

Bus 274: Further Statistics For Business

Harald Schmidbauer



About These Slides

- The present slides are not self-contained; they need to be explained and discussed. This will be done in the lectures.
- Even though being a “work in progress” and subject to revision, the slides constitute copyrighted material.
If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer **harald** at **hs-stat** dot **com**
Angi Rösch **angi** at **angi-stat** dot **com**

- The slides were produced using \LaTeX (www.latex-project.org) and R (www.R-project.org) on a GNU/Linux system, all of which are free and open source software (FOSS).
- R files used for this course are available upon request.



Chapter 15:

Simple Linear Regression



15.1 Simple Linear Regression: Goals

Goals of Simple Linear Regression.

Once again, given are points (x_i, y_i) , from a bivariate metric variable (X, Y) .

How can we establish a *functional relationship* between X and Y ? Most importantly:

- Which straight line is “good”? —
What does “good” mean?
- How can the parameters of a “good” line be computed?



15.1 Simple Linear Regression: Goals

Goals of Simple Linear Regression.

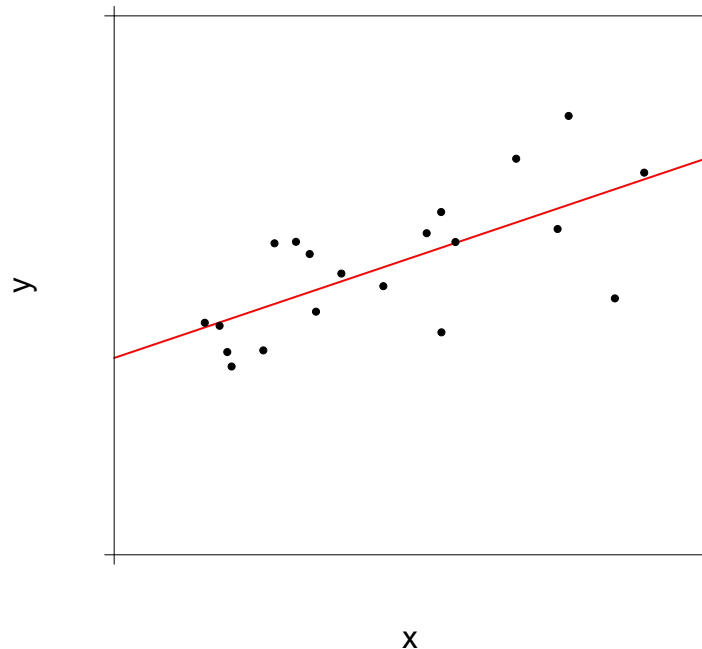
Why would we want to fit a line to a cloud of points?

- In order to quantify the relationship between X and Y , using a simple model.
- In order to forecast Y for a given value of X .



15.2 The Regression Line

Finding a “good” line. . .



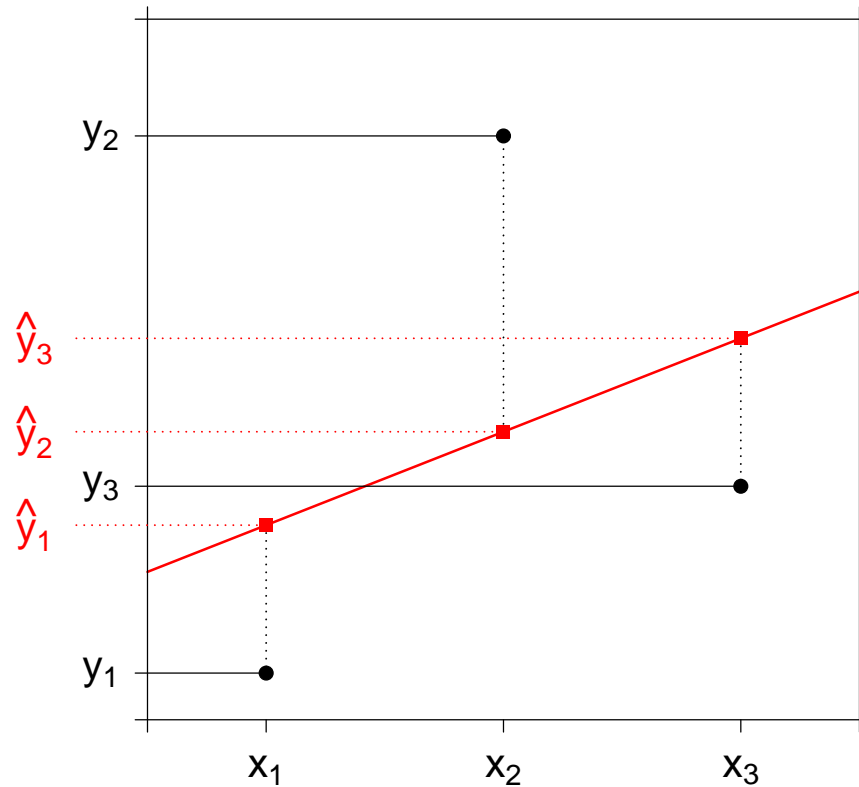
. . . and how can we find a “good” line?

— A *criterion* is needed!



15.2 The Regression Line

A very simple scatterplot.



- observed points:

$$(x_i, y_i)$$

- points on the line:

$$(x_i, \hat{y}_i)$$



15.2 The Regression Line

The method of least squares.

