" and X_1 =0 corresponding to "unexposed"

Suppose (counter to the fact) that we could run an experiment in which

- first *every* individual is exposed (i.e. X_1 =1) and the outcome Y_1 is observed
- then, turning back the clock, *every* individual is unexposed (i.e. X_1 =0) and the outcome Y_0 is observed

All other characteristics of the individuals are assumed to be the same in the two parts of the hypothetical experiment

In real life, we can not turn back the clock, so one of the two experimental outcomes for every individual is an *unobserved counterfactual*

The *causal effect* (in a statistical sense) of the exposure is defined as the difference in population means under the two parts of the counterfactual experiment:

$$\text{Causal effect} = E(Y_1) - E(Y_0)$$

If the means differ, we say that the exposure is a *causal determinant* of the outcome

A simple model for the counterfactual experiment

To make the argument simple, we assume that all other characteristics of the individuals are captured by a binary covariate X_2 which also has a causal effect on the outcome

Further we assume that the (counterfactual) outcome for individual i when exposed take the form

$$y_{1i} = \beta_0 + \beta_1^c + \beta_2^c x_{2i} + \varepsilon_{1i}$$

while when unexposed it becomes

$$y_{0i} = \beta_0 + \beta_2^c x_{2i} + \varepsilon_{0i}$$

Then the population means for the two parts of the counterfactual experiment become

$$\text{exposed:} \quad E(Y_1) = E(\beta_0 + \beta_1^c + \beta_2^c X_2 + \varepsilon_1)$$
$$= \beta_0 + \beta_1^c + \beta_2^c E(X_2)$$

$$\text{unexposed:} \quad E(Y_0) = E(\beta_0 + \beta_2^c X_2 + \varepsilon_0)$$
$$= \beta_0 + \beta_2^c E(X_2)$$

In the counterfactual experiment the distribution of X_2 is the same in both parts of the experiment, and hence its mean is the same

Hence the causal effect of the exposure becomes

$$\text{Causal effect} = E(Y_1) - E(Y_0)$$
$$= \beta_0 + \beta_1^c + \beta_2^c E(X_2) - \left\{\beta_0 + \beta_2^c E(X_2)\right\}$$
$$= \beta_1^c$$

Confounding

In reality we cannot observe the counterfactuals

We can only observe the outcome for an individual under one of the two conditions (exposed/unexposed)

In practice we therefore have to compare the mean values of the outcome in two distinct populations, one exposed and one unexposed

But then there is no guarantee that the mean value of X_2 will be the same in the exposed and unexposed populations

Let $E_1(X_2)$ denote the mean of X_2 among the exposed,
and let $E_0(X_2)$ denote the mean of X_2 among the unexposed

For the exposed population:

$$E(Y_1) = \beta_0 + \beta_1^c + \beta_2^c E_1(X_2)$$

For the unexposed population:

$$E(Y_0) = \beta_0 + \beta_2^c E_0(X_2)$$

Thus

$$\begin{aligned} E(Y_1) - E(Y_0) &= \beta_0 + \beta_1^c + \beta_2^c E_1(X_2) - \left\{\beta_0 + \beta_2^c E_0(X_2)\right\} \\ &= \beta_1^c + \beta_2^c \left\{E_1(X_2) - E_0(X_2)\right\} \end{aligned}$$

If we perform a study where we sample from the exposed and unexposed populations, and estimate the difference based on the exposed and unexposed samples, we will estimate

$$\beta_1^c + \beta_2^c \left\{E_1(X_2) - E_0(X_2)\right\}$$

If the mean value of X_2 differs between the exposed and unexposed populations, we will get a biased estimate of the causal effect β_1^c

We say that the (causal) effect of X_1 is confounded by X_2

No confounding

If the distribution of X_2 is independent of the level of exposure (i.e. X_1 = 0,1), then $E_1(X_2) = E_0(X_2)$ and there will be no confounding

In particular this will be the case in an experiment where individuals are randomly allocated to exposure/no exposure

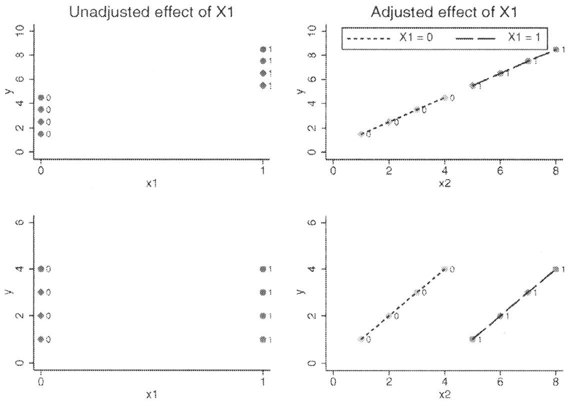
Conditions for confounding

A covariate X_2 is a confounder for the causal effect of X_1 provided that

- X_2 is a causal determinant of the outcome Y (or a proxy for such determinants)
- X_2 is a causal determinant of X_1 (or they share a common causal determinant)

