

**PhD course in Basic Biostatistics - Day 5**

Erik Parner, Department of Biostatistics, Aarhus University®

**Regression models in general****The simple linear regression**

Lung function (PEFR) and height

The model, estimation, inference

Changing the reference height

Checking the model: predicted values and residuals

Point wise confidence and prediction intervals

**Comparing two groups after adjustment for a covariate**

Sex difference in PEFR

**Correlations**

The Pearson correlation

The Spearman rank correlation

**Why you should not use correlation in the comparison of measurement methods**

01-10-2015

Basic Biostatistics - Day 5

1

**Overview**

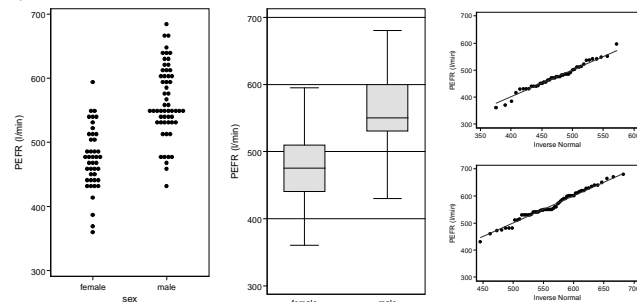
Data to analyse	Type of analysis	Unpaired/Paired	Type	Day
Continuous	One sample mean	Irrelevant	Parametric	Day 1
			Nonparametric	Day 3
	Two sample mean	Non-paired	Parametric	Day 2
			Nonparametric	Day 2
		Paired	Parametric	Day 3
			Nonparametric	Day 3
	<b>Regression</b>	<b>Non-paired</b>	<b>Parametric</b>	<b>Day 5</b>
	Several means	Non-paired	Parametric	Day 6
			Nonparametric	Day 6
	Binary	Irrelevant	Parametric	Day 4
			Parametric	Day 4
		Paired	Parametric	Day 4
Time to event	Regression	Non-paired	Parametric	Day 7
			Nonparametric	Day 8
	Regression: Rate/hazard ratio	Non-paired	Semi-parametric	Day 8

Correlation is seen as a topic associated with regression.

01-10-2015

Basic Biostatistics - Day 5

2

**Lung function men and women (Exampel 4 later)****Question:** How does the *PEFR* differ for men and women ?

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
female	43	474.0698	7.4829	49.0687	458.9687	489.171
male	57	564.2807	7.4236	56.0471	549.4094	579.152
diff		<b>-90.21093</b>	10.73949		<b>-111.5231</b>	<b>-68.89877</b>

01-10-2015

Basic Biostatistics - Day 5

3

**Example: Lung function men and women****Question:** How does the *PEFR* differ for men and women?**First answer:**The mean lung function among men is **90(69;112)l/min** larger than among women!**BUT:**

We know that PEFR depends on height and that men are higher than women (in average).

How much of the above difference can explained by this ?

How large is the "height adjusted" difference in *PEFR* ?**In the regression model we aim at comparing men and women with the same height (adjusting for height).**

01-10-2015

Basic Biostatistics - Day 5

4

### Regression in general

A regression model can be many things!

In general it **models** the relationship between:

$y$ : **dependent**/response

and a set of

$x$ 's: **independent**/explanatory variables.

The dependent variable is **modelled** as a function of the independent variable plus some unexplained random variation:

$$y = f(x; \theta) + e(\sigma)$$

Diagram illustrating the components of the regression model equation  $y = f(x; \theta) + e(\sigma)$ :

- Systematic part**:  $f(x; \theta)$
- Random part**:  $e(\sigma)$
- Unknown Parameters**:  $\theta$  (in  $f(x; \theta)$ ) and  $\sigma$  (in  $e(\sigma)$ )

01-10-2015

Basic Biostatistics - Day 5

5

### Regression in general

$$y = f(x; \theta) + e(\sigma)$$

Some examples:

$$pefr = \beta_0 + \beta_1 \cdot height + E$$

$$pefr = \beta_0 + \beta_1 \cdot height + \beta_2 \cdot height^2 + E \quad \text{and } E \sim N(0, \sigma^2)$$

$$gfr = \exp(\beta_0 + \beta_1 \cdot \ln[Cr]) + E$$

$$conc(t) = dose \cdot V \cdot [\exp(-\lambda_{abs} \cdot t) - \exp(-\lambda_{eli} \cdot t)] + E$$

The first **two** are **linear** regressions, the last two **non-linear**.

In this course we will **focus** on the **linear** regressions.

01-10-2015

Basic Biostatistics - Day 5

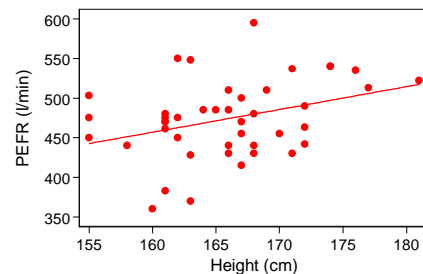
6

### Example: lung function and height

**Purpose:** Describe the association between lung function and height among young women:

**Data:**  $PEFR$  (l/min) and  $height$  (cm) for 43 female medical students.

Figure 5.1



**A model:**  $PEFR = \text{line} + \text{some random variation}$  seems to be valid.

01-10-2015

Basic Biostatistics - Day 5

7

### Simple linear regression: The model

Let  $PEFR_i$  and  $height_i$  be the data for the  $i$ th woman.

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

This model is based on the **assumptions**:

1. The **expected** value of  $PEFR$  is a **linear function** of  $height$ .
2. The **unexplained** random deviations are **independent**.
3. The unexplained random deviations have the **same distributions**.
4. This distribution is **normal**.

