University of St.Gallen

EFMD EQUIS ACCREDITED

Methods: Statistics (4,120)

11. Linear Regression

Spring 2022

Prof. Dr. Roland Füss

Swiss Institute of Banking and Finance (s/bf)

Contents

Learning Objectives

After this lecture, you know how:

- a **linear regression analysis** is performed and which **conditions** must apply for a valid estimation.
- **correlations in a data set** can be investigated and interpreted by means of a regression analysis.

Literature

Hill, R.C., W.E. Griffiths, and G.C. Lim (2011). *Principles of Econometrics*, 4th ed. United States: Wiley, **Chapter 2.****

Levine, D.M., K. A. Szabat, and D.F. Stephan. (2016). *Business Statistics: A First Course*, 7th ed. United States: Pearson, **Chapter 12**.*

Stinerock, R. (2018). *Statistics with R*. United Kingdom: Sage. **Chapter 12**.

Shira, Joseph (2012). *Statistische Methoden der VWL und BWL*, 4th ed. Munich et al.: Pearson Studium, **Chapter 17**.

Weiers, R. M. (2011). *Introductory Business Statistics*,7th ed., Canada: Thomson South-Western, **Chapter 15**.

*Mandatory literature **Very recommendable

I. Basic Idea

Regression analysis examines whether variables are related to each other and in which direction this **relationship** goes. It extends the analysis of variance and is often used for predictions in forecasts or choice models.

One examines whether the change of one variable (x) leads to a change of another variable (y) (**causal relationship**):

- x_k = Independent variable, predictor variable ($k = 1,2,...,K$)
- y = Dependent variable, target variable

Regression analysis is divided according to the number of involved predictors:

- **Simple Regression** with $K = 1$
- **Multiple Regression** with $K > 1$

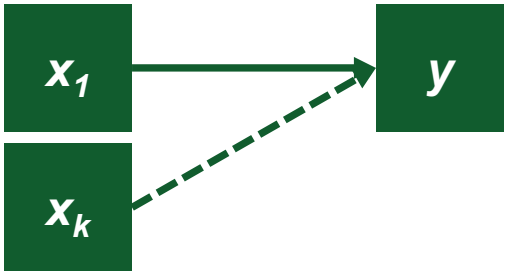
II. Correlation vs. Regression

Correlation analysis:

The **relationship** between the metric variable *x* and the metric variable *y is* investigated (non-directional relationship).

Regression analysis:

The **influence** of the metric or dummy-coded variable x_1 and possible further variables x_k on the metric variable *y is* investigated (directional relationship).



II. Correlation vs. Regression

In contrast to correlation analysis, regression analysis allows the examination of **dependencies** between two (or more) metrically scaled variables. (exception: dummy-coded variables).

There is an assumed **cause-effect relationship** between *x* and *y,* where *x is* the cause and *y* is the effect. Variable *y* is therefore dependent on variable *x*. This relationship is described as follows:

$$y = f(x)$$

The reasoning for this cause-effect relationship is not derived from regression analysis but from **theory** (e.g., from an economic, behavioral theory).

III. Basic Model (Introductory Example)

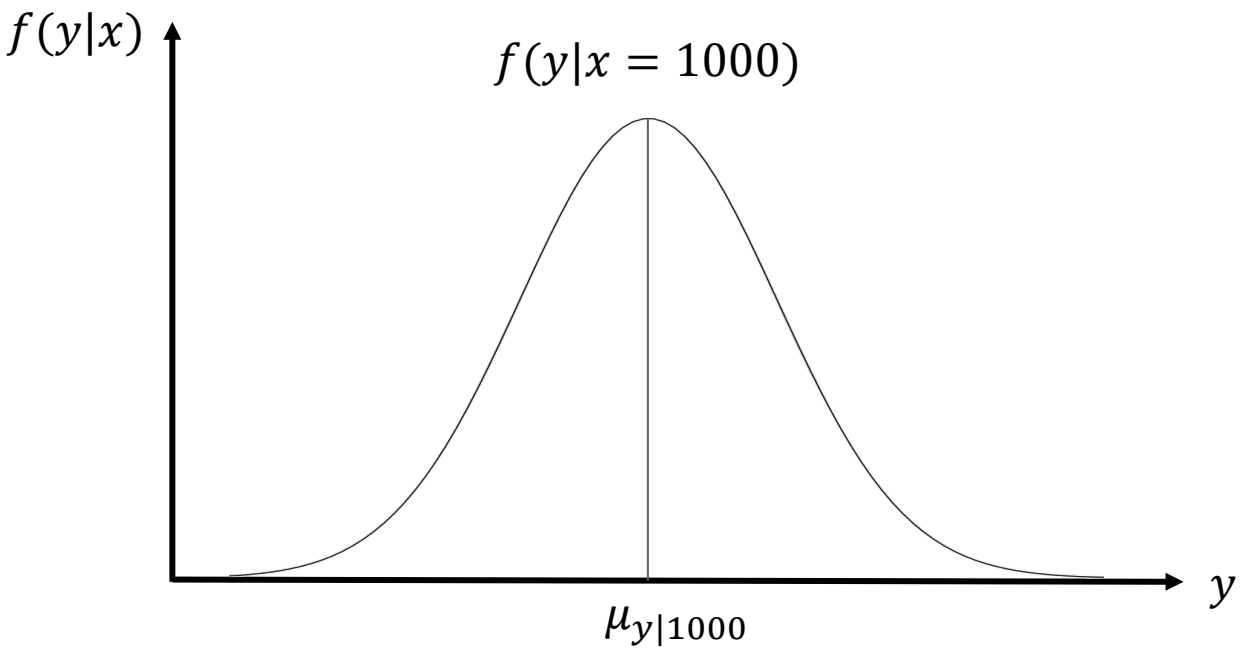
Income and food expenses:

An economic theory describes the relationship between the weekly income and food expenditure of students. With the help of a regression analysis, the strength and form of this relationship could be determined. **The following questions, among others, should be clarified**:

- Is there a significant relationship between income and food expenditure?
- Can the variance in weekly income explain the variance in food expenditure?
- How do food expenses increase when the weekly income rises by 100 CHF?
- What food expenses should a student with a weekly income of 500 CHF have?

III. Basic Model (Introductory Example)

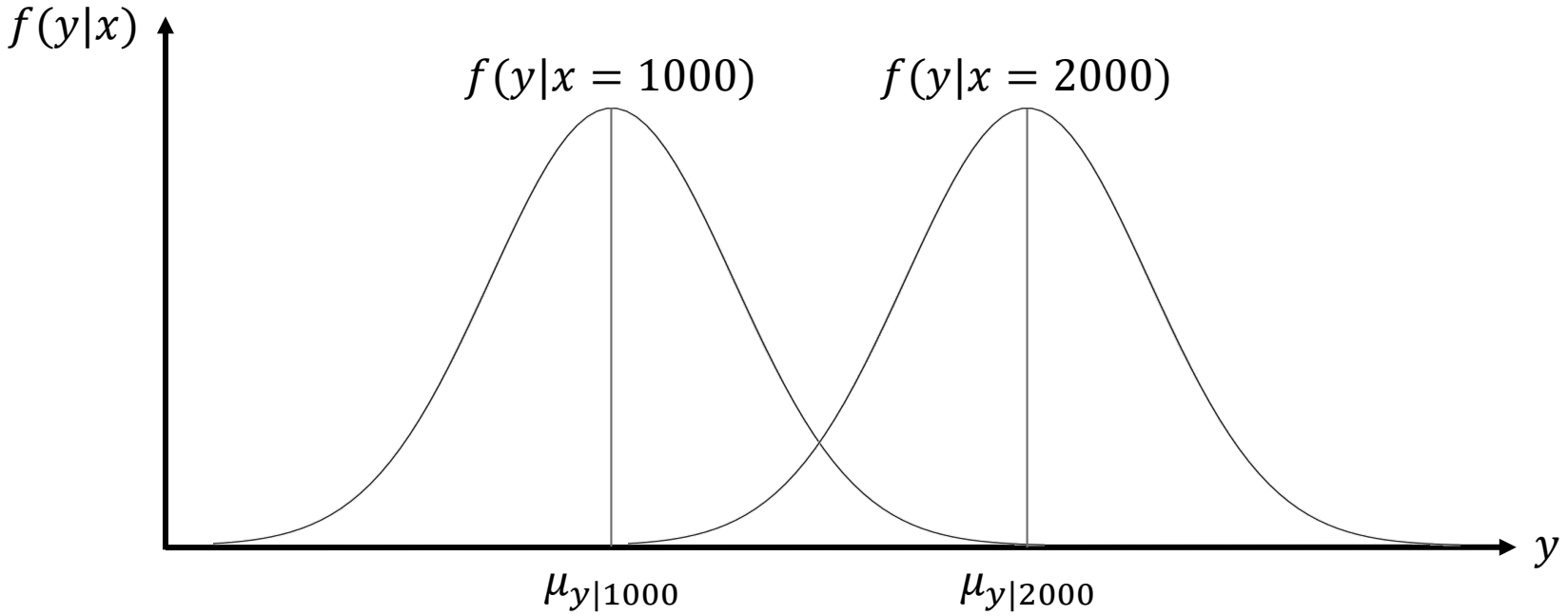
Consider the relationship between *x* = income and *y* = food expenditure per week. The chart below shows the probability distribution of food expenses for students with an income of *x* = 1000 CHF.



