

STK4900/9900 - Lecture 4

Program

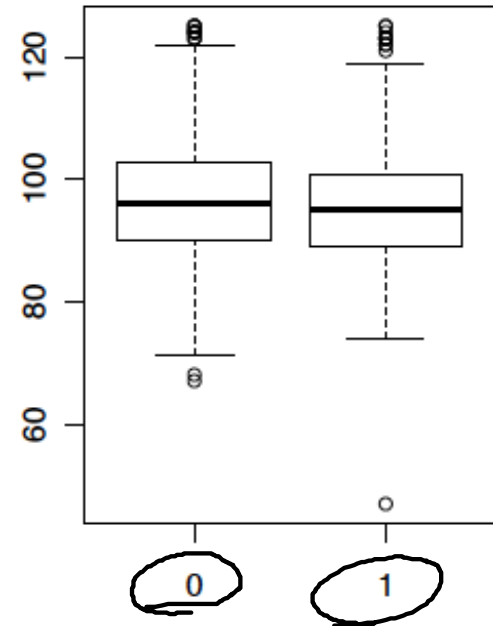
1. Causal effects
2. Confounding
3. Interaction
4. More on ANOVA
5. Prediction

- Sections 4.1, 4.4, (4.5), 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

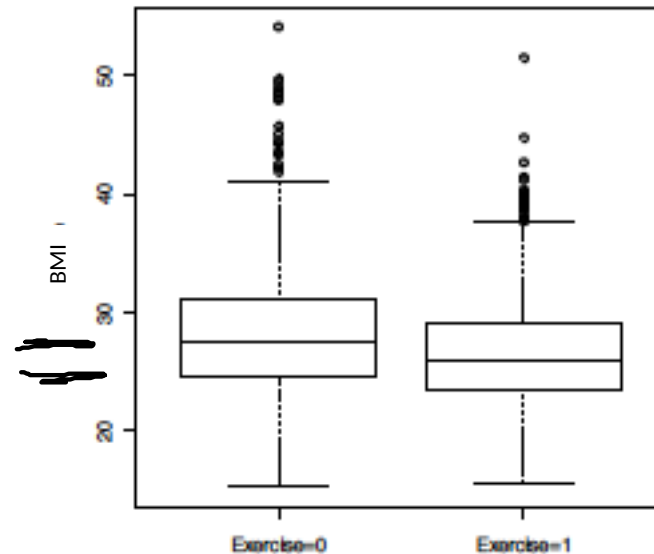
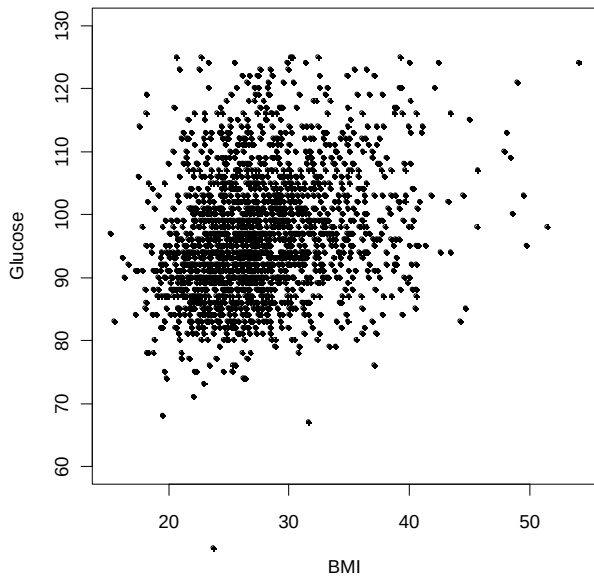
Can we conclude that exercise on average decreases the blood glucose level with 1.7 mg/dL?

Problem:

The women who exercise are not a random sample of all women in the cohort (as they would have been in a randomized 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

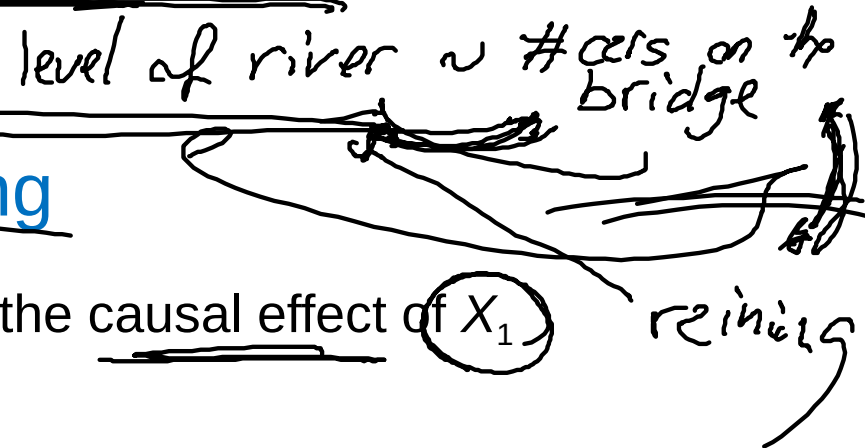
Illustration for BMI:



Confounding

It is possible that the observed significant association between exercise and glucose levels is due to the dependency between exercise and BMI and other covariates, i.e. not causal.