01-10-2015

Basic Biostatistics - Day 5

8

## Simple linear regression: The parameters

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

The model have **three** unknown parameters:

1. The **intercept**  $\beta_0$
2. The **slope** (or **regression coefficient**)  $\beta_1$
3. The **residual variance**  $\sigma^2$  or **residual standard deviation**  $\sigma$ .

The **interpretation** of the parameters:

$\beta_0$  is expected **PEFR** of a woman with **height**=0.

Obviously, this does not make sense.

We will later look at how one can get a meaningful estimate of the general level of **PEFR** !

01-10-2015

Basic Biostatistics - Day 5

9

## Simple linear regression: The parameters

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

$\beta_1$  is the **expected difference** in **PEFR** for two women, who differ with **one unit** (here cm) in **height**.

If a woman is 6 cm higher than another, then we will expect that her **PEFR** is  $6\beta_1$  higher than the other.

$\sigma$  is best understood by the fact that a **95%-prediction** interval around the line is given by  $\pm 1.96\sigma$ .

01-10-2015

Basic Biostatistics - Day 5

10

## Simple linear regression: The estimates (by hand)

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

Estimates of the parameters are found by the method of **least square**, which, for this model, is equivalent to the **maximum likelihood** method.

The estimates are found using a computer program;

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2} = \sqrt{\frac{1}{n-2} \sum r_i^2}$$

01-10-2015

Basic Biostatistics - Day 5

11

## Simple linear regression: The standard errors

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$se(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

## Some comments:

The precision of the estimate of  $\beta_1$  and  $\beta_0$  depends on the size of the variation around the line, i.e.  $\sigma$ .

- The precision of the estimate of  $\beta_1$  **increases** with increasing variation of the  $x$ 's.
- The precision of the estimate of  $\beta_0$  **increases** with increasing variation of the  $x$ 's and with sample size.
- The precision of the estimate of  $\beta_0$  **decreases** with increasing average of the  $x$ 's.

01-10-2015

Basic Biostatistics - Day 5

12

### Simple linear regression: Confidence intervals

Approx. 95% CI for  $\beta_1$ :  $\hat{\beta}_1 \pm 1.96 \cdot \text{se}(\hat{\beta}_1)$

Approx. 95% CI for  $\beta_0$ :  $\hat{\beta}_0 \pm 1.96 \cdot \text{se}(\hat{\beta}_0)$

Exact 95% confidence intervals, CI's, for  $\beta_0$  and  $\beta_1$  are found from the estimates and standard errors

95% CI for  $\beta_1$ :  $\hat{\beta}_1 \pm t_{n-2}^{0.975} \cdot \text{se}(\hat{\beta}_1)$

95% CI for  $\beta_0$ :  $\hat{\beta}_0 \pm t_{n-2}^{0.975} \cdot \text{se}(\hat{\beta}_0)$

Where  $t_{n-2}^{0.975}$  is the upper 97.5 percentile in the t-distribution  $n-2$  degrees of freedom.

These confidence intervals are found in the output.

01-10-2015

Basic Biostatistics - Day 5

13

### Simple linear regression: test

As usual we can perform a test of hypothesis of the type:

Hypothesis:  $\beta_i = \beta_i^0$

by calculating  $z_{obs} = \frac{\hat{\beta}_i - \beta_i^0}{\text{se}(\hat{\beta}_i)}$

The p-value is found by checking a t-distribution  $n-2$  degrees of freedom.

$$p = 2 \cdot \Pr(t(n-2) \geq |z_{obs}|)$$

You will find tests for

$\beta_1 = 0$ , i.e.  $y$  is independent of  $x$

and

$\beta_0 = 0$ , i.e. the line goes through  $(0, 0)$

in the **regression output**.

01-10-2015

Basic Biostatistics - Day 5

14

### Stata: Simple linear regression

```
regress PEFR height if sex==1
```

Source   SS df MS				
Model	12116.1257	1	12116.1257	
Residual	89008.665	41	2170.94305	
Total	101124.791	42	2407.73311	

pefr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height	2.871025	1.215288	2.36	0.023	.4167005 5.325349
_cons	-2.38683	201.8064	-0.01	0.991	-409.9432 405.1696

$n$ : Always check this

43

 $\hat{\beta}_1$  $\hat{\beta}_0$  $\hat{\sigma}^2$  $\hat{\sigma}$ 

Standard errors

95% confidence intervals

01-10-2015

Basic Biostatistics - Day 5

15

### The example: Summarising

$$PEFR_i = \beta_0 + \beta_1 \cdot \text{height}_i + E_i \quad E_i \sim N(0, \sigma^2)$$

The estimates:  $\beta_1$ : 2.87 (0.42;5.33) l/min/cm

$\beta_0$ : -2.39 (-410;405) l/min

$\sigma$ : 46.6 l/min

The difference in mean PEFR between two women who differ one cm in height is in interval from 0.42 to 5.33 l/min - the best guess is 2.87 l/min.

The mean PEFR for a woman who is 0 cm is in the interval -410 to 405 l/min - the best guess is -2.39 l/min.

A 95% prediction interval is given as  $\pm 93$  l/min.

01-10-2015

Basic Biostatistics - Day 5

16

**Stata: changing the intercept**

$$PEFR_i = \beta_0 + \beta_1 \cdot (\text{height}_i - 170) + E_i \quad E_i \sim N(0, \sigma^2)$$

Let us fit the model with a **meaningful** intercept/constant:

```
generate height170=height-170
regress PEFR height170 if sex==1
```

Source	SS	df	MS	
Model	12116.1257	1	12116.1257	
Residual	89008.665	41	2170.94305	
Total	101124.791	42	2407.73311	

Number of obs = 43  
F( 1, 41) = 5.58  
Prob > F = 0.0230  
R-squared = 0.1198  
Adj R-squared = 0.098  
Root MSE = 46.593

PEFR	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height170	2.871025	1.215288	2.36	0.023	.4167005 5.325349
_cons	485.6874	8.641215	56.21	0.000	468.2361 503.1387

Nothing is changed except this

The expected PEFR for a woman with height = 170cm is:  
**486 (468;503) l/min**

01-10-2015 Basic Biostatistics - Day 5 17

**Predicted values and residuals**

$$Y_i = \beta_0 + \beta_1 \cdot x_i + E_i \quad E_i \sim N(0, \sigma^2)$$

Based on the estimates we can calculate the **predicted** (fitted) values and the **residuals**:

Predicted value:  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$

Residual:  $r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)$

The **predicted value** is the best guess of  $y_i$  (based on the estimates) for the  $i$ th person.