$$\mu_{y|x=1000} = E[y|x = 1000]$$

III. Basic Model (Introductory Example)

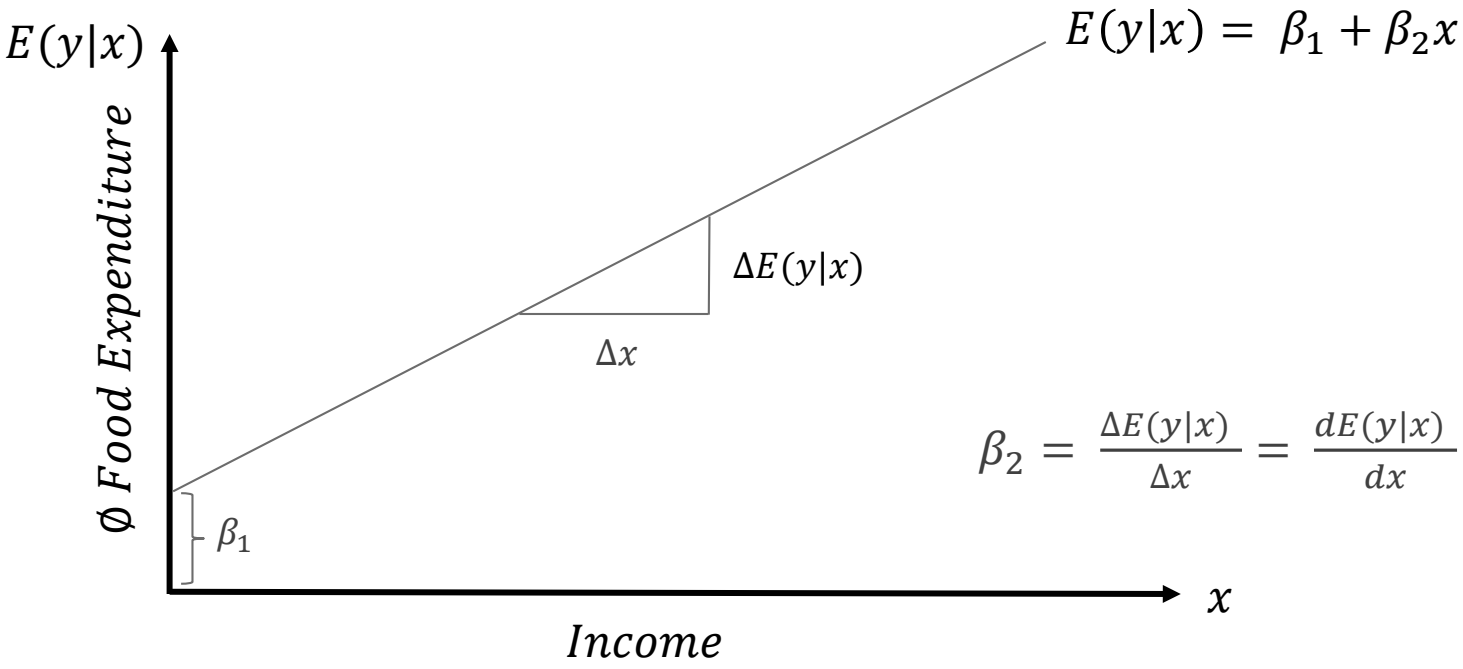
Our economic theory tells us that the average weekly food expenditure increases with income!