In such case we say that the association is spurious and that BMI and the other covariates are confounding variables, more precisely we have



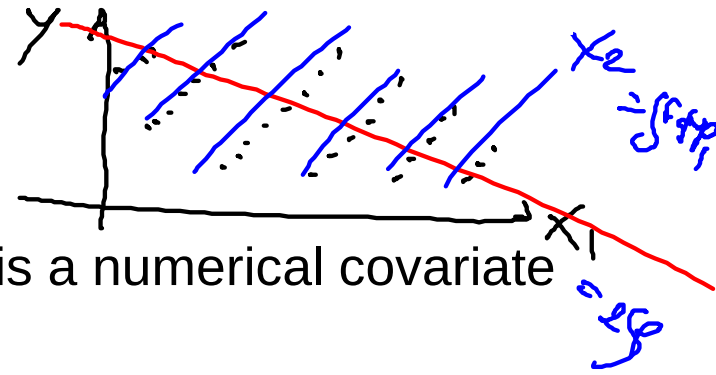
Conditions for confounding

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

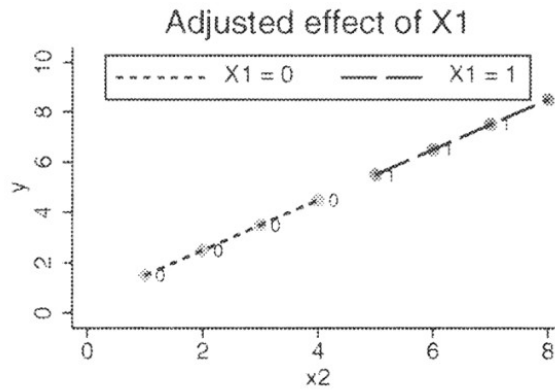
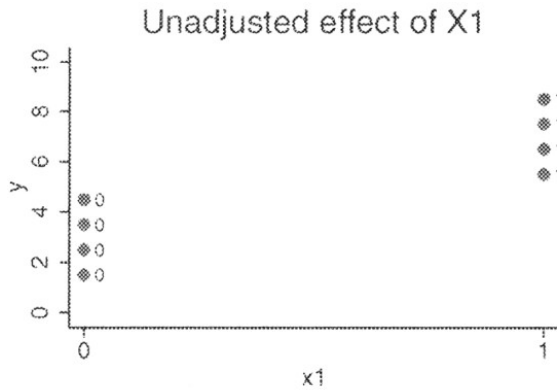
- X_2 is a plausible cause of the outcome Y
(or a proxy for such determinants)
- X_2 is also a plausible cause of predictor X_1
(or they share a common causal determinant)

Confounding patterns

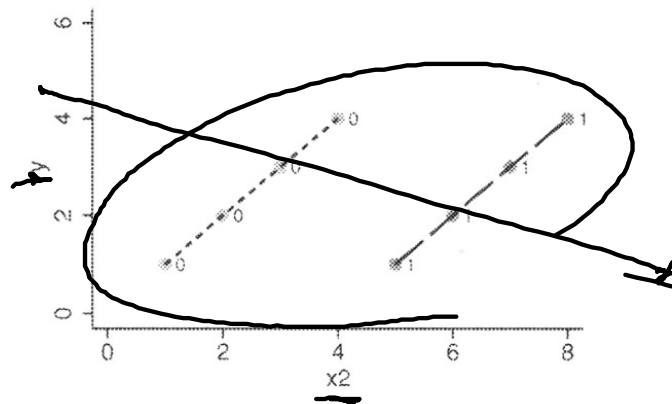
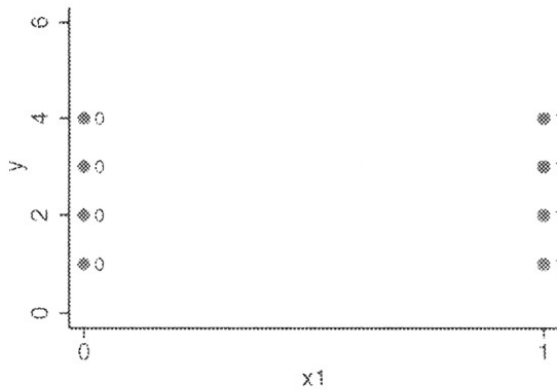
Simpson paradox



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 be estimated by comparing the means of exposed and unexposed within levels of X_2

In practice this is obtained by fitting the linear model

$$E(Y|X) \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

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

AIC
↓
20.57

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

In particular:

Suppose that the true model is given by

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad (1)$$

~~$U(x_1, x_2) \rightarrow 0$~~

but the data are analyzed with a model omitting x_2 , thus as

$$E(Y) = a + bx_1 \quad (2)$$

We then have

$$\hat{b} = \hat{\beta}_1 + \hat{\beta}_2 r_{12} \frac{s_2}{s_1} \quad (3)$$

where \hat{b} is the least squares estimate of b under model (2),

$\hat{\beta}_1$ and $\hat{\beta}_2$ are least squares estimates of model (1),

r_{12} is the Pearson correlation between x_1 and x_2

and the s_j the empirical standard deviations of x_1 and x_2

It follows:

When the two covariates are correlated, $r_{12} \neq 0$,
and when there is a causal effect of x_2 on Y , so $\hat{\beta}_2 \neq 0$,
then we estimate different effects of x_1 under model (1) and (2),
that is: $\hat{b} \neq \hat{\beta}_1$

However when the two covariates are weakly correlated, $r_{12} \approx 0$,
or when there is no important causal effect of x_2 on Y , so $\hat{\beta}_2 \approx 0$,
then the estimates differ little, $\hat{b} \approx \hat{\beta}_1$

From equation (3) it follows that inclusion of a new covariate x_2 can
both make the association between x_1 and Y weaker as well as
stronger.

Example (contd)

We will demonstrate equation (3) on the glucose data. Note that BMI had a strongly significant association with glucose. It turns out that BMI is the essential confounder of the exercise.

↳ However, there were 2 subjects with unknown (missing) BMI. These have to be removed from the data before the comparison

R code for removing the missing:

↳ hers.nob=hers.no[!is.na(hers.no\$BMI),]

We then fit models with and without BMI:

```
fit.a=lm(glucose~exercise,data=hers.nob)
```

```
fit.a$coef
```

```
(Intercept)
```

```
97.370059
```

```
exercise
```

```
-1.701807
```

```
fit.b=lm(glucose~exercise+BMI,data=hers.nob)
```

```
fit.b$coef
```

```
(Intercept)
```

```
83.9422021
```

```
exercise
```

```
-0.9172885
```

```
BMI
```

```
0.4736147
```

Example (contd)

We then calculate the correlation between, and the standard deviations of, exercise and BMI

```
>r12=cor(hers.nob$exercise,hers.nob$BMI)
```

```
>r12
```

```
-0.1587467
```

```
>s1=sd(hers.nob$exercise)
```

```
>s1
```

```
0.4927197
```

```
>s2=sd(hers.nob$BMI)
```

```
>s2
```

```
5.141301
```

$$-0.92 + 0.47 \cdot (-0.15) \cdot \frac{5.14}{0.49}$$

$$-0.92 - 0.78 = -1.70$$

$$\frac{5.14 \cdot 0.15}{0.49} = \frac{2620}{5140} = 0.7760$$

Finally we demonstrate that equation (3) holds in the example

$$\hat{b} = \hat{\beta}_1 + \hat{\beta}_2 r_{12} \frac{s_2}{s_1}$$

```
fit.b$coef[2]+fit.b$coef[3]*r12*s2/s1
```


```
exercise
```

```
-1.701807
```

The answer is identical to the estimate \hat{b} for the simple model!

Control of confounding

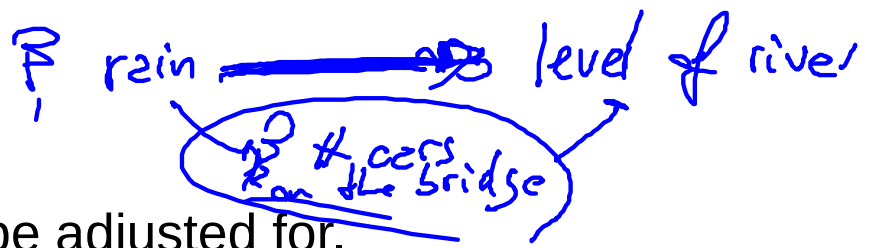
If all confounding variables are recorded and included adequately in a multiple regression model we should then identify the causal effects also in an observational study.

 But there is of course no way we can know that all confounders have been identified and measured without error.

We should therefore be cautious about concluding about causal effects from observational studies.

Still we may hope that we are closer to identifying causality after adjusting (or controlling) for known confounders

Mediation, Sec. 4.5



Not all measured variables should be adjusted for.

Example: Statin drugs may reduce (bad) cholesterol which in turn may reduce risk of heart attack.

Adjusting for cholesterol measured after taking statins may then hide a causal effect of statins on risk of heart attack.

In this case cholesterol is a mediator, or intermediate variable. It is likely correlated (caused by) to statin use and causally related to heart attack. However, since it is on the causal pathway between statin use and heart attack we should not adjust for it.

statin → lower cholesterol → reduced risk of heart attack

heart attack ~ statin + ~~cholesterol~~

Why randomization works

In a study where subjects are randomized to different treatments we can ignore confounding.

This can be deduced from equation (3) $\hat{b} = \hat{\beta}_1 + \hat{\beta}_2 \frac{r_{12} s_2}{s_1}$

After randomization the treatment x_1 and the confounder x_2 will be (approximately) uncorrelated, thus $r_{12} \approx 0$ and $\hat{b} \approx \hat{\beta}_1$.

Hence the causal effect is estimated after randomization!