The **residual** is a guess of  $E_i$  (based on the estimates) for the  $i$ th person.

```
predict fitfemale if e(sample),xb
predict resfemale if e(sample),resid
```

01-10-2015 Basic Biostatistics - Day 5 18

**Checking the model: Independent errors ?**

**Assumption no. 2:** *the errors should be independent*, is mainly checked by considering **how the data was collected**.

The assumption is **violated** if

- some of the persons are **relatives** (and some are not) and the dependent variable has some **genetic** component.
- some of the persons were **measured** using one instrument and others using another.
- in general if the persons were sampled in **clusters**.

01-10-2015 Basic Biostatistics - Day 5 19

**Checking the model:  
Linearity and identical distributed errors**

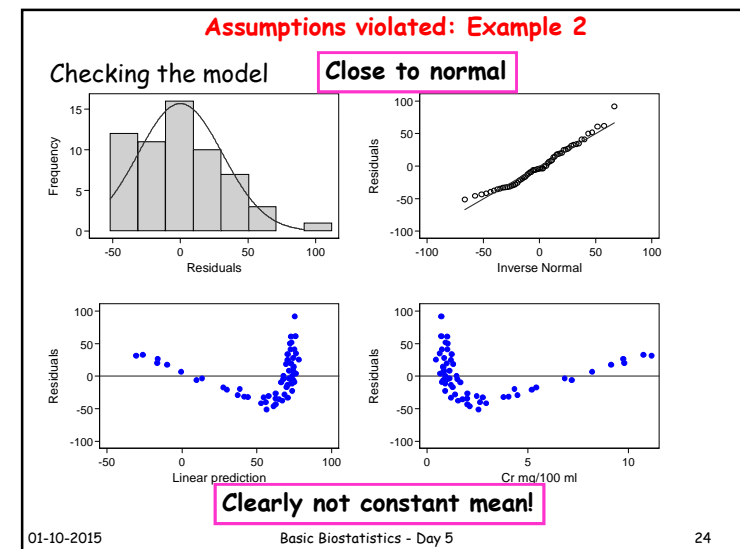
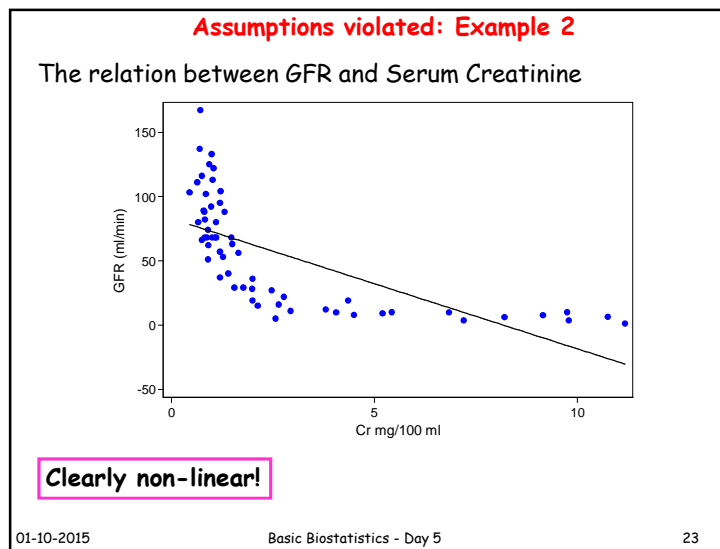
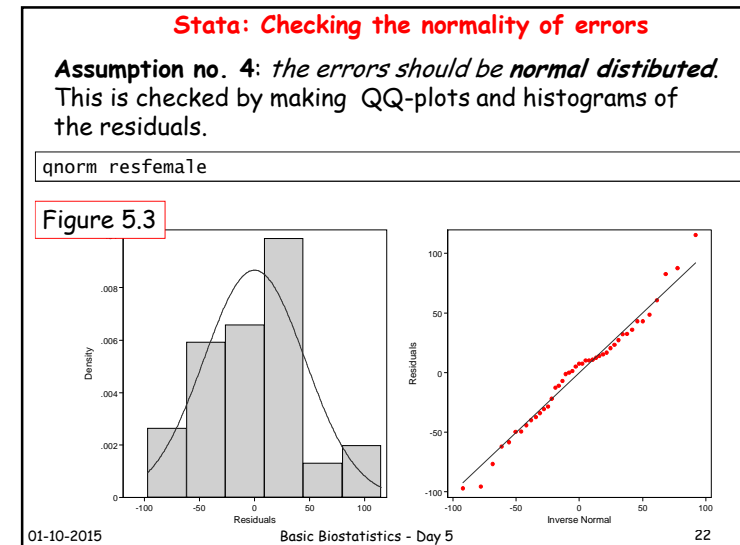
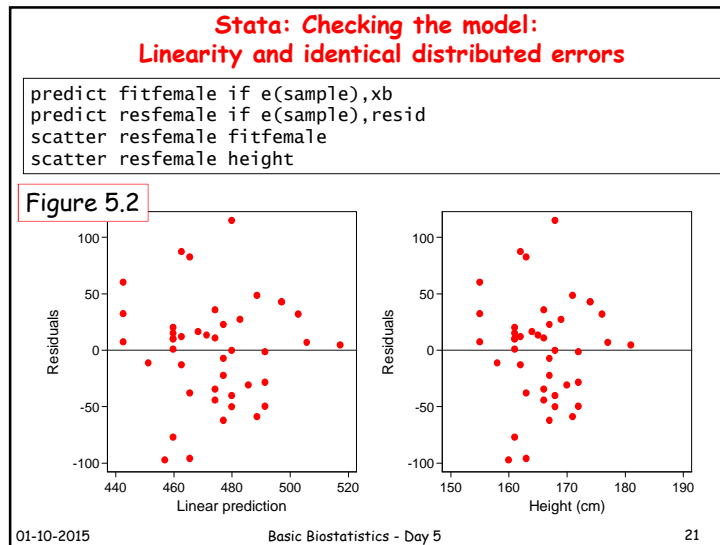
**Assumption no. 1:**  
The **expected** value of  $Y$  is a **linear** function of  $x$ .