$\mu_{y|x=1000} = E[y|x = 1000]$ $\mu_{y|x=2000} = E[y|x = 2000]$

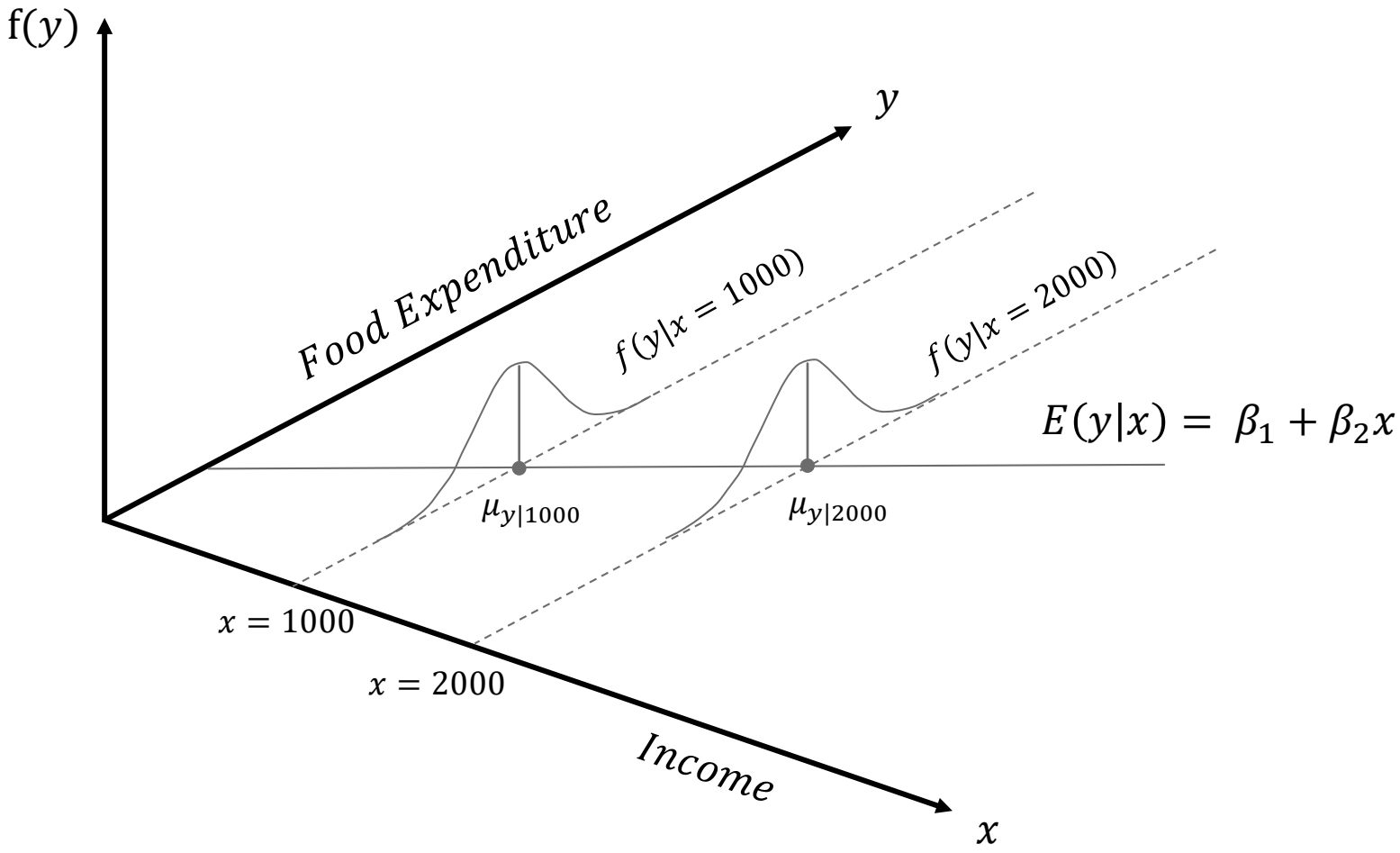
III. Basic Model (Introductory Example)

If there is a linear relationship between income and food expenditure per week, the true regression model can be drawn as follows:



III. Basic Model (Introductory Example)

The "complete picture" of the true model looks like this:



III. Basic Model

The random error in this regression model is given as :

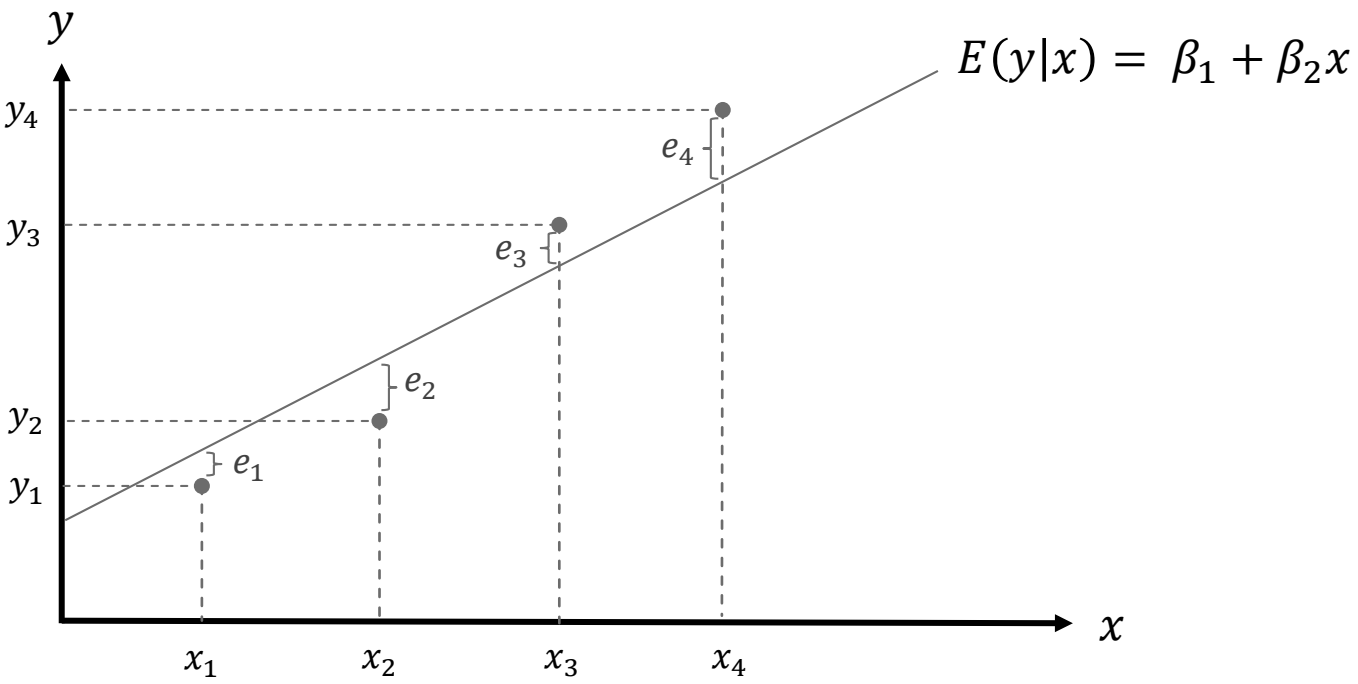
$$e_i = y_i - E[y_i|x_i] = y_i - \beta_1 - \beta_2 x_i$$