- Define $\hat{y}_i = a + bx_i$ and $e_i = y_i - \hat{y}_i$.
- a, b such that the sum of squared distances is minimized:

$$Q(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

- The line $y = a + bx$ with these parameters a and b is called the regression line of Y with respect to X .
- b : regression coefficient



15.2 The Regression Line

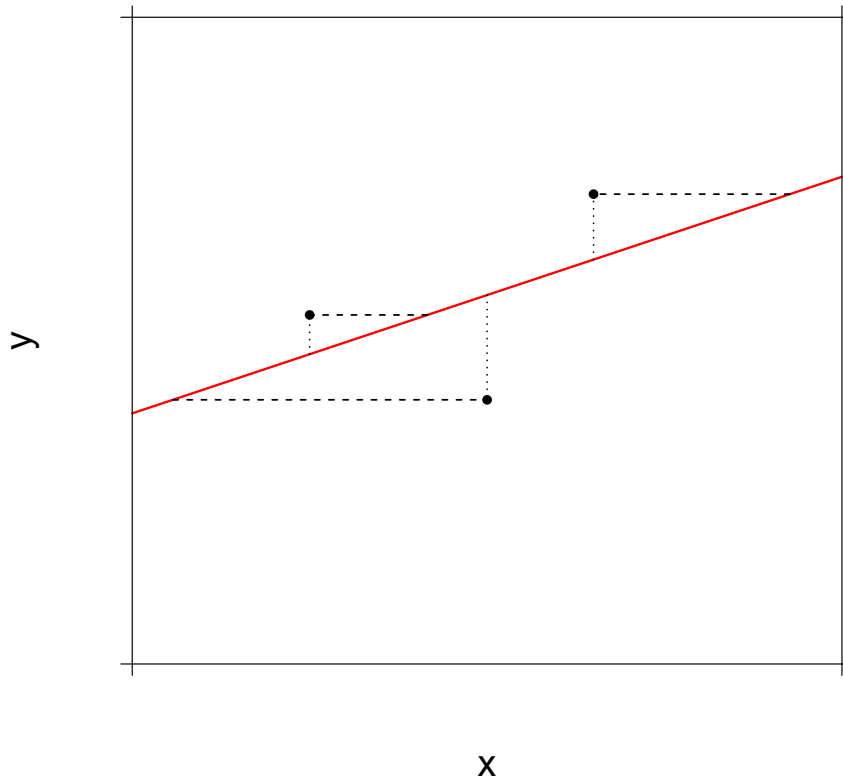
Regression: some first comments.

- “Good” means: The sum of squared distances, *parallel to the y-axis*, is minimized.
- This procedure is asymmetric!
- It conforms to the idea:
 Given X , what is Y ?
- X : “independent variable”, Y : “dependent variable”



15.2 The Regression Line

Regression is asymmetric.



The regression lines. . .

- . . . of Y w.r.t. X and
 - . . . of X w.r.t. Y
- are usually different.



15.2 The Regression Line

Y w.r.t. X , or rather X w.r.t. Y ?

Example:

X = body-height of a person;

Y = body-weight of a person

- Regression of Y w.r.t. X looks quite natural.
- Regression of X w.r.t. Y would be strange.



15.2 The Regression Line

Y w.r.t. X , or rather X w.r.t. Y ?

Example: Change in percent of price indices, on the corresponding month of the previous year:

X = change of housing price index;

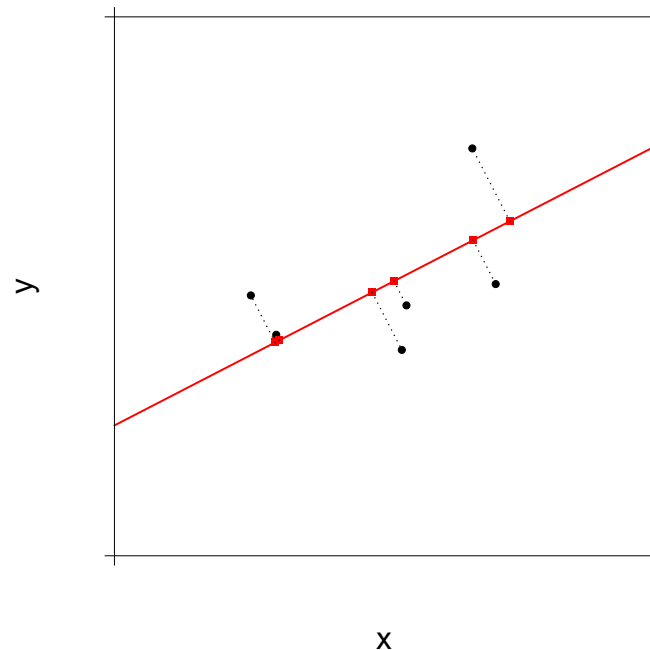
Y = change of clothing price index

- Neither regression — Y w.r.t. X nor X w.r.t. Y — looks very convincing.
- A *symmetric* procedure is more appropriate than regression here.



15.2 The Regression Line

A symmetric procedure: principal components analysis.



This illustrates how a “good” line can be found in such a case.



15.2 The Regression Line

Computing the regression line.

Minimizing Q leads to the following equations for the slope b and the intercept a :

$$\begin{aligned} b &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} \\ &= \frac{\text{cov}(X, Y)}{\text{var}(X)}, \end{aligned}$$

$$a = \bar{y} - b\bar{x}.$$



15.2 The Regression Line

Example: (This is a toy example. . .)