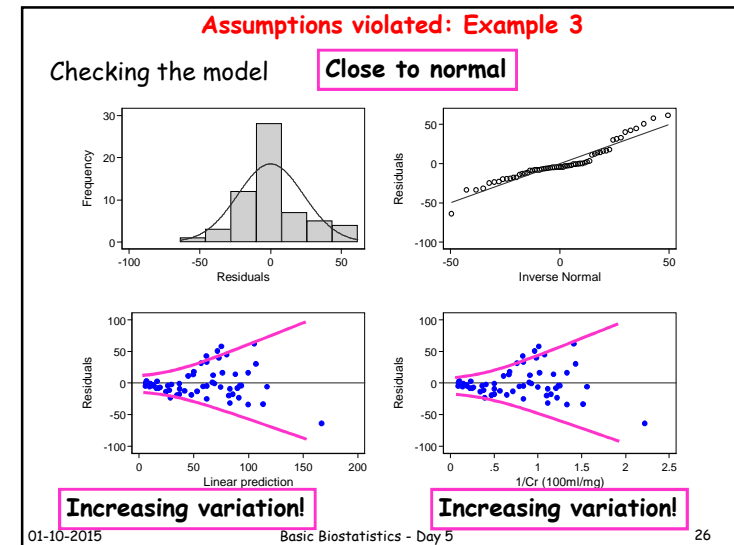
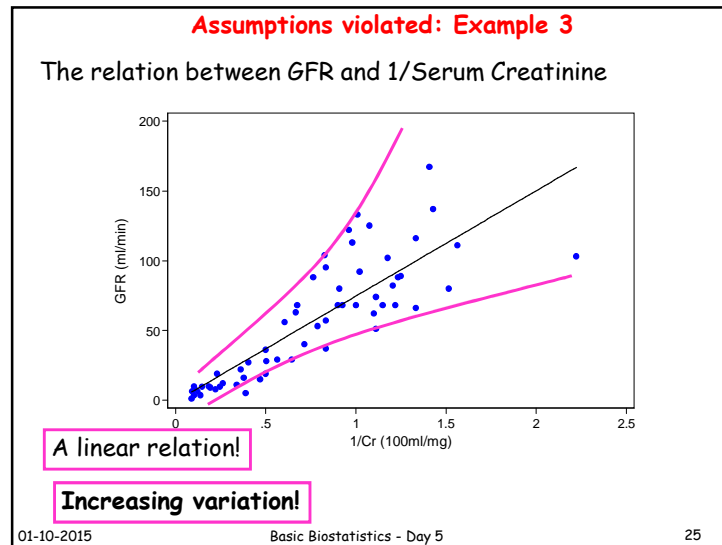
**Assumption no. 3:**  
The unexplained random deviations have the **same distributions**.

These are checked by inspecting the following plots of:

- **Residuals** versus **predicted**
- **Residuals** versus  $x$

01-10-2015 Basic Biostatistics - Day 5 20





**Confidence interval for the estimated line**

The **true line** is given as :  $y = \beta_0 + \beta_1 \cdot x$

and **estimated** by plugging in the estimates  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$

The **standard error** of this estimate is given by:

$$se(\hat{\beta}_0 + \hat{\beta}_1 \cdot x) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

with the 95% (pointwise) **confidence interval**

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x \pm t_{n-2}^{0.975} \cdot se(\hat{\beta}_0 + \hat{\beta}_1 \cdot x)$$

Many programs can make a plot with the fitted line and its confidence limits.

In Stata its done by the **lfitci** graph command.

01-10-2015 Basic Biostatistics - Day 5 27

**Prediction interval for future value**

The **true line** is given as :  $y = \beta_0 + \beta_1 \cdot x$

and **estimated** by plugging in the estimates  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$

The **standard deviation** for a **new observation** is given by:

$$sd(\hat{\beta}_0 + \hat{\beta}_1 \cdot x + E) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

with the 95% (pointwise) **prediction interval**

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x \pm t_{n-2}^{0.975} \cdot sd(\hat{\beta}_0 + \hat{\beta}_1 \cdot x + E)$$

Many programs can make a plot with the fitted line and its prediction limits.

