University of St.Gallen

EFMD EQUIS ACCREDITED

Methods: Statistics (4,120)

12. Multiple Regression

Spring 2021

Prof. Dr. Roland Füss

Swiss Institute of Banking and Finance (s/bf)

Contents

Learning Objectives

After this lecture, you know how:

- the results of a multiple regression analysis are interpreted and its **model quality** is determined.
- **dummy variables** can model interaction effects.
- **potential errors in** regression analysis can be detected and avoided.

Literature

Hill, R.C., W.E. Griffiths, and G.C. Lim (2011). *Principles of Econometrics*, 4th ed. United States: Wiley, **Chapter 5.****

Levine, D.M., K. A. Szabat, and D.F. Stephan. (2016). *Business Statistics: A First Course*, 7th ed. United States: Pearson, **Chapter 13**.*

Stinerock, R. (2018). *Statistics with R: A Beginner's Guide*. United Kingdom: Sage. **Chapter 13**.*

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 16**.

*Mandatory literature **Very recommendable

I. Basic Model (Introductory Example)

Fast food chain:

A fast-food company wants to investigate its hamburger sales empirically. The managing director reads about an economic theory that the price and the advertising expenditures have a major impact on the hamburger turnover. Among others, he asks himself the following questions:

- Do hamburger sales increase with higher advertising expenditure?
- If yes, do the higher sales compensate for increased advertising expenditures?
- Does a reduction of the price lead to an increase in sales or a loss of turnover? Are hamburger sales price-inelastic or price-elastic?

I. Basic Model

A relationship between a dependent variable and several independent variables, such as the introductory example, can be examined closely by using OLS in a **multiple regression analysis.** This results in the following assumptions and the multiple regression model can be written as:

- **Assumption 1**: $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{ik} + e_i$
- **Assumption 2**: $E[e_i|x_i] = 0$ where $x_i = (x_{i2}, x_{i3}, \ldots, x_{ik})'$
- **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 there is no exact linear relationship among any of the independent variables.
- **Assumption 6**: $e_i, \ldots, e_n$ are normally distributed.

I. Basic Model

Therefore, in a multiple regression model there are multiple slope coefficients. These coefficients are always interpreted individually. For example, β_k measures the effect of the change in the independent variable x_k on the expected value of *y* **if all other variables are held constant (ceteris paribus).**

True regression model: $E(y|x_{i2}, \dots, x_{ik}) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{ik}$

Estimated regression model: $\hat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + \cdots + b_K x_{ik}$

OLS residuals: $\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_{i2} - \cdots - b_K x_{ik}$

The OLS estimates of the coefficients are obtained by minimizing* the residual sum of squares as with simple regression.

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{i2} - \cdots - b_K x_{ik})^2$$

* Solving the first-order conditions for a minimum yields messy expressions for the least squares estimators, even when K is small. Morover, it requires complex linear algebra and matrix notation.

I. Basic Model (Introductory Example, R-Example 1)

The managing director now wants to empirically investigate the influence of price and advertising expenditure on hamburger sales. Therefore, the fast food chain offers burgers in 75 cities at different prices, varies advertising expenditure in each city and measures the sales figures. This cross-sectional sample looks as follows:

Town	Sales (CHF1000)	Price (CHF)	Advertising (CHF1000)
i	y	x_2	x_3
1	73.2	5.69	1.3
2	71.8	6.49	2.9
…	…	…	…
74	81.3	5.45	2.0
75	75.0	6.05	2.2
		Descriptive statistics	
Mean	77.37	5.69	1.84
Median	76.50	5.69	1.80
Max.	91.20	6.49	3.10
Min.	62.40	4.83	0.50
SDV	6.49	0.52	0.83

I. Basic Model (Introductory Example, R-Example 1)

Open the file "L12-Example_1.R" in R-Studio and reproduce the R-Code, which leads to the following results.

```
# example: hamburger sales
#-------------------------------------
# the managing director of a fast-food chain suspects a connection between hamburger sales,
# the price and the advertising expenditure. therefore, he varies the prices and advertising expenses in 75 branches.

# Open the dataset "andy.rda".
load("andy.rda")

# estimate the multiple regression model:
mlr <- lm(sales ~ price + advert, data = andy)
summary(mlr)
```

```
Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722  < 2e-16 ***
price        -7.9079     1.0960  -7.215 4.42e-10 ***
advert        1.8626     0.6832   2.726  0.00804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.886 on 72 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4329
F-statistic: 29.25 on 2 and 72 DF,  p-value: 5.041e-10
```