This results in the following assumptions and the simple linear regression model can be written as follows:

- **Assumption 1**: $y_i = \beta_1 + \beta_2 x_i + e_i$
- **Assumption 2**: $E[e_i|x_i] = 0$
- **Assumption 3**: $\mathrm{Var}(e_i|x_i) = \sigma^2$
- **Assumption 4**: $\mathrm{Cov}(e_i, e_j|x_i, x_j) = 0$ for $i \neq j$
- **Assumption 5**: $x_i, \ldots, x_n$ are random and not all the same.
- **Assumption 6**: $e_i, \ldots, e_n$ are normally distributed.

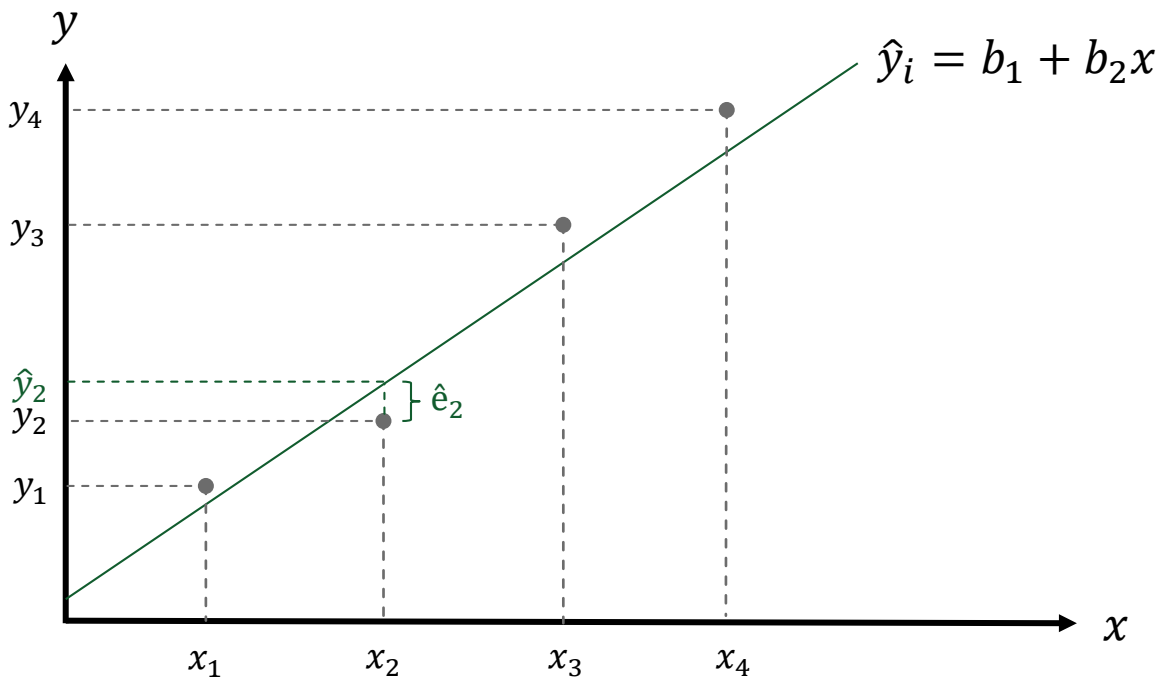
III. Basic Model

Hence, the relationship between y_i, e_i and the true regression line (population regression line) is as follows:



III. Method of Ordinary Least Squares (OLS)

The **Ordinary Least Squares** method tries to find the true regression line. The estimated line has the smallest distance to the individual observations.



III. Method of Ordinary Least Squares (OLS)

Mathematically, the OLS estimates b_1 and b_2 for the constant and the slope of the true regression model are based on the OLS residuals.

True regression model: $E(y|x) = \beta_1 + \beta_2 x$

Estimated regression model: $\hat{y}_i = b_1 + b_2 x$

OLS residuals: $\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$

The residual sum of squares is minimized. The sum of squares is used because the residuals can be positive or negative.