In Stata its done by the **lfitci** and graph command, the option **stdf**

01-10-2015 Basic Biostatistics - Day 5 28

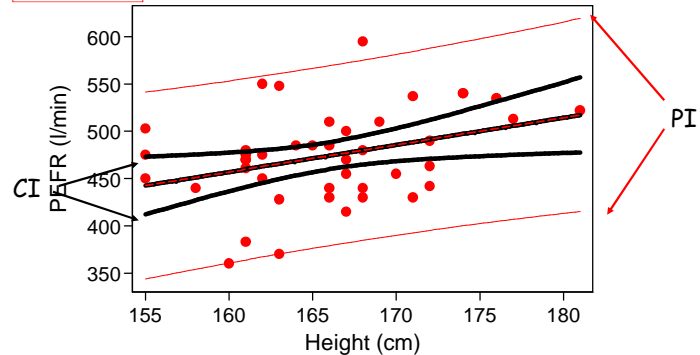
## Stata: graph confidence and prediction intervals

```

twoway (scatter PEFR height if sex==1 ) ///
      (lfitci PEFR height if sex==1 ) ///
      (lfitci PEFR height if sex==1, stdf )

```

Figure 5.4



01-10-2015

Basic Biostatistics - Day 5

29

## Example 4: Lung function men and women

**Question:** How does the *PEFR* differ for men and women ?

We know that *PEFR* depends on height and that men are higher than women (in average).

How much of the above difference can explained by this ?

How large is the "height adjusted" difference in *PEFR* ?

Note, we can only adjust for height, if the *PEFR* - height relationship is the same for men and women.

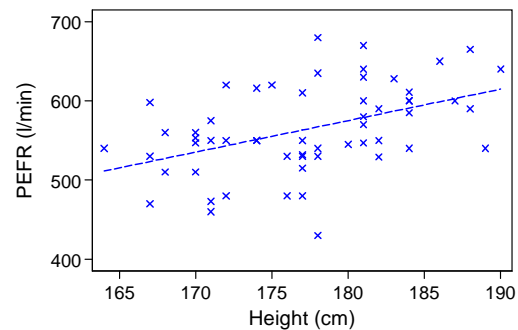
Let us first fit a linear regression to the data for men.

01-10-2015

Basic Biostatistics - Day 5

30

## PEFR and height among males



Root MSE = 50.4

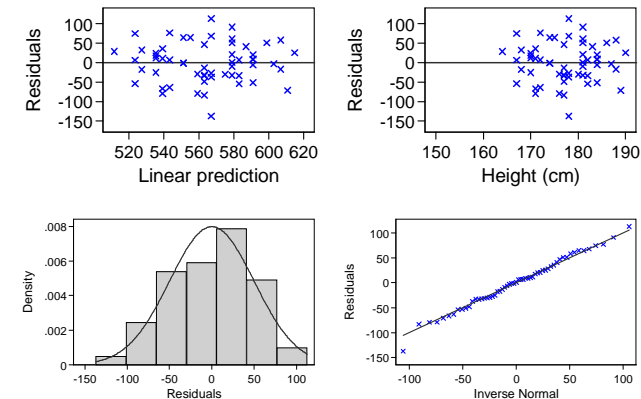
PEFR	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height170	3.974479	1.052755	3.78	0.000	1.864712	6.084247
_cons	535.274	10.17822	52.59	0.000	514.8764	555.6716

01-10-2015

Basic Biostatistics - Day 5

31

## PEFR and height among males: model check



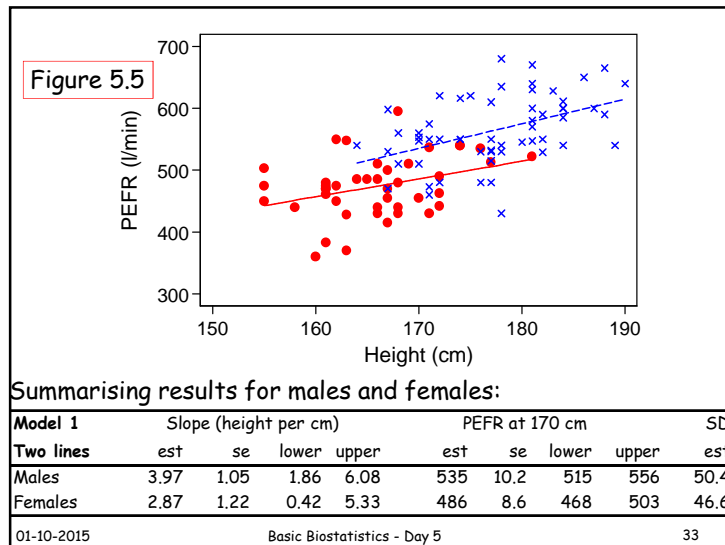
Model ok

01-10-2015

Basic Biostatistics - Day 5

32





Model 1	Slope (height per cm)				PEFR at 170 cm				SD
Two lines	est	se	lower	upper	est	se	lower	upper	est
Males	3.97	1.05	1.86	6.08	535	10.2	515	556	50.4
Females	2.87	1.22	0.42	5.33	486	8.6	468	503	46.6

Here we will focus on the **slopes** and the **intercepts** (PEFR at 170 cm) and **assume** that the size of the unexplained variation is the same for the two sexes, i.e. **identical SD's**. Under this additional assumption we have **Model 2**:

$$PEFR_i = \begin{cases} \beta_0 + \beta_1 \cdot height_i + E_i & \text{female} \\ \alpha_0 + \alpha_1 \cdot height_i + E_i & \text{males} \end{cases} \quad E_i \sim N(0, \sigma^2)$$

Model 2	Slope (height per cm)				PEFR at 170 cm				SD
Same SD	est	se	lower	upper	est	se	lower	upper	est
Males	$\alpha$ 3.97	1.02	1.95	6.00	535	9.9	516	555	48.8
Females	$\beta$ 2.87	1.27	0.34	5.40	486	9.1	468	504	

Only the standard errors, CI's and the SD changed.