Confounding patterns

Examples of confounding patterns when X_2 is a numerical covariate



Complete confounding

Negative confounding

Fig. 4.1 in the book

Control of confounding

Consider the situation where all causal determinants other than X_1 are captured by the binary covariate X_2

Then, given the level of X_2 (= 0,1), there is no more confounding and the causal effect of X_1 may estimated by comparing the means of exposed and unexposed within levels of X_2

In practice this is obtained by fitting the linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

since here β_1 is the effect of one unit's increase in X_1 keeping the value of X_2 constant

In general we may use multiple linear regression to correct for a number of confounders by including them as covariates in the model (assuming that all relevant confounders are recorded in the data)

Example (contd)

We fit a multiple regression model with blood glucose level as response and exercise, age, alcohol use, and body mass index (BMI) as covariates

Multiple linear regression:

```
              Estimate     Std. Error    t value    Pr(>|t|)
(Intercept)   78.96        2.592         30.45      <2e-16
exercise      -0.950       0.429         -2.22      0.0267
age           0.064        0.03           2.02      0.0431
drinkany      0.680        0.422          1.61      0.1071
BMI           0.489        0.042         11.77      <2e-16

Residual standard error: 9.389 on 2023 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.072,    Adjusted R-squared: 0.070
F-statistic: 39.22 on 4 and 2023 DF,  p-value: < 2.2e-16
```

We now find that exercise on average decreases the blood glucose level with 1.0 mg/dL

This should be closer to the causal effect of exercise

Interaction for binary covariates

We have considered the situation where two binary predictors X_1 and X_2 have a causal effect on the outcome

We could then estimate the (causal) effects by fitting the linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Note that we assume that the effect of X_1 is the same for both levels of X_2 (and vice versa):

X_1	X_2	$E(y \mid \mathbf{x})$
0	0	β_0
1	0	$\beta_0 + \beta_1$
0	1	$\beta_0 + \beta_2$
1	1	$\beta_0 + \beta_1 + \beta_2$

If the effect of X_1 depends on the level of X_2 we have an *interaction*

We may then fit a model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

The effect for different values of the covariates are then given by:

X_1	X_2	$X_1 X_2$	$E(y \mid \mathbf{x})$
0	0	0	β_0
1	0	0	$\beta_0 + \beta_1$
0	1	0	$\beta_0 + \beta_2$
1	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Example

Use the HERS data to study how low-density lipoprotein cholesterol after one year (LDL1) depends on hormone therapy (HT) and statin use (both binary)

R commands:

```
ht.fit=lm(LDL1~HT+statins+HT:statins, data=hers)
summary(ht.fit)
```

R output (edited):

```
              Estimate     Std. Error    t value    Pr(>|t|)
(Intercept)   145.157      1.326         109.507    < 2e-16
HT            -17.73       1.87          -9.477     < 2e-16
statins       -13.81       2.15          -6.416      1.65e-10
HT:statins      6.24       3.08           2.030      0.0425
```

(In the model formula HT:statin specifies the interaction term "HT*statin")

The effect of HT seems to be lower among statin users

```
              Estimate    Std. Error    t value     Pr(>|t|)
(Intercept)   145.157     1.326         109.507    < 2e-16
HT            -17.73      1.87          -9.477     < 2e-16
statins       -13.81      2.15          -6.416      1.65e-10
HT:statins      6.24      3.08           2.030       0.0425
```

HT reduces LDL cholesterol for non-users of statins by 17.7 mg/dl

For users of statins the estimated reduction is 17.7 - 6.2 = 11.5 mg/dl

To obtain the uncertainty, we use the "contrast" library

R commands:

```
library(contrast)
par1= list(HT=1,statins=1)    # specify one set of values of the covariates
par2= list(HT=0,statins=1)    # specify another set of values of the covariates
contrast(ht.fit, par1,par2)   # compute the difference between the two sets
```

R output (edited):

```
 Contrast      S.E.     Lower    Upper     t      df      Pr(>|t|)
-11.48         2.44     -16.27   -6.69    -4.7    2604     0
```

Interaction for one binary and one numerical covariate

We now consider the situation where X_1 is a binary predictor and X_2 is numerical

As an illustration we consider the HERS data, and we will see how baseline LDL cholesterol depends on statin use (X_1) and BMI (X_2)

The model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

assumes that the effect of BMI is the same for statin users and those who don't use statins