We don't even need to know the confounding factors

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

Handwritten annotations: Blue arrows point from the text to the corresponding rows in the table. The text "only HT" points to the row (1, 0). The text "only static" points to the row (0, 1). The text "both" points to the row (1, 1).

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	$X_1 X_2$	$E(y \mathbf{x})$
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)
```

$lm(LDL1 \sim HT * statins, data=)$

$$\beta_0 + \beta_1 HT + \beta_2 statins + \beta_3 HT * statins$$

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

$\beta_0 + \beta_1$

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
```

$$\begin{array}{r} -31.54 + \\ \hline 8.24 \\ \hline -25.30 \end{array}$$

$$\begin{array}{r} -17.73 \\ -13.81 \\ \hline +6.24 \end{array}$$

R output (edited):

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

Another options for interpreting interactions can be to construct a new categorical variable with one level for each combination of levels of the original factors.

In the Hypertension-Statin example we construct a variable with 4 levels:

- Level 1: HT=0 and statins=0
- Level 2: HT=1 and statins=0
- Level 3: HT=0 and statins=1
- Level 4: HT=1 and statins=1

```
hers$HTstat=1*(hers$HT==0&hers$statins==0)+2*(hers$HT==1&hers$statins==0)
+3*(hers$HT==0&hers$statins==1)+4*(hers$HT==1&hers$statins==1)
hers$HTstat=factor(hers$HTstat)
```

```
ht.fit.b=lm(LDL1~HTstat, data=hers)
summary(ht.fit.b)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	145.2	1.33	109.5	< 2e-16 ***
HTstat2	-17.7	1.87	-9.5	< 2e-16 ***
HTstat3	-13.8	2.15	-6.4	1.65e-10 ***
HTstat4	-25.3	2.20	-11.5	< 2e-16 ***

Level 4 estimates the effect of HT=1 and statins=1 compared to HT=statins=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

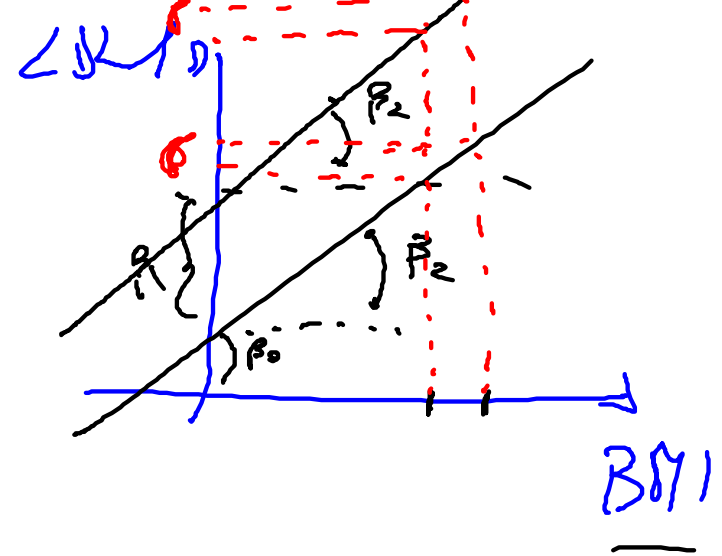
It may be of interest to consider a model where the effect of BMI may differ between statin users and those who don't use statins, i.e. where there is an *interaction*

We then consider the model

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

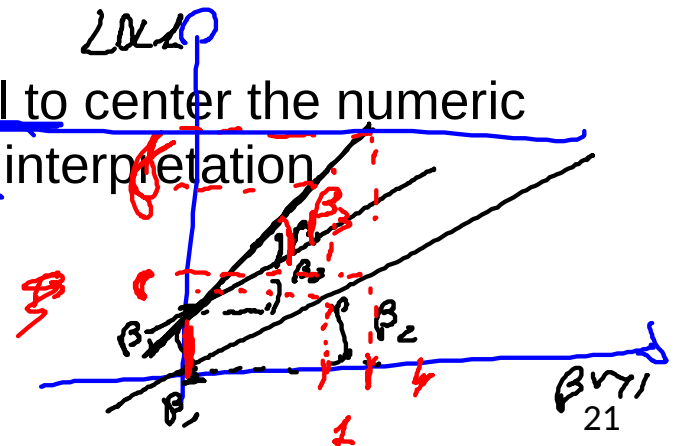
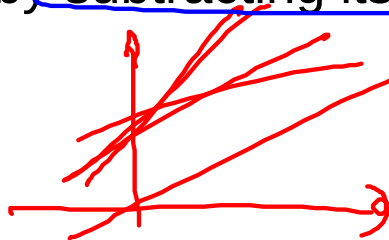
Note that the model may be written

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



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

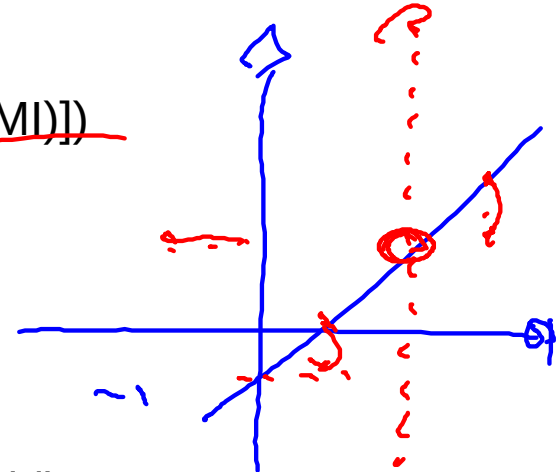
When considering such a model, it is useful to center the numeric covariate (by subtracting its mean) to ease interpretation.



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

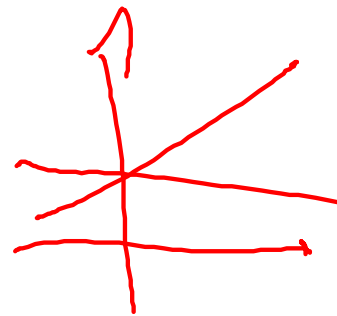
R commands:

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



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



$$\frac{0.640}{+0.721}$$

$$\rightarrow 0.89$$

$$BMI - BMI$$

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:

$$x_{1i} = \begin{cases} 0 & \text{if individ } i \text{ has level } a_1 \text{ for factor A (reference)} \\ 1 & \text{if individ } i \text{ has level } a_2 \text{ for factor A} \end{cases}$$

$$x_{2i} = \begin{cases} 0 & \text{if individ } i \text{ has level } b_1 \text{ for factor B (reference)} \\ 1 & \text{if individ } i \text{ has level } b_2 \text{ for factor B} \end{cases}$$

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):

$$x_{2i} = \begin{cases} 1 & \text{if individ } i \text{ has level } b_2 \text{ for factor B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{if individ } i \text{ has level } b_3 \text{ for factor B} \\ 0 & \text{otherwise} \end{cases}$$



A model with interaction then takes the form

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

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

We recapitulate:

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

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

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

We may then rewrite model (*) as

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

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

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

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

catalyst

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

Temp.

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$fcats=factor(polymer$cat)
fit=lm(rate~ftemp+fcats+ftemp:fcats,data=polymer)
summary(fit)
```