01-10-2015 Basic Biostatistics - Day 5 34

$$PEFR_i = \begin{cases} \beta_0 + \beta_1 \cdot height_i + E_i & \text{female} \\ \alpha_0 + \alpha_1 \cdot height_i + E_i & \text{males} \end{cases}$$

If we let  $\delta_0 = \alpha_0 - \beta_0$  and  $\delta_1 = \alpha_1 - \beta_1$  then we can write the model

$$PEFR_i = \begin{cases} \beta_0 + \beta_1 \cdot height_i + E_i & \text{female} \\ (\beta_0 + \delta_0) + (\beta_1 + \delta_1) \cdot height_i + E_i & \text{males} \end{cases}$$

Model 2	Slope (height per cm)				PEFR at 170 cm				SD
Same SD	est	se	lower	upper	est	se	lower	upper	est
Males	$\alpha$ 3.97	1.02	1.95	6.00	535	9.9	516	555	48.8
Females	$\beta$ 2.87	1.27	0.34	5.40	486	9.1	468	504	
Differens	$\delta$ 1.10	1.63	-2.13	4.34	49.6	13.4	23.0	76.2	

The standard errors are based on complicated formulas - the computer does it for you.

01-10-2015 Basic Biostatistics - Day 5 35

$$PEFR_i = \begin{cases} \beta_0 + \beta_1 \cdot height_i + E_i & \text{female} \\ (\beta_0 + \delta_0) + (\beta_1 + \delta_1) \cdot height_i + E_i & \text{males} \end{cases}$$

The same  $PEFR$  -  $height$  relationship for male and females corresponds to  $\delta_1 = 0$ .

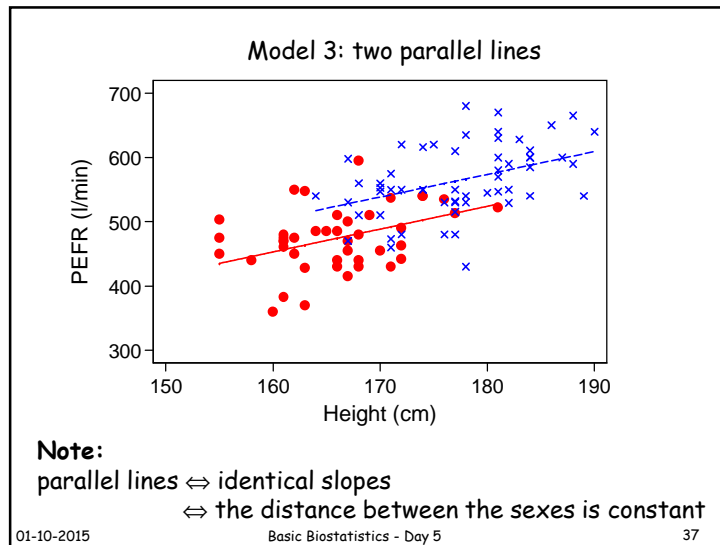
We have the estimate 1.10 (-2.13;4.34)

The confidence interval says this can be accepted (pval=0.55).

**Model 3**  $PEFR_i = \begin{cases} \beta_0 + \beta_1 \cdot height_i + E_i & \text{female} \\ (\beta_0 + \delta_0) + \beta_1 \cdot height_i + E_i & \text{males} \end{cases}$

Model 3	Slope (height per cm)				PEFR at 170 cm				SD
Same Slope	est	se	lower	upper	est	se	lower	upper	est
Males	3.54	0.79	1.97	5.12	538	8.7	521	556	48.7
Females					488	8.1	472	504	
Differens	0.00				50.0	13.3	23.6	76.5	

01-10-2015 Basic Biostatistics - Day 5 36



Model 0		mean PEFR				SD
Two groups		est	se	lower	upper	est
Males		564	7.4	549	579	56.0
Females		474	7.5	459	489	49.1
Differens		90.2	10.7	68.9	111.5	

Model 1		Slope (height per cm)				PEFR at 170 cm				SD
Two lines		est	se	lower	upper	est	se	lower	upper	est
Males		3.97	1.05	1.86	6.08	535	10.2	515	556	50.4
Females		2.87	1.22	0.42	5.33	486	8.6	468	503	46.6

Model 2		Slope (height per cm)				PEFR at 170 cm				SD
Same SD		est	se	lower	upper	est	se	lower	upper	est
Males		3.97	1.02	1.95	6.00	535	9.9	516	555	48.8
Females		2.87	1.27	0.34	5.40	486	9.1	468	504	
Differens		1.10	1.63	-2.13	4.34	49.6	13.4	23.0	76.2	

Model 3		Slope (height per cm)				PEFR at 170 cm				SD
Same Slope		est	se	lower	upper	est	se	lower	upper	est
Males		3.54	0.79	1.97	5.12	538	8.7	521	556	48.7
Females		3.54	0.79	1.97	5.12	488	8.1	472	504	
Differens		0.00				50.0	13.3	23.6	76.5	

01-10-2015 Basic Biostatistics - Day 5 38

### Regression some comments

- The models 2 and 3 are examples of **multiple regression** models:

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + \delta_0 \cdot male + \delta_1 \cdot male \cdot height_i + E_i$$

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + \delta_0 \cdot male + E_i$$

- Notices that the difference between the sexes is smaller after adjustment for the height.
- The methods of adjusting for a continuous variable when comparing two (or several) groups are called **Analysis of Covariance**.

01-10-2015

Basic Biostatistics - Day 5

39

### Stata: summary of regression analysis code

```
use PEFR.dta,clear
* Scatter plot
twoway (scatter PEFR height if sex==1) ///
      (lfit PEFR height if sex==1) ///
* Fitting the regression
generate height170=height-170
regress PEFR height170 if sex==1
* Generating fitted values and residuals
* (the if e(sample) ensures that it is only done for the
* observations actually used in the regression)
predict fitfemale if e(sample), xb
predict resfemale if e(sample), res
scatter resfemale fitfemale
scatter resfemale height
* we will go through the analysis comparing the men and the
* women at the exercises.
* Comparing the slopes:
regress PEFR b1.sex##c.height170
* The height adjusted sex difference.
regress PEFR b1.sex c.height170
```

01-10-2015

Basic Biostatistics - Day 5

40

**Stata: summary of regression analysis code**

b1: sex=1 is set to be the ref.