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	\hat{y}_i	e_i
1	5	15	25	225	75	13.9	1.1
2	10	8	100	64	80	11.3	-3.3
3	15	12	225	144	180	8.7	3.3
4	20	5	400	25	100	6.1	-1.1
Σ	50	40	750	458	435	40	0

Then,

$$b = \frac{4 \cdot 435 - 50 \cdot 40}{4 \cdot 750 - 50^2} = -0.52, \quad a = \frac{40}{4} - (-0.52) \cdot \frac{50}{4} = 16.5$$

The regression line is: $y = 16.5 - 0.52x$.

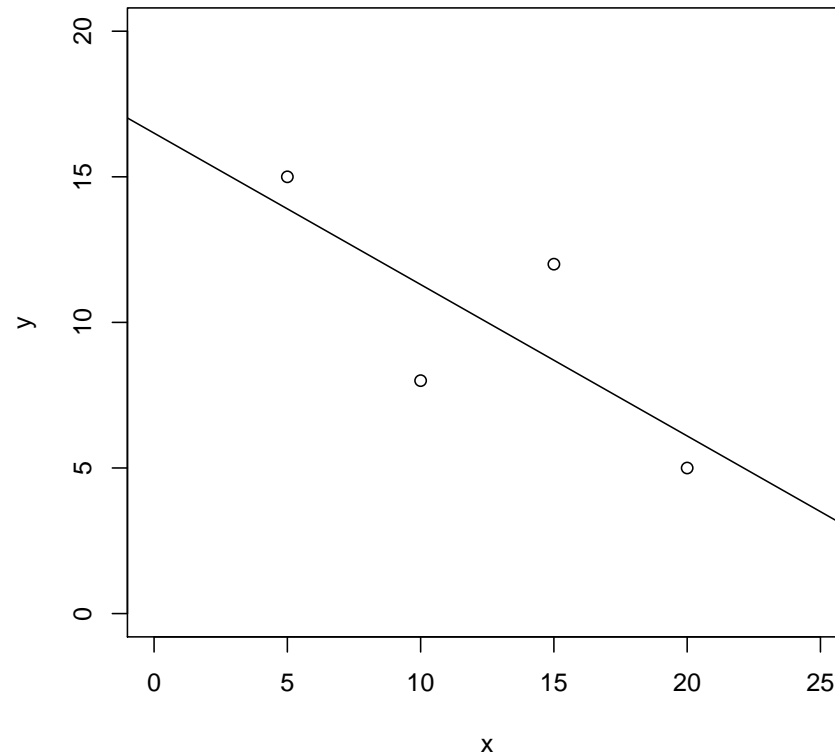
Using this regression line, the \hat{y}_i and the e_i can be computed.

We observe: $\bar{\hat{y}} = \bar{y}$, $\bar{e} = 0$. (This is always the case.)



15.2 The Regression Line

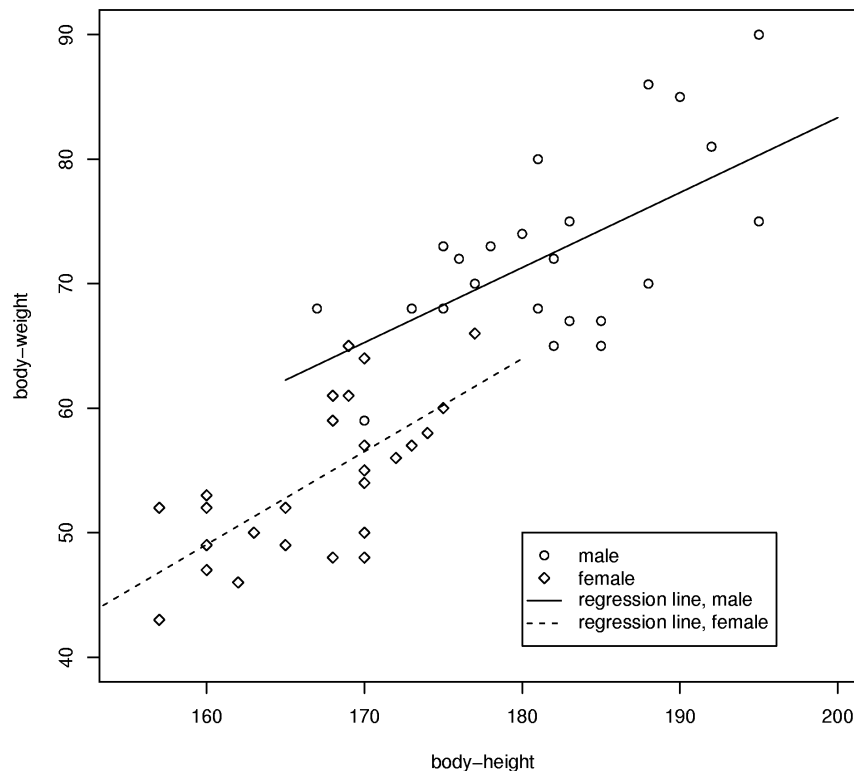
A plot of the toy example.



15.2 The Regression Line

Example: (This example is much better.)