II. Goodness of Fit

The **coefficient of determination *R²*** either remains the same or improves when adding a further independent variable (even if it has no economic justification). Hence, in the case of multiple regression, R^2 does not allow for a fair comparison of different regression models.

$$R^2_{adj} = 1 - [(1 - R^2)\frac{n-1}{n-k-1}]$$

However, the ***Adjusted R²*** overcomes this weakness by including the sample size and the number of independent variables in the calculation. In contrast to the quality measure R^2, **the *Adjusted* R^2 can also decrease with the addition of variables with little or no explanatory content (penalization).** This allows a fair comparison of different models.

! The ***Adjusted* R^2** should **not** be used as a criterion for adding or deleting variables in the regression model!

II. Goodness of Fit

In the case of a multiple regression, an *F*-test *can also be used* to check whether the regression model is significant overall. It checks whether the prediction of the dependent variable is improved by adding an independent variable (slope test). In the case of multiple regressions, however, the calculation changes slightly:

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

Mean Square Error: $$MSE = \frac{\text{SSE}}{\text{n} - k - 1}$$

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

The calculated *F*-Ratio is then compared with the critical value F_c, which follows a *F*-distribution with *k* and *n - k -1* degrees of freedom. If the *F*-Ratio exceeds the critical value, the null hypothesis (H_0: $\beta_2 = \beta_3 = \cdots = \beta_K = 0$) can be rejected. Hence, the model is significant and the linear relationships can be confirmed.

III. Properties of OLS Estimators

In the multiple regression framework, OLS estimators are also random variables with means and variances. If **assumptions 1 to 5** hold, the estimators fulfill the **Gauss-Markov Theorem**. $b_1, b_2, \dots, b_k$ then have the smallest variance of all linear and unbiased estimators of $\beta_1, \beta_2, \dots, \beta_k$ and are **BLUE** (Best Linear Unbiased Estimators).*

As the error variance is also unknown in the case of multiple regressions, the true variance of the coefficients has to be estimated again. The estimated variance (**standard error**) of each coefficient has to be derived from error variance seperately. However, these calculations are messy and are not mentioned in detail.

* In the case of multiple regression, the Gauss-Markov Theorem is also not dependent on assumption 6.

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:

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

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

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 $n - \mathrm{k} - 1$ degrees of freedom, which is the sample size minus the total number of independent variables minus one.

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.

V. Dummy Variables (R-Example 2)

House prices:

The price of a house depends on the living space in square feet (*SQFT)*. However, a broker suspects that this relationship could be influenced by the presence of a pool. Therefore, he would like to investigate this relationship empirically and takes a sample of 1000 house sales in a city. In addition to price and living space, this sample collects a variable which measures whether the house has a pool or not.

This variable is called a **dummy variable** which reflects the respective categories (no pool, pool) with the values 0 and 1:

$POOL_i = 0$ if house i has no pool
$POOL_i = 1$ if house i has a pool

Dummy variables are often included in different ways in regression models and can, therefore, measure different effects.

V. Dummy Variables (R-Example 2)

Open the file "L12-Example_1.R" in R-Studio and reproduce the R-Code.