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

Solving the first-order conditions for a minimum yields:

$$b_1 = \bar{y} - b_2 \bar{x} \qquad b_2 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i{}^2 - (\sum_{i=1}^{n} x_i)^2}$$

III. OLS (Introductory Example, R-Example 1)

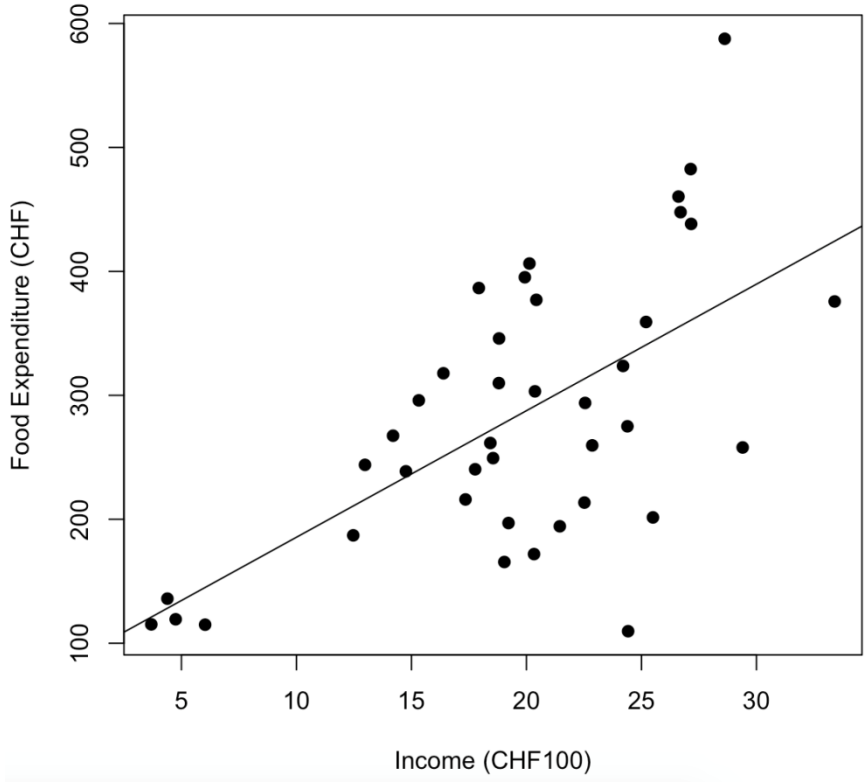
In order to investigate the theoretical relationship between the weekly income and the food expenditure of students, a sample is collected at the HSG. 40 students state their income and food expenses per week.

Student	Food expenditure (CHF)	Income (CHF100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
…	…	…
39	257.95	29.40
40	375.73	33.40

Descriptive statistics		
Mean	283.574	19.605
Median	264.480	20.030
Max.	587.660	33.400
Min.	109.710	3.690
SDV	112.675	6.848

III OLS (Introductory Example, R-Example 1)

Open the file "L11-Example_1.R" in R-Studio and reproduce the R-code and interpret the following regression line.



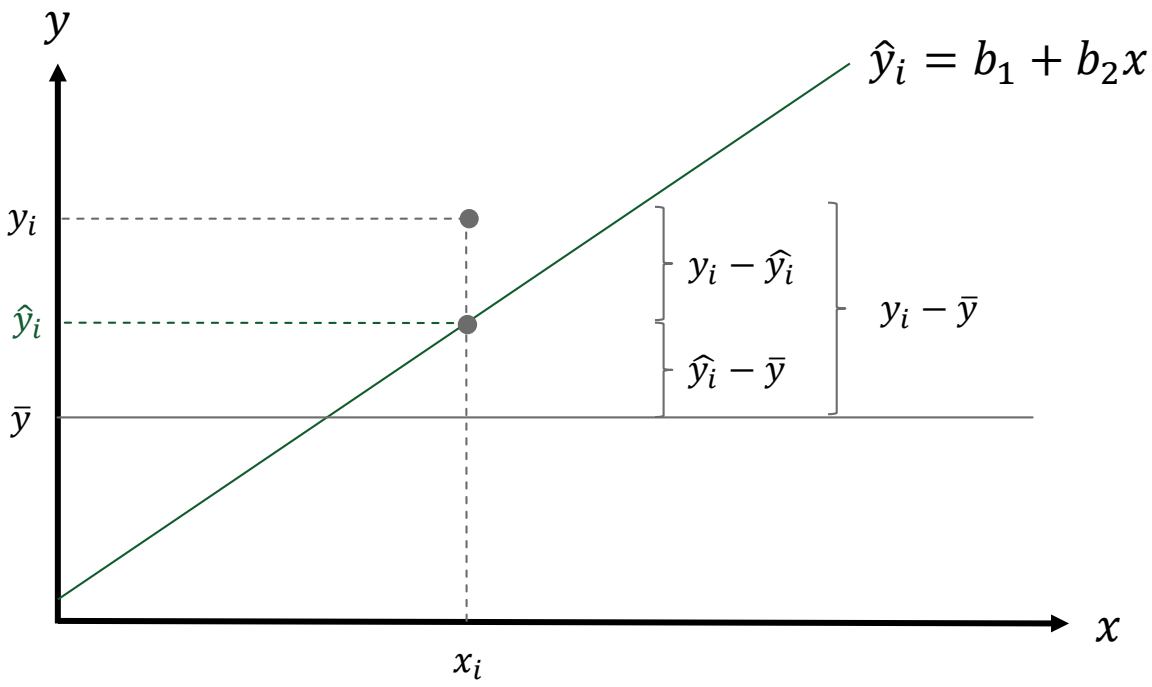
$$\hat{y}_i = 83.42 + 10.21x$$

```
Call:
lm(formula = food_exp ~ income, data = food)

Coefficients:
(Intercept)       income
      83.42        10.21
```