X = body-height (in cm) of a person;
 Y = body-weight (in kg) of a person



- regression line, males:

$$y = 0.60x - 37.1$$

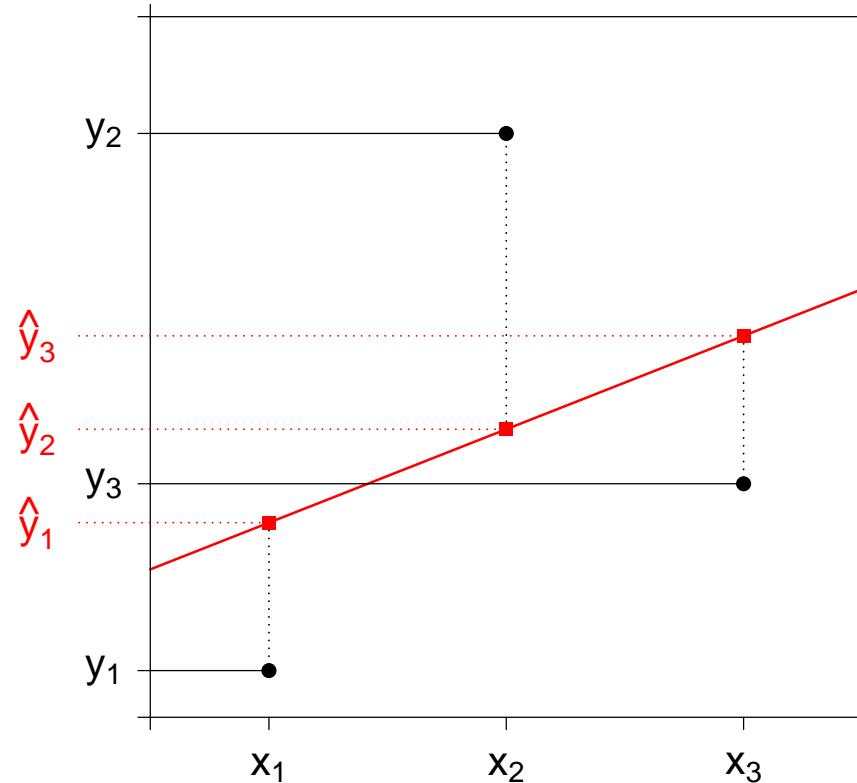
- regression line, females:

$$y = 0.75x - 70.4$$



15.3 Explanatory Power of the Model

Variability of the y_i and the \hat{y}_i .



15.3 Explanatory Power of the Model

The explanatory power of the regression model. . .

We observe:

- There is (in general) less variability in the \hat{y}_i than in the y_i ! — That is, the regression line cannot explain the *entire* variability in the observed y_i .
- The regression could provide a complete explanation if all points (x_i, y_i) were *on* the regression line.



15.3 Explanatory Power of the Model

Decomposition of variance.

- It holds that:

$$\begin{array}{rcl} \sum (y_i - \bar{y})^2 & = & \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \\ \text{SST} & = & \text{SSR} + \text{SSE} \end{array}$$

- Here,

SST: total sum of squares

SSR: regression sum of squares

SSE: error sum of squares



15.3 Explanatory Power of the Model

The coefficient of determination.

It is defined as:

$$\frac{SSR}{SST}$$

- The coefficient of determination is the share of variability in the data which is explained by the regression.
- It holds that $\frac{SSR}{SST} = r^2 = \text{cor}^2(X, Y)$.
- $r^2 = 100\%$ if and only if all observed points are on the regression line.
- $r^2 = 0\%$ if and only if X and Y are uncorrelated.

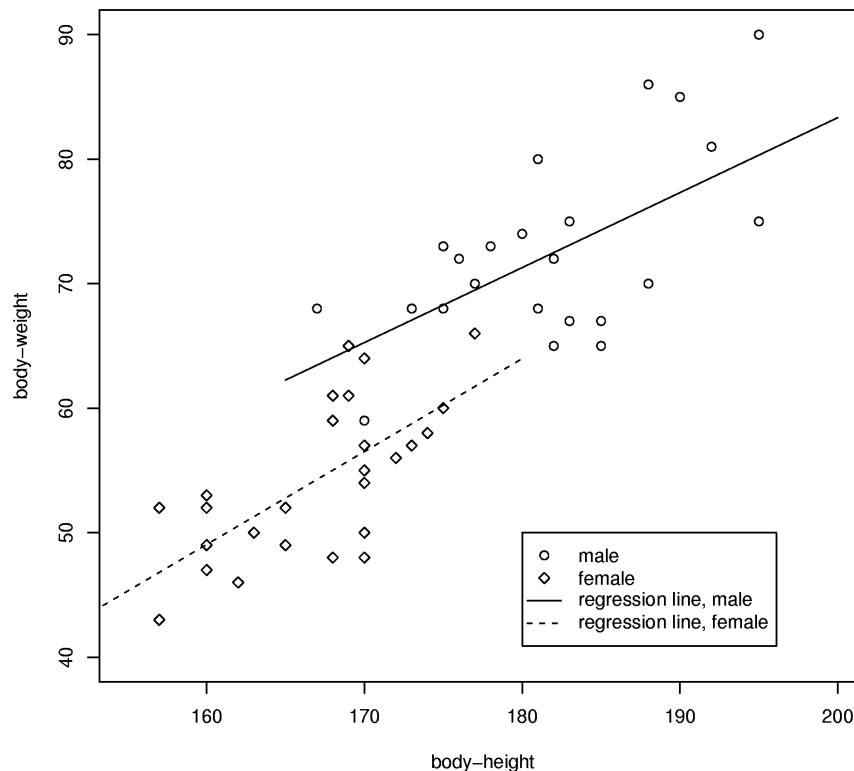


15.3 Explanatory Power of the Model

Example.