##: we allow for different slopes

c: height170 is considered continuous with linear effect

```
regress PEFR b1.sex##c.height170
*** output omitted ***
```

	PEFR	Coef.	Std. Err.	t	P> t	[95%]
sex						
male		49.58657	13.38325	3.71	0.000	23.1
height170		2.871025	1.273115	2.26	0.026	.34
sex#c.height170						
male		1.103455	1.631048	0.68	0.500	-2.1
_cons		485.6874	9.052385	53.65	0.000	467

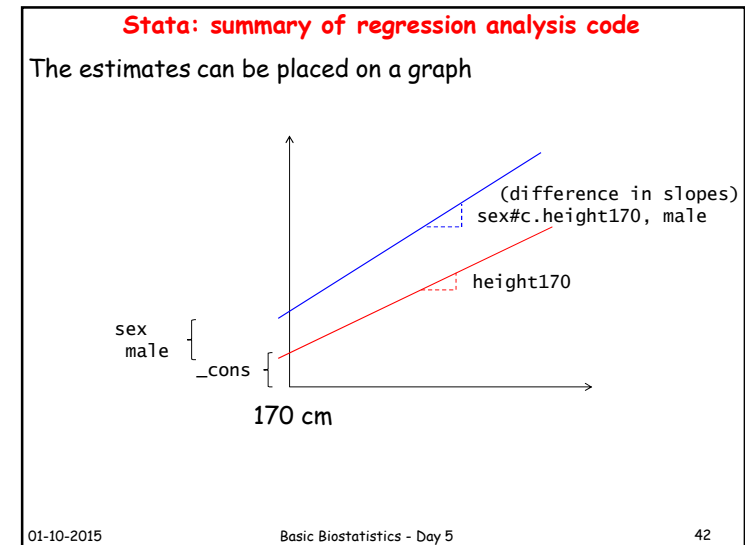
Difference for men and women at 170 cm

Slope for women

Difference in slope for men and women

Expected value for women 170 cm

01-10-2015 Basic Biostatistics - Day 5 41



**Stata: summary of regression analysis code**

```
. regress PEFR b1.sex c.height170
*** output omitted ***
```

	PEFR	Coef.	Std. Err.	t	P> t	[95%]
sex						
male		50.0129	13.33098	3.75	0.000	23.1
height170		3.543314	.7935878	4.46	0.000	1.9
_cons		488.4078	8.087545	60.39	0.000	472

Difference in intercept for men and women

Slope for men and women

Intersection for women

01-10-2015 Basic Biostatistics - Day 5 43

**PEFR and Gender - formulations**

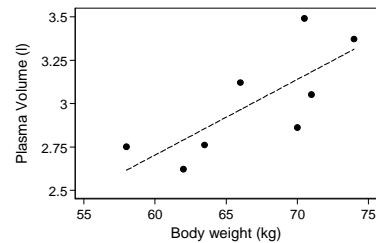
**Methods**  
The gender difference in PEFR was estimated as the difference in mean PEFR after linear adjustment for height. The model was checked by diagnostic plots of the residuals. Estimates... CI....

**Results**  
After adjustment for height men had a mean PEFR that was **50(24;77)l/min** higher than women.

**Conclusion**  
The sex difference in PEFR cannot solely be attributed to the difference in heights.

01-10-2015 Basic Biostatistics - Day 5 44

### Example 10.1 Body weight and plasma volume



Source	SS	df	MS			
Model	.390684335	1	.390684335	Number of obs =	8	
Residual	.287265681	6	.047877614	F( 1, 6) =	8.16	
Total	.677950016	7	.096850002	Prob > F =	0.0289	
				R-squared =	0.5763	
				Adj R-squared =	0.5057	
				Root MSE =	.21881	
plasma	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bweight	<b>.0436153</b>	.0152684	2.86	0.029	<b>.006255</b>	<b>.0809757</b>
_cons	<b>.0857244</b>	1.023998	0.08	0.936	<b>-2.419909</b>	<b>2.591358</b>

01-10-2015

Basic Biostatistics - Day 5

45

### The (Pearson) correlation coefficient

The (Pearson) correlation coefficient,  $\rho$ , is a measure of the strength of the **linear relationship** between two variables  $x$  and  $y$  following a **bivariate normal** distribution.

It only make sense if both  $x$  and  $y$  have a normal distribution and there is a linear relationship between  $x$  and  $y$ .

The correlations coefficient has the **following properties**:

- It is symmetric in  $x$  and  $y$ , and a change of scale of  $x$  and/or  $y$  will not change  $\rho$ .
- $\rho = \pm 1$  if the observation line exactly on a straight line.
- $-1 \leq \rho \leq 1$
- **If**  $x$  and  $y$  are independent, **then**  $\rho = 0$

01-10-2015

Basic Biostatistics - Day 5

46

### The (Pearson) correlation coefficient

The correlation is best understood as the coefficient of determination.

$\rho^2$  = how much of the variation in one of the variables that can be explained by the variation of the other.

So if  $\rho = 0.8$  then  $\rho^2 = 0.64 = 64\%$ , i.e. 64% of the variation in  $y$  can be explained by the variation in  $x$  and vice versa.

$\rho$  is **estimated** by the empirical correlation coefficient  $r$ :

$$\hat{\rho} = r = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

01-10-2015

Basic Biostatistics - Day 5

47

### The (Pearson) correlation coefficient

It is possible to make **approximate confidence** intervals for the Pearson correlation (see p95-96 in Kirkwood & Sterne).

Very few programs (not Stata!) will do this for you!

It is possible to make an exact test of the hypothesis:  $\rho = 0$

The test is **identical** to the test of **zero slope** in the simple linear regression.