IV. Goodness of Fit

However, the question arises as to how well the OLS estimate resembles the true regression line. The variance decomposition gives an indication:



Total Sum of Squares (SST) = Regression Sum of Squares (SSR) + Residual Sum of Squares (SSE)

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

IV. Goodness of Fit

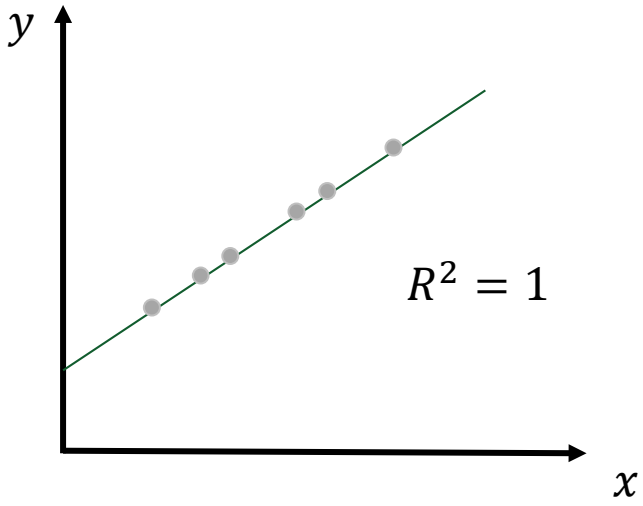
Based on the variance decomposition, the **coefficient of determination R^2** measures the proportion of the variation in the dependent variable that is explained by the regression model. R^2 only assumes values between 0 and 1.

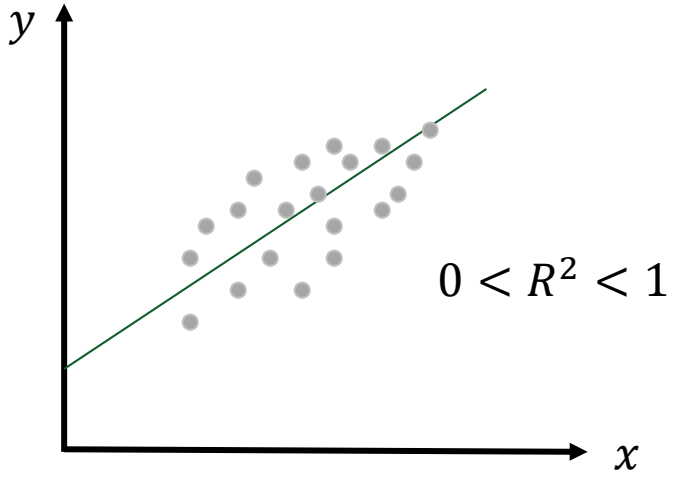
$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

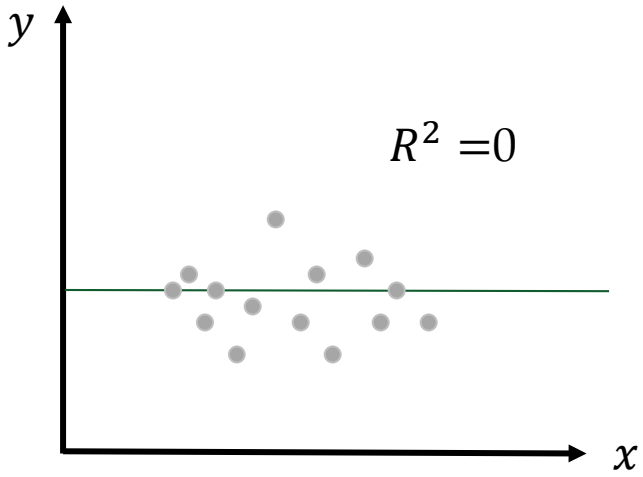
The coefficient of determination R^2 can be interpreted as the **share of dispersion of the *y*-values that is explained by the regression**. For example, the $R^2 = 0.37$ in the previous regression means that 37% of the dispersion in weekly food expenditures of students can be explained by the linear dependency on income.