X = body-height (in cm) of a person;

Y = body-weight (in kg) of a person



● males:

$$r^2 = 37\%$$

● females:

$$r^2 = 48\%$$



15.3 Explanatory Power of the Model

Example. An interpretation of our results:

Why are not all women (or men) equally heavy?

- It is not by pure chance that women's (or men's) body-weights are different.
- It is because people are not equally tall!
- Is this the only reason? — No, but difference in body-height explains about 48% (37%) of the variability in body-weight.
- So what might be other reasons? This is not investigated here. . . .
(a guess: caloric intake, hours of exercise per week, . . .)

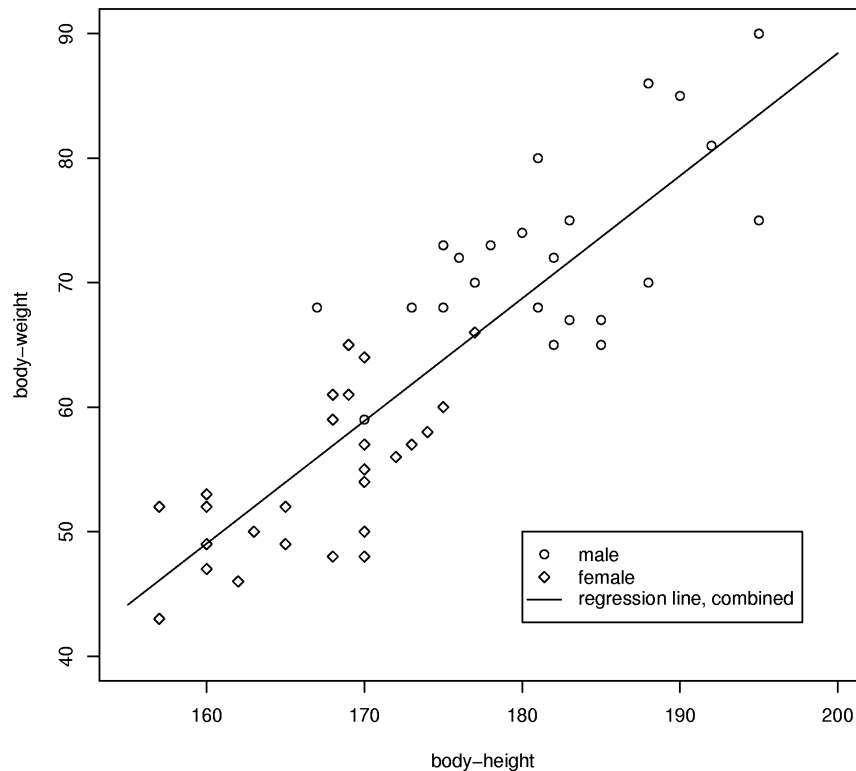


15.3 Explanatory Power of the Model

Example. (This example is doubtful!)

X = body-height (in cm) of a person;

Y = body-weight (in kg) of a person



- regression line:

$$y = 0.98x - 108.5$$

- coef. of determination:

$$r^2 = 76\%$$



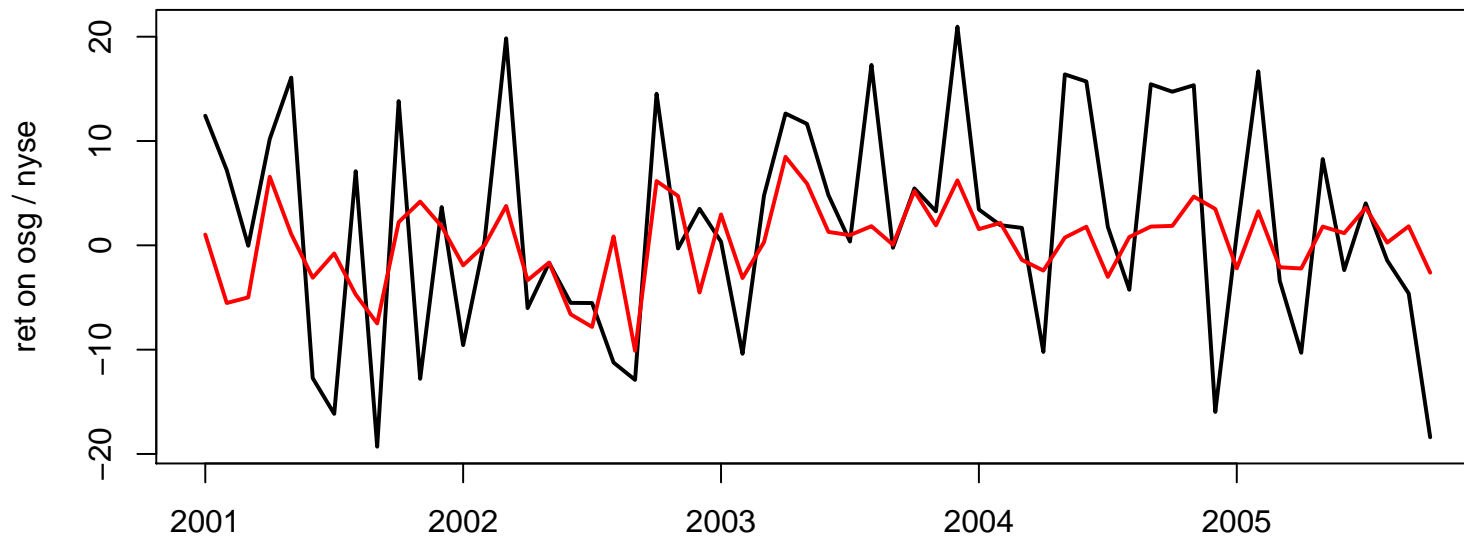
15.3 Explanatory Power of the Model

Example. Overseas Shipholding Group, Inc. (“OSG”), is a marine transportation company whose stock is listed at New York Stock Exchange (NYSE).