0 0 1
0 1 1

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
fcats0.6	6.5	1.73	3.75	0.005
fcats0.7	18.5	1.73	10.68	2.06e-06
ftemp60:fcats0.6	6.5	2.45	2.65	0.026
ftemp70:fcats0.6	6.0	2.45	2.45	0.037
ftemp60:fcats0.7	7.5	2.45	3.06	0.014
ftemp70:fcats0.7	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 $J \times K$ 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):

$$\underline{TSS} = \underline{SSA} + \underline{SSB} + \underline{SSAB} + \underline{RSS}$$

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 not balanced, the decomposition of the total sum of squares is not 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	$SSA / (J - 1)$	$F = \frac{SSA / (J - 1)}{RSS / (n - JK)}$
Factor B	$K - 1$	SSB	$SSB / (K - 1)$	$F = \frac{SSB / (K - 1)}{RSS / (n - JK)}$
Interaction	$(J - 1)(K - 1)$	$SSAB$	$SSAB / (J - 1)(K - 1)$	$F = \frac{SSAB / (J - 1)(K - 1)}{RSS / (n - JK)}$
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:

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

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

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

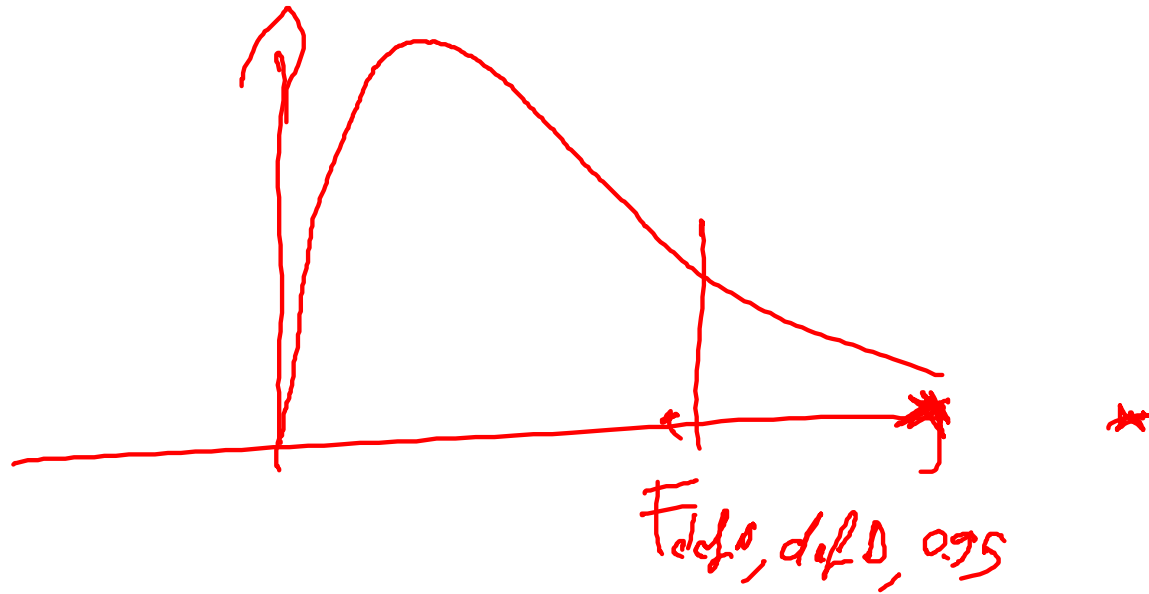
For the example:

R commands:

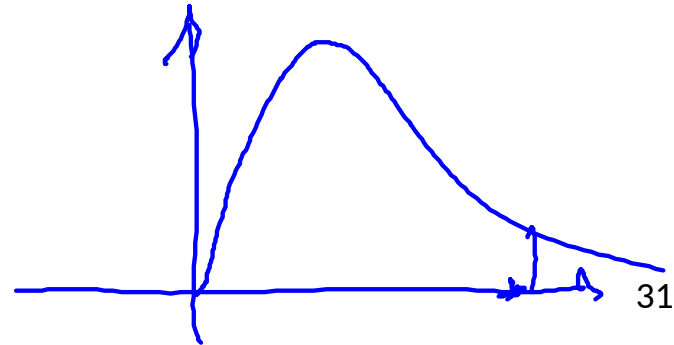
anova(fit)

R output:

Analysis of Variance Table



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



Higher level ANOVA

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

Data:

y_{ijkl}

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		<u>SSA</u>	<u>SSA / df</u>	<u>F_A</u>
Factor B		<u>SSB</u>	<u>SSB / df</u>	<u>F_B</u>
Factor C		<u>SSC</u>	<u>SSC / df</u>	<u>F_C</u>
Interaction AB		<u>SSAB</u>	<u>SSAB / df</u>	<u>F_{AB}</u>
Interaction AC		<u>SSAC</u>	<u>SSAC / df</u>	<u>F_{AC}</u>
Interaction BC		<u>SSBC</u>	<u>SSBC / df</u>	<u>F_{BC}</u>
Interaction ABC		<u>SSABC</u>	<u>SSABC / df</u>	<u>F_{ABC}</u>
Residual		<u>RSS</u>	<u>RSS / df</u>	
Total	$n - 1$	<u>TSS</u>		

*) 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.

$$\text{volume} = \beta_0 + 1.02 \text{ height} + 2.00 \text{ diameter} + \epsilon$$

Expected values and prediction with new covariate

Example: Consider a new tree with measured diameter (and height)

What is the expected volume of the tree?

How certain is the estimate of the expected volume?

How certain are we about the volume of the actual tree?

Example: Systolic blood pressure and age

What is the expected blood pressure at age 50?

What is the confidence interval for this expected blood pressure?

What is the level of uncertainty in blood pressure for a new patient aged 50 years?

The confidence intervals for the expected values will only depend on uncertainties in the estimated regression coefficients.

The prediction intervals for new observations also requires the individual variation!

Confidence intervals expected values

Consider a new covariate vector $\mathbf{x}^{new} = (x_1^{new}, x_2^{new}, \dots, x_p^{new})$

The expected outcome with this covariate is given by

$$\mu^{new} = \beta_0 + \beta_1 x_1^{new} + \beta_2 x_2^{new} + \dots + \beta_p x_p^{new}$$

which is naturally estimated by plugging in least squares estimates:

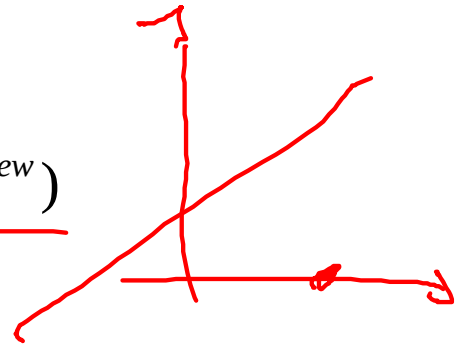
$$\hat{\mu}^{new} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{new} + \hat{\beta}_2 x_2^{new} + \dots + \hat{\beta}_p x_p^{new}$$

The variance of $\hat{\mu}^{new}$ only depends on the variances of (and covariances between) the least squares parameter estimates and with standard error $se(\hat{\mu}^{new})$ for $\hat{\mu}^{new}$ we have that

$$t = \frac{\hat{\mu}^{new} - \mu^{new}}{se(\hat{\mu}^{new})} \sim t_{n-p-1}$$

i.e. t-distributed with $n-p-1$ degrees of freedom and a CI of μ^{new} is given as

$$\hat{\mu}^{new} \pm c \cdot se(\hat{\mu}^{new})$$



Prediction intervals for a new outcome

A new outcome with the covariate $\mathbf{x}^{new} = (x_1^{new}, x_2^{new}, \dots, x_p^{new})$ is given as

$$Y^{new} = \mu^{new} + \varepsilon^{new} = \beta_0 + \beta_1 x_1^{new} + \beta_2 x_2^{new} + \dots + \beta_p x_p^{new} + \varepsilon^{new}$$

where the new error term ε^{new} is independent of the previous data and so of the least squares parameter estimates.