It may be of interest to consider a model where the effect of BMI may differ between statin users and those who don't use statins, i.e. where there is an *interaction*

We then consider the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

Note that the model may be written

$$y_i = \begin{cases} \beta_0 + \beta_2 x_{2i} + \varepsilon_i & \text{when} \quad x_{1i} = 0 \\ \beta_0 + \beta_1 + (\beta_2 + \beta_3) x_{2i} + \varepsilon_i & \text{when} \quad x_{1i} = 1 \end{cases}$$

This is a model with different intercepts and different slopes for the numerical covariate depending on the value of the binary covariate

When considering such a model, it is useful to center the numeric covariate (by subtracting its mean) to ease interpretation

In the example, we let X_2 correspond to the centered BMI-values, denoted cBMI

R commands:

```
hers$cBMI=hers$BMI - mean(hers$BMI[!is.na(hers$BMI)])
stat.fit=lm(LDL~statins+cBMI+statins:cBMI,data=hers)
summary(stat.fit)
par1=list(statins=1,cBMI=1)
par2=list(statins=1,cBMI=0)
contrast(stat.fit,par1,par2)
```

R output (edited):

```
               Estimate     Std. Error    t value     Pr(>|t|)
(Intercept)    151.09       0.881         171.58     < 2e-16
statins        -16.72       1.463         -11.43     < 2e-16
cBMI             0.640      0.156           4.09     4.41e-05
statins:cBMI    -0.721      0.269          -2.68     0.0075

 Contrast    S.E.    Lower     Upper     t      df     Pr(>|t|)
 -0.081      0.219   -0.511     0.349   -0.37   2743   0.712
```

Interaction for two numerical covariates

We finally consider the situation where X_1 and X_2 are both numerical

A model with interaction is then given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

For such a model, it is useful to center the covariates

But even then the interpretation of the estimates is a bit complicated

Two-way ANOVA

Consider the situation where the outcome y_i for an individual depends on two factors, A and B, each with two levels, denoted a_1, a_2 and b_1, b_2

One such example is how LDL cholesterol depends on HT (with levels "placebo" and "hormone therapy") and statin use (with levels "no" and "yes"); cf. slide 16

We may here introduce the covariates:

$$x_{1i} = \begin{cases} 0 & \text{if individ } i \text{ has level } \mathrm{a}_1 \text{ for factor A (reference)} \\ 1 & \text{if individ } i \text{ has level } \mathrm{a}_2 \text{ for factor A} \end{cases}$$

$$x_{2i} = \begin{cases} 0 & \text{if individ } i \text{ has level } \mathrm{b}_1 \text{ for factor B (reference)} \\ 1 & \text{if individ } i \text{ has level } \mathrm{b}_2 \text{ for factor B} \end{cases}$$

Then a regression model with interaction takes the form (cf slide 15)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

If (e.g.) factor B has three levels b_1, b_2 , b_3, we need to introduce two *x*'s for this factor (cf slide 26 of Lecture 3):

$$x_{2i} = \begin{cases} 1 & \text{if individ } i \text{ has level } \mathrm{b}_2 \text{ for factor B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{if individ } i \text{ has level } \mathrm{b}_3 \text{ for factor B} \\ 0 & \text{otherwise} \end{cases}$$

A model with interaction then takes the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \varepsilon_i \qquad (*)$$

It becomes quite complicated to write the model like this, so it is common to use an alternative formulation

We recapitulate:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \varepsilon_i \qquad (*)$$

In order to rewrite model (*), we denote the outcomes for level a_j of factor A and level b_k of factor B by

$$y_{ijk} \quad \text{for} \quad i = 1, \ldots, n_{jk}$$

We may then rewrite model (*) as

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \qquad (**)$$

We have the following relations between the parameters in model (*) and model (**)

(*)	β_0	β_1	β_2	β_3	β_4	β_5
(**)	μ	α_2	β_2	β_3	$(\alpha\beta)_{22}$	$(\alpha\beta)_{23}$

In model (**) the parameters for the reference levels are 0 :

$$\alpha_1 = \beta_1 = (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{13} = (\alpha\beta)_{21} = 0$$

Note that the model formulation

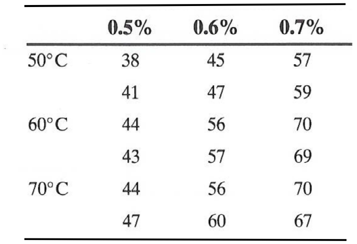
$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \qquad (**)$$

works equally well when factor A has J levels and factor B has K levels, while the formulation (*) would become much more complicated

In Lecture 3 (cf. slide 30), we considered a study of how the extraction rate of a certain polymer depends on temperature and the amount of catalyst used.