Let monthly returns in percent be defined as

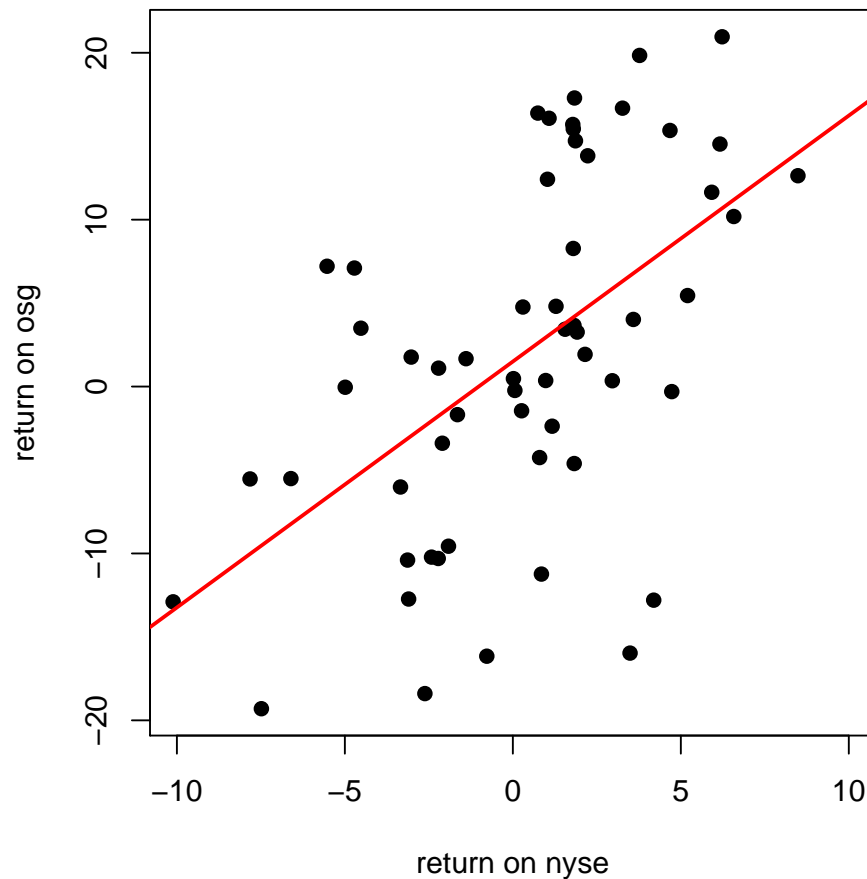
osg.ret = on OSG stock (black in the figure below);

nyse.ret = on the NYSE Composite Index (red)



15.3 Explanatory Power of the Model

Scatterplot and regression results.



- regression line:

$$\text{osg.ret} = 1.50 + 1.47 \cdot \text{nyse.ret}$$

- coef. of determination:

$$r^2 = 29\%$$



15.3 Explanatory Power of the Model

An interpretation of our results.

Why are there fluctuations in OSG stock price changes?

- It is not by pure chance that OSG stock price fluctuates.
- It is because the market index NYSE Composite fluctuates!
- Is this the only reason? — No, but fluctuations in NYSE Composite explain about 29% of the variability in OSG stock price.
- So what might be other reasons? This is not investigated here. . . .
(a guess: import/export quantities, decisions of the CEO, condition of competitors, . . .)



15.4 A Stochastic SLR Model

SLR in descriptive and inductive statistics.

- So far, we have seen SLR from a purely *descriptive* point of view.
(There were no probabilities, no stochastic models.)
- Advantage of this approach: simplicity
- Disadvantage: We obtain no insight into the mechanism which created the data —
for this purpose, we need a stochastic model and the methods of inductive statistics!



15.4 A Stochastic SLR Model

A *stochastic* simple linear regression model.

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

- The random variable Y_i represents the observation belonging to x_i .
- α and β are unknown parameters (to be estimated).
- x_i is the observation of the independent variable X .
- ϵ_i is a random variable; it contains everything not accounted for in the equation $y = \alpha + \beta x$.



15.4 A Stochastic SLR Model

Assumptions about ϵ .

We shall assume that the ϵ_i in

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

are a sequence of independent and identically distributed random variables:

$$\epsilon_i \sim N(0, \sigma_\epsilon^2) \quad \text{iid}$$

The “normality assumption” is very strong.



15.4 A Stochastic SLR Model

Assumptions about ϵ .

- The assumption of independence may well be fulfilled if observations are made from isolated objects.
- The assumption of independence may be violated if observations constitute a time series. (But what we say below also largely holds if the ϵ_i are correlated.)
- A great deal of statistical theory is concerned with relaxing these assumptions.



15.4 A Stochastic SLR Model

Computing estimators.

The method of least squares leads to the following estimators for β and α :

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2},$$

$$\hat{\alpha} = \bar{y} - b\bar{x}.$$

These are the same *formulas* as before — but what they *mean* is completely different!



15.4 A Stochastic SLR Model

The estimators $\hat{\alpha}$ and $\hat{\beta}$.