All programs can make this test.

01-10-2015

Basic Biostatistics - Day 5

48

### Spearman's rank correlation

Subject	Body weight		Plasma volume	
	Obs	Rank	Obs	Rank
1	58.0	1	2.75	2
2	70.0	5	2.86	4
3	74.0	8	3.37	7
4	63.5	3	2.76	3
5	62.0	2	2.62	1
6	70.5	6	3.49	8
7	71.0	7	3.05	5
8	66.0	4	3.12	6

The body weight and the plasma volume are ranked separately.

**Spearman's rank** correlation is found as the correlation of the ranks!

It has the same properties as the correlation, but it has no interpretation.

The **test** of no association based on Spearman rank correlations is in general **valid**.

01-10-2015

Basic Biostatistics - Day 5

49

### Correlations some comments

The Pearson correlation is only a valid measure of association if:

1. We have independent observations, i.e. the pairs  $(x, y)$  are **independent**.
2. Both the  $x$ 's and the  $y$ 's have a **normal distribution**.
3. There is a **linear relationship** between  $x$  and  $y$ .

Note, these assumptions are **stronger** than the ones behind the simple linear regression.

The **test** of no association based on **Spearman** rank correlation is valid if 1. and

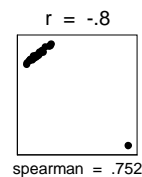
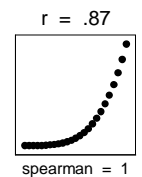
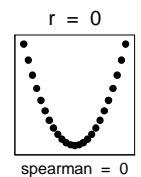
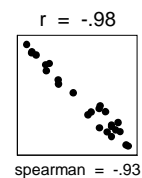
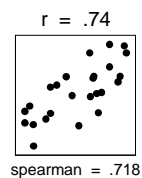
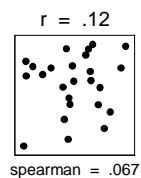
- 3b. There is a **monotone relationship** between  $x$  and  $y$ .

01-10-2015

Basic Biostatistics - Day 5

50

### Example of Pearson and Spearman correlations



**Remember: Always plot the data !!!!**

01-10-2015

Basic Biostatistics - Day 5

51

### Body weight and plasma volume

The **(Pearson)** correlation: **0.76(0.12;0.95)**

The **(Pearson)** correlation squared: **0.58(0.014;0.91)**

The hypothesis:  $\rho = 0$  gives  $p = 0.029$

The Spearman rank correlation is **0.81**

The test of no association based on this gives  $p = 0.015$

01-10-2015

Basic Biostatistics - Day 5

52

### Comparison of the measurement methods

A correlation coefficient is often seen in the literature as a way to compare two measurements.

**A correlation coefficient cannot be used to measure the agreement of two methods.**

We will illustrate this on the next overheads by showing that the correlation

- Does not measure a systematic difference.
- Does not measure a random difference.

01-10-2015

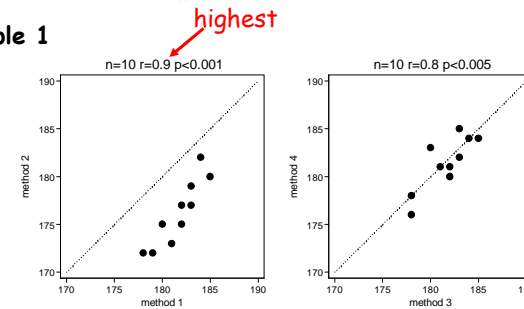
Basic Biostatistics - Day 5

53

### Comparison of the measurement methods

Two studies, each comparing two methods of measuring height on men. In both studies 10 men were measured twice, once with each method.

#### Example 1



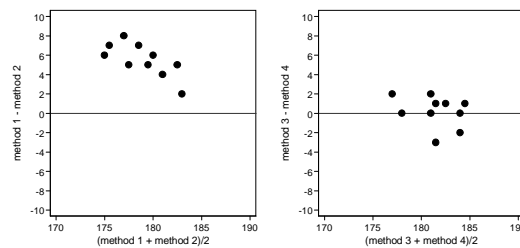
Is a higher correlation evidence of higher agreement?

01-10-2015

Basic Biostatistics - Day 5

54

Is a higher correlation evidence of higher agreement? **NO!!!**



Average difference:

5.6cm

0.2cm

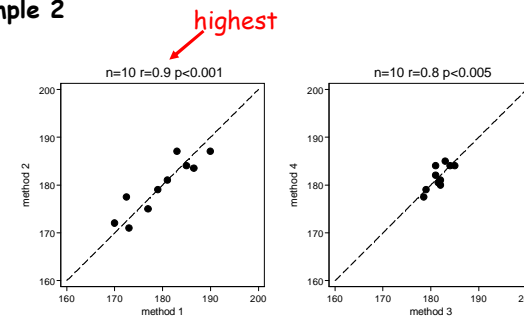
The correlation does not give you any information on whether the observations are located around  $y=x$  !!!!!

01-10-2015

Basic Biostatistics - Day 5

55

#### Example 2



Note, both data sets are located around  $y=x$  !

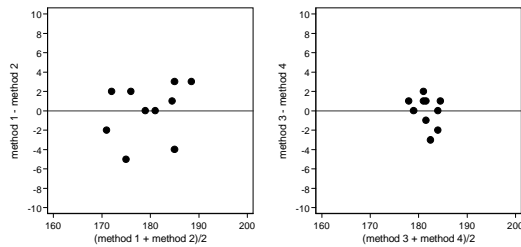
Is a higher correlation evidence of higher agreement?

01-10-2015

Basic Biostatistics - Day 5

56

Is a higher correlation evidence of higher agreement? **NO!!!**



SD of the  
difference: 2.8cm

1.6cm

The random differences are a bit smaller in the right plot!