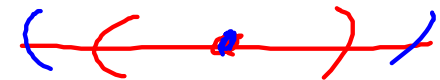
The natural point estimate (best guess) for Y^{new} also equals

$$\hat{\mu}^{new} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{new} + \hat{\beta}_2 x_2^{new} + \dots + \hat{\beta}_p x_p^{new}$$

$$\varepsilon \sim N(0, \sigma^2)$$
$$E(\varepsilon) = 0$$

But an interval for the prediction of the new outcome also needs to incorporate the random noise ε^{new} and so becomes

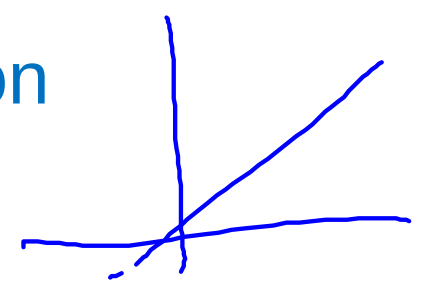
$$\hat{\mu}^{new} \pm c \cdot \sqrt{s_{Y|x}^2 + se(\hat{\mu}^{new})^2}$$



where again the c is a percentile in the t-distribution with $n-p-1$ degrees of freedom.

In particular for simple linear regression

$$\text{Var}(\hat{\mu}^{new}) = \sigma_{\varepsilon}^2 \left(\frac{1}{n} + \frac{(x^{new} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)$$



Hence the confidence interval for the expected value $\mu^{new} = \beta_0 + \beta_1 x^{new}$ becomes

$$\hat{\mu}^{new} \pm c \cdot s_{Y|x} \sqrt{\frac{1}{n} + \frac{(x^{new} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

P.i. $+ \varepsilon_i$
 $\text{Var}(\varepsilon_i) = \sigma_{\varepsilon}^2$

whereas the prediction interval for the new outcome value is given as

$$\hat{\mu}^{new} \pm c \cdot s_{Y|x} \sqrt{1 + \frac{1}{n} + \frac{(x^{new} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

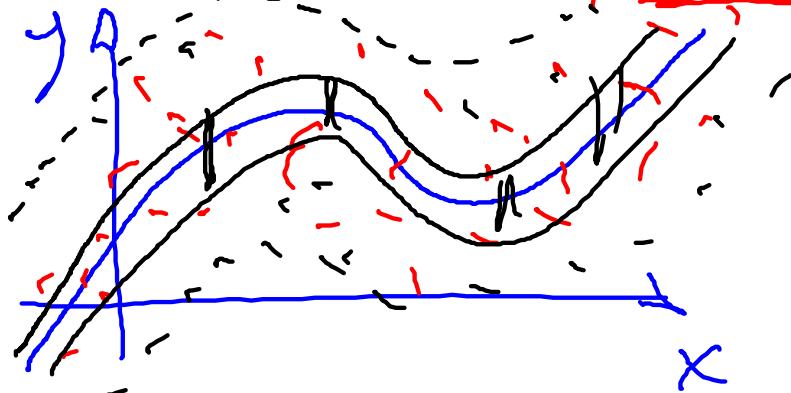
Note that both the confidence and prediction intervals are most narrow when $x^{new} = \bar{x}$

Example: Blood pressure and age

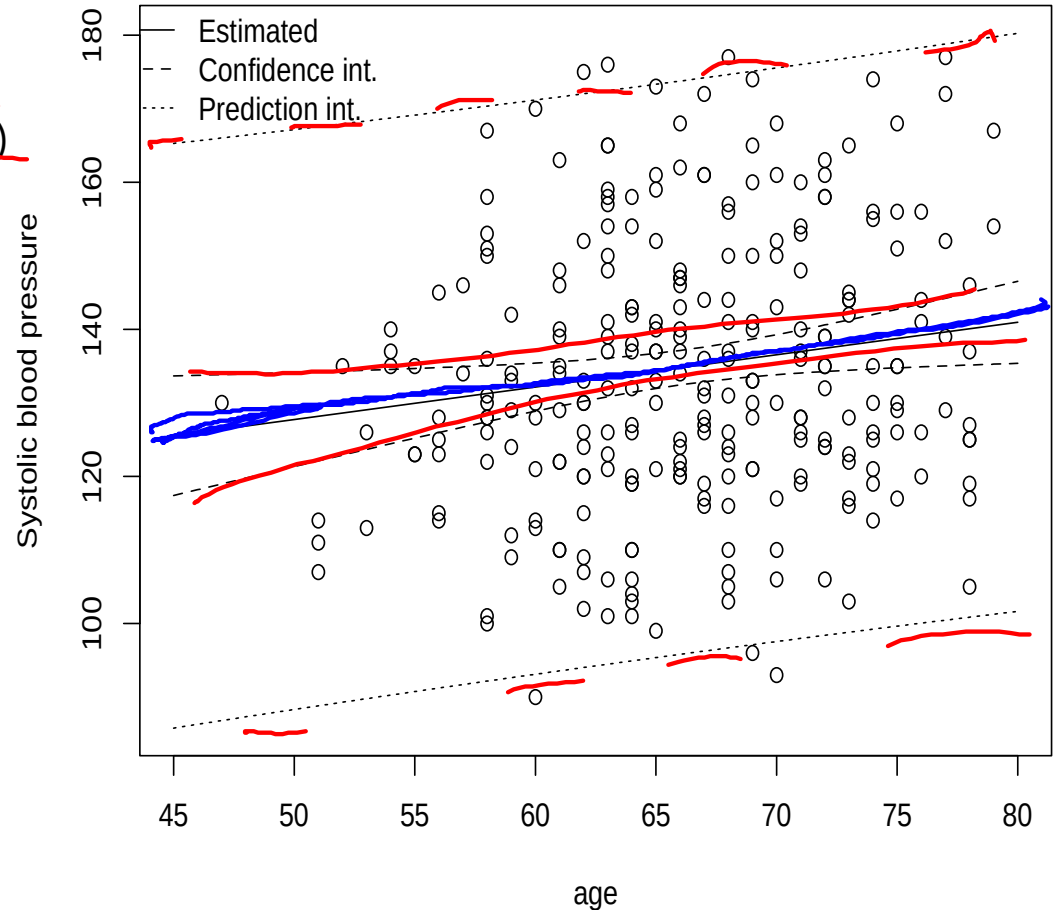
$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \text{age}$$