- $\hat{\alpha}$ and $\hat{\beta}$ are functions of the sample data (x_i, Y_i) .
- A function of sample data is called a *statistic*.
- Just as a random variable (representing an observation), a statistic has a probability distribution.
- These distributional properties can help us to learn about the unknown parameters α and β .



15.4 A Stochastic SLR Model

The estimators $\hat{\alpha}$ and $\hat{\beta}$.

We shall now:

- look at some distributional properties of the estimators $\hat{\alpha}$ and $\hat{\beta}$;
- find out under what circumstances β can be estimated reliably;
- look at examples, with a focus on understanding computer output.



15.4 A Stochastic SLR Model

The estimator $\hat{\beta}$. (Often more important than $\hat{\alpha}$.)

Statistical inference about β is based on the following property:

$$\frac{\hat{\beta} - \beta}{s_{\beta}} \sim t_{n-2},$$

where s_{β} is the standard error of $\hat{\beta}$:

$$s_{\beta}^2 = \frac{s_{\epsilon}^2}{\sum (x_i - \bar{x})^2} \quad \text{with} \quad s_{\epsilon}^2 = \frac{\text{SSE}}{n - 2}$$

(The latter estimates σ_{ϵ}^2 .)



15.4 A Stochastic SLR Model

The estimator $\hat{\beta}$.

In other words:

$$\hat{\beta} \sim \mathbf{N}(\beta, s_{\beta}^2) \quad \text{approximately.}$$

Under what circumstances can β be estimated reliably? — This is the case when

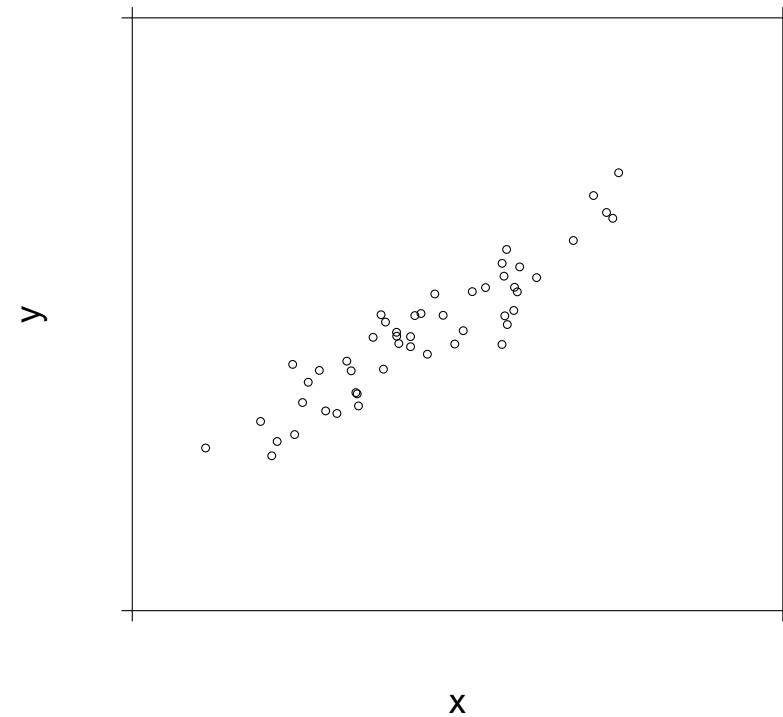
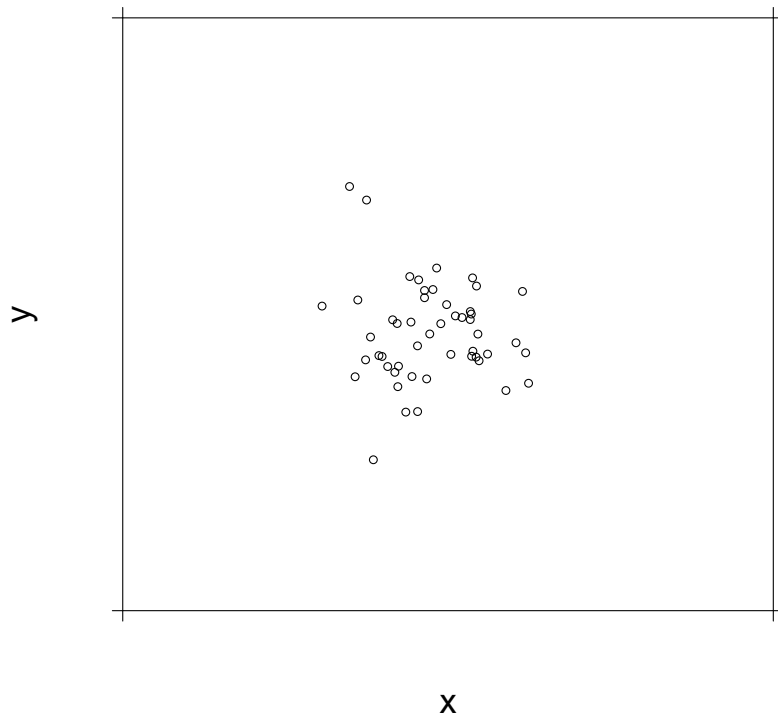
$$s_{\beta}^2 = \frac{s_{\epsilon}^2}{\sum (x_i - \bar{x})^2}$$

is small! Therefore, it is desirable that $\sum (x_i - \bar{x})^2$ should be large.