We there assumed a linear effect of temperature and the amount of catalyst

We will here consider temperature and catalyst as factors, each with three levels



	0.5%	0.6%	0.7%
50°C	38	45	57
	41	47	59
60°C	44	56	70
	43	57	69
70°C	44	56	70
	47	60	67

R commands:

```
polymer=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/polymer.txt",header=T)
polymer$ftemp=factor(polymer$temp)
polymer$fcat=factor(polymer$cat)
fit=lm(rate~ftemp+fcat+ftemp:fcat,data=polymer)
summary(fit)
```

R output:

```
                  Estimate        Std. Error      t value     Pr(>|t|)
(Intercept)        39.5           1.23            32.25       1.30e-10
ftemp60             4.0           1.73             2.31       0.046
ftemp70             6.0           1.73             3.46       0.007
fcat0.6             6.5           1.73             3.75       0.005
fcat0.7            18.5           1.73            10.68       2.06e-06
ftemp60:fcat0.6     6.5           2.45             2.65       0.026
ftemp70:fcat0.6     6.0           2.45             2.45       0.037
ftemp60:fcat0.7     7.5           2.45             3.06       0.014
ftemp70:fcat0.7     4.5           2.45             1.84       0.099

Residual standard error: 1.73 on 9 degrees of freedom
Multiple R-squared: 0.986,     Adjusted R-squared: 0.973
F-statistic: 78.78 on 8 and 9 DF,  p-value: 2.012e-07
```

In a planned experiment we can make sure that we have the same number of observations for all the *J* x *K* combinations of levels of factor A and factor B

We then have a *balanced* design, and the total sum of squares (TSS) may be uniquely decomposed as a sum of squares for each of the two factors (SSA, SSB), a sum of squares for interaction (SSAB), and a residual sum of squares (RSS):

$$TSS = SSA + SSB + SSAB + RSS$$

To each of these sum of squares there correspond a degree of freedom as given in the ANOVA table on the next slide

NB! If the design is <u>not</u> balanced, the decomposition of the total sum of squares is <u>not</u> unique

The result of a two-way ANOVA may be summarized in the table

Source	df	Sum of squares	Mean sum of squares	F statistics
Factor A	$J-1$	SSA	$SSA/(J-1)$	$F = \dfrac{SSA/(J-1)}{RSS/(n-JK)}$
Factor B	$K-1$	SSB	$SSB/(K-1)$	$F = \dfrac{SSB/(K-1)}{RSS/(n-JK)}$
Interaction	$(J-1)(K-1)$	$SSAB$	$SSAB/(J-1)(K-1)$	$F = \dfrac{SSAB/(J-1)(K-1)}{RSS/(n-JK)}$
Residual	$n-JK$	RSS	$RSS/(n-JK)$	
Total	$n-1$	TSS		

The F-statistics (with their appropriate degrees of freedom) may be used to test the following null hypotheses:

$$H_0:\ \text{all}\ \ (\alpha\beta)_{jk} = 0 \quad \text{(no interaction)}$$

$$H_0:\ \text{all}\ \ \alpha_j = 0 \quad \text{(no main effect of A)}$$

$$H_0:\ \text{all}\ \ \beta_k = 0 \quad \text{(no main effect of B)}$$

For the example:

R commands:

```
anova(fit)
```

R output:

```
Analysis of Variance Table

              Df     Sum Sq      Mean Sq      F value      Pr(>F)
ftemp          2     332.11      166.06       55.35        8.76e-06
fcat           2    1520.11      760.06       253.35       1.23e-08
ftemp:fcat     4      38.56        9.64       3.213        0.067
Residuals      9      27.00        3.00
```

Higher level ANOVA

Consider for illustration the situation with three factors, A, B, and C.

Data:

y_{ijkl} = observation number i for level a_j of factor A,
level b_k of factor B, and level c_l of factor C

Model with interaction:

$$y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + \varepsilon_{ijkl}$$

The result of a three-way ANOVA may be summarized in the table

Source	df *	Sum of squares	Mean sum of squares	F statistics
Factor A		SSA	SSA/df	F_A
Factor B		SSB	SSB/df	F_B
Factor C		SSC	SSC/df	F_C
Interaction AB		$SSAB$	$SSAB/df$	F_{AB}
Interaction AC		$SSAC$	$SSAC/df$	F_{AC}
Interaction BC		$SSBC$	$SSBC/df$	F_{BC}
Interaction ABC		$SSABC$	$SSABC/df$	F_{ABC}
Residual		RSS	RSS/df	
Total	$n-1$	TSS		

*) can be found on computer output

The decomposition of the total sum of squares is unique if the design is balanced

Hypothesis testing is similar to two-way ANOVA