R commands:

```
sbpage=lm(sbp~age,data=hers.sample)
age=45:80
newage=as.data.frame(age)
estsbp=predict(sbpage,newage,int="conf")
predsbp=predict(sbpage,newage,int="pred")
```

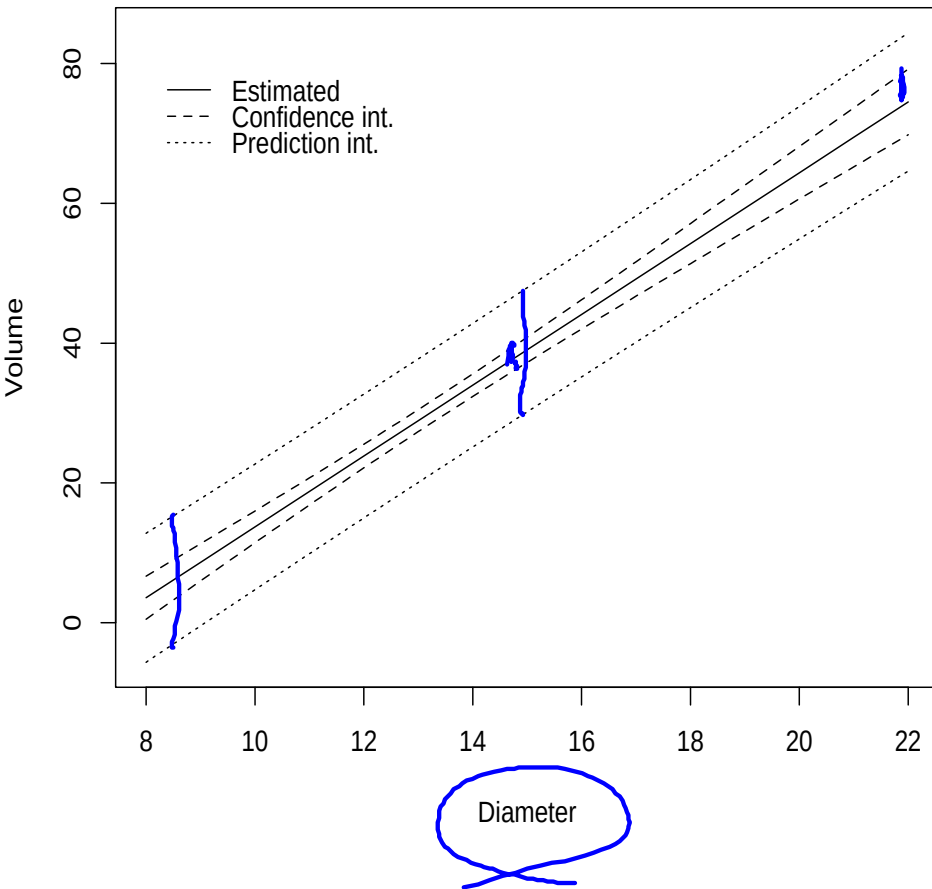


```
yint=c(min(predsbp[,2]),max(predsbp[,3]))
plot(age,estsbp[,1],type="l",ylim=yint,
      ylab="Systolic blood pressure")
lines(age,estsbp[,2],lty=2)
lines(age,estsbp[,3],lty=2)
lines(age,predsbp[,2],lty=3)
lines(age,predsbp[,3],lty=3)
points(hers.sample)
legend(42,186,c("Estimated","Confidence
int.,"Prediction int."),lty=1:3,bty="n")
```



Example: Diameter and tree volume

Model: Volume = $b_0 + b_1$ Diameter



Model: Volume = $b_0 + b_1$ Diameter + b_2 Diameter²