15.4 A Stochastic SLR Model

The estimator $\hat{\beta}$.



15.4 A Stochastic SLR Model

The estimator $\hat{\alpha}$. (Often less important than $\hat{\beta}$.)

Statistical inference about α is based on the following property:

$$\frac{\hat{\alpha} - \alpha}{s_{\alpha}} \sim t_{n-2},$$

where s_{α} is the standard error of $\hat{\alpha}$:

$$s_{\alpha}^2 = s_{\epsilon}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \quad \text{with} \quad s_{\epsilon}^2 = \frac{\text{SSE}}{n - 2}$$



15.4 A Stochastic SLR Model

Example: Body-height and body-weight again; here: males.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-37.1232	30.0065	-1.237	0.22851
height.m	0.6023	0.1647	3.657	0.00131 **

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

Residual standard error: 5.849 on 23 degrees of freedom

Multiple R-Squared: 0.3676, Adjusted R-squared: 0.3402

F-statistic: 13.37 on 1 and 23 DF, p-value: 0.001314

- The estimated regression model is:
$$\text{weight.m} = -37.1 + 0.60 \cdot \text{height.m} + \text{random error}$$
- Approximate 95% confidence bounds for β are given by $0.60 \pm 2 \cdot 0.165$; the corresponding 95% confidence interval is $[0.27, 0.93]$.
- The slope β is significantly different from 0.
(The null hypothesis $H_0 : \beta = 0$ is rejected against $H_1 : \beta \neq 0$.)



15.4 A Stochastic SLR Model

Example (continued):

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-37.1232	30.0065	-1.237	0.22851
height.m	0.6023	0.1647	3.657	0.00131 **

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

Residual standard error: 5.849 on 23 degrees of freedom

Multiple R-Squared: 0.3676, Adjusted R-squared: 0.3402

F-statistic: 13.37 on 1 and 23 DF, p-value: 0.001314

- The intercept α is not significantly different from 0.
- The error variance is estimated as $s_{\epsilon}^2 = (5.85)^2 = 34$.



15.4 A Stochastic SLR Model

Example: Body-height and body-weight again; here: females.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-70.3885	24.8816	-2.829	0.0087	**
height.f	0.7465	0.1494	4.996	3.08e-05	***

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

Residual standard error: 4.425 on 27 degrees of freedom

Multiple R-Squared: 0.4804, Adjusted R-squared: 0.4612

F-statistic: 24.96 on 1 and 27 DF, p-value: 3.075e-05

Comments. . . ??



15.4 A Stochastic SLR Model

Example: Overseas Shipholding Group, Inc. (“OSG”), and the NYSE Composite Index.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.4989	1.1801	1.270	0.209
nyse.ret	1.4737	0.3067	4.805	1.2e-05 ***

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

Residual standard error: 8.962 on 56 degrees of freedom

Multiple R-Squared: 0.2919, Adjusted R-squared: 0.2793

F-statistic: 23.09 on 1 and 56 DF, p-value: 1.200e-05

- The estimated regression model is:
$$\text{osg.ret} = 1.50 + 1.47 \cdot \text{nyse.ret} + \text{random error}$$
- Approximate 95% confidence bounds for β are given by $1.47 \pm 2 \cdot 0.31$; the corresponding 95% confidence interval is $[0.86, 2.08]$.
- The slope β is significantly different from 0.
(The null hypothesis $H_0 : \beta = 0$ is rejected against $H_1 : \beta \neq 0$.)



15.5 Prediction Based on SLR

Point prediction vs. interval prediction.

Let x be given. The outcome of the random variable $Y = \alpha + \beta x + \epsilon$ can be predicted in terms of. . .

- a single point: $\hat{Y} = \hat{\alpha} + \hat{\beta}x$
 - This has disadvantages similar to those of a point estimate.
- a prediction interval.
It has to cope with two sources of uncertainty:
 - The parameters α, β are unknown.
 - There is a random error ϵ , which has an unknown variance σ_ϵ^2 .



15.5 Prediction Based on SLR

Prediction intervals.

Given x_{n+1} (an out-of-sample value), a 95% prediction interval for the corresponding Y_{n+1} has bounds

$$\hat{Y}_{n+1} \pm t_{n-2, 0.975} \cdot s_{\epsilon} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

These are the bounds of an interval which will contain the random variable $Y_{n+1} = \alpha + \beta x_{n+1} + \epsilon$ with probability 95%.

Here, \hat{Y}_{n+1} is a point prediction, obtained as $\hat{Y}_{n+1} = \hat{\alpha} + \hat{\beta} x_{n+1}$.



15.5 Prediction Based on SLR

Example: Body-height and body-weight again; here: males.

Our model estimation was based on a sample of size $n = 25$. Now let the body height of a 26th person be given as $x_{26} = 180$ cm.

A point prediction of this person's body-weight is:

$$\hat{Y}_{26} = -37.1 + 0.60 \cdot 180 = 70.9$$

(Don't forget this was a sample of *young students*.)

An approximate 95% prediction interval has bounds

$$70.9 \pm 2 \cdot 5.85 \cdot \sqrt{1 + \frac{1}{25} + \frac{(180 - 182.04)^2}{1260.96}}$$

The corresponding prediction interval is: [58.9,82.9].



15.5 Prediction Based on SLR

The length of prediction intervals.

Prediction intervals become longer as x_{n+1} , for which Y_{n+1} is to be forecast, moves away from \bar{x} . This is illustrated in the following figures.