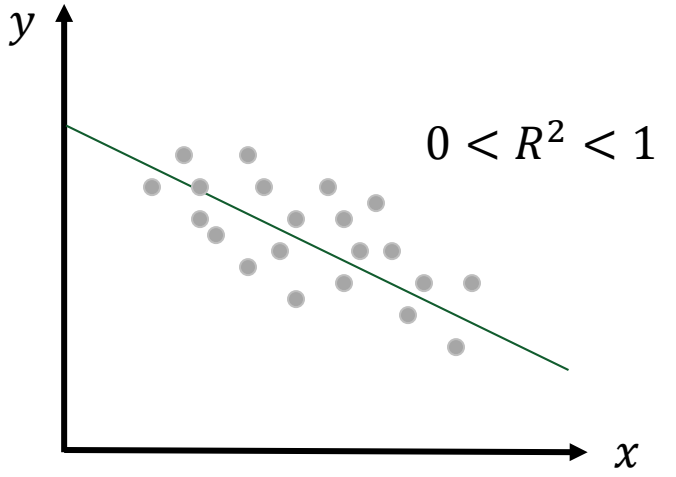
IV. Goodness of Fit

These four graphs illustrate possible values for the coefficient of determination R^2:









IV. Goodness of Fit

To check whether the regression model is significant overall, an *F*-test* can be used. This test checks whether the prediction of the dependent variable is improved by adding the independent variable (H_0: $\beta_2 = 0$, slope test). Hence, it clarifies whether the model as a whole has explanatory power.

Mean Square Regression: $$MSR = \frac{\text{SSR}}{1}$$

Mean Square Error: $$MSE = \frac{\text{SSE}}{n-2}$$

$$F\text{-}Ratio = \frac{MSR}{MSE}$$

The calculated *F*-Ratio is then compared with the critical value F_c, *which* follows an *F*-distribution with 1 and *n - 2* degrees of freedom. If the *F*-Ratio exceeds the critical value, the model is significant and the linear relationship can be confirmed.

* The results of a *t*-test and a *F*-test are identical for a simple regression.
The *F*-test in case of multiple regression will be covered in the next lecture.

V. Properties of OLS Estimators

The OLS estimators are random variables with means and variances. If **assumptions 1 to 5** hold, then the following applies:

$$E(b_1) = \beta_1 \qquad \text{Var}(b_1) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$E(b_2) = \beta_2 \qquad \text{Var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\text{and } \text{Cov}(b_1, b_2) = \sigma^2 \left[\frac{-\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

According to the **Gauss-Markov Theorem**, this means that the OLS estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2. They are **BLUE** (Best Linear Unbiased Estimators).

The Gauss-Markov Theorem does not depend on assumption 6. If assumption 6 does happen to hold, the OLS estimators are also normally distributed.

V. Variance of OLS estimators

In practice, the error variance is unknown, which means the variances of the OLS estimators are also unknown. An unbiased estimator of the error variance is given as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-2}$$

Therefore, the estimated variances of the OLS estimators are:

$$\widehat{\mathrm{Var}(b_1)} = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \widehat{\mathrm{Var}(b_2)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The square roots of the estimated variances are called the **standard errors.**

VI. Confidence Intervals and *t*-Tests for OLS Estimators

If assumptions 1 to 6 are met, confidence intervals for the OLS estimators can be calculated as follows:

$$t = \frac{b_k - \beta_k}{s_{b_k}} \sim t_{n-2} \text{ for k= } 1, 2$$

$$b_k \pm t_c s_{b_k}$$

The critical value t_c is determined based on the significance level α and can be found in the table of the *t*-distribution. It should be pointed out that we have $n - 2$ degrees of freedom, which is the sample size minus the total number of independent variables minus 1.

It is also possible to create **one-tailed and two-tailed *t*-tests** for the OLS estimators with the *t*-statistics stated above. The procedure of hypothesis testing remains the same as in the previous lectures.

VI. OLS Estimator (Introductory Example, R-Example 1)

Open the file "L11-Example_1.R" in R-Studio and reproduce the R-Code, which leads to the following results.

```
# example: food expenses
#--------------------------------------
# to investigate the theoretical relationship between weekly income and food expenditure
# of students, a random sample is taken at  HSG. 40 students state their income and food expenses per week.

# open the dataset "food.rda"
load("food.rda")

# quick inspection of the new dataset.
head(food, 3)

# create a scatter-plot to verify that the two variables are linearly related.
plot(food$income, food$food_exp, pch = 19, xlab="Income (CHF100)",ylab="Food Expenditure (CHF)")

# estimate the simple linear regression model and add the OLs estimate to the graph.
slr <- lm(food_exp ~ income, data = food)
abline(slr)

# detailed results of the regression model.
summary(slr)

# How will food expenses change if income increases by CHF100 per week?
# How can the estimate of the constant be interpreted?
# How are the t-values in the regression output interpreted? H0? H1?
# What does R2 say?
# How can the F-test in the output be interpreted?

# 95% confidence interval for the estimated coefficients
confint(slr,level=0.95)

# Use the results of the confidence interval to test if the slope is
# significantly (5% level) different from 15.
```

```
> confint(slr, level=0.95)
                2.5 %    97.5 %
(Intercept) -4.463279 171.29528
income       5.972052  14.44723
```

```
Call:
lm(formula = food_exp ~ income, data = food)

Residuals:
     Min       1Q   Median       3Q      Max
-223.025  -50.816   -6.324   67.879  212.044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.416     43.410   1.922   0.0622 .
income        10.210      2.093   4.877 1.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3688
F-statistic: 23.79 on 1 and 38 DF,  p-value: 1.946e-05
```