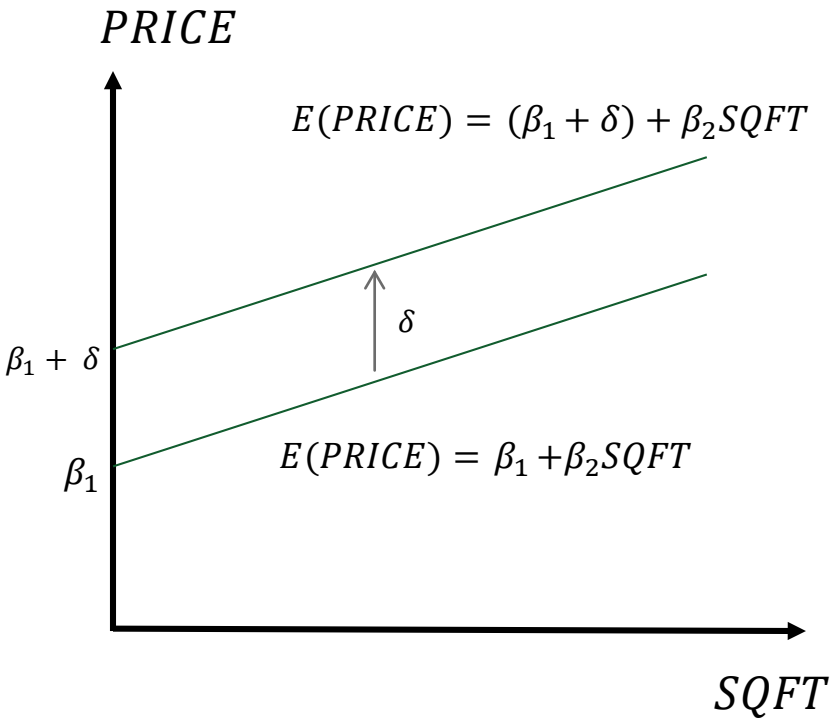
House	Price (1000 USD)	Living space (sqft/10)	Pool
i	***y***	x_2	***D***
1	205.452	23.46	0
2	185.328	20.03	0
…	…	…	…
999	300.728	28.74	0
1000	220.987	20.93	1

	Descriptive statistics		
Mean	247.66	25.21	0.204
Median	245.66	25.36	0
Max.	345.20	30.00	1
Min.	134.32	20.03	0
SDV	42.19	2.92	0.40

V. Dummy Variables (R-Example 2)

Model 1: Intercept Dummy

The following house price model allows the **intercept** to vary with the presence or absence of a pool: $PRICE_i = \beta_1 + \delta \boldsymbol{POOL_i} + \beta_2 SQFT_i + e_i$

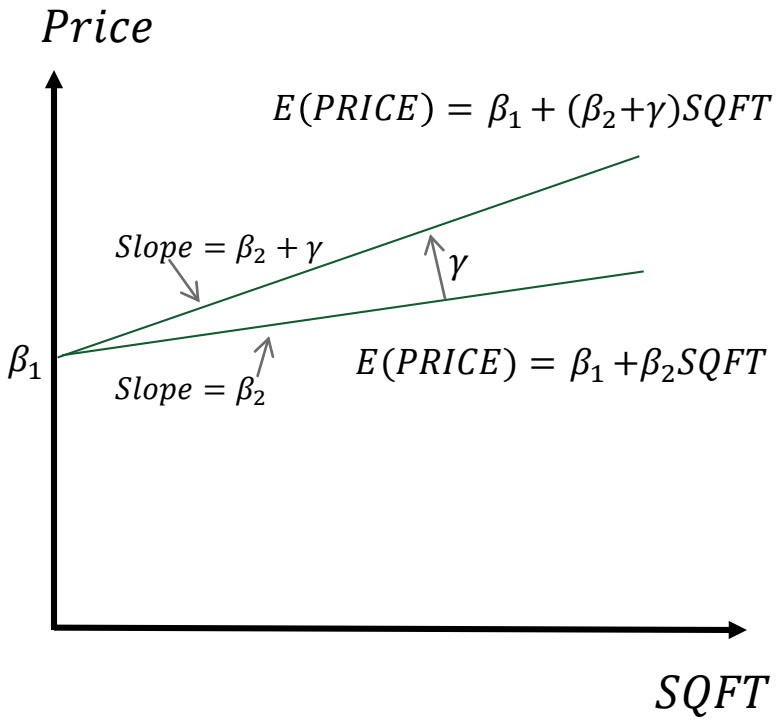



```
Call:
lm(formula = price ~ pool + sqft, data = utown)

Residuals:
    Min      1Q  Median      3Q     Max
-77.044 -30.262   3.746  30.302  74.530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.6767     9.3384   3.178  0.00153 **
pool          5.6904     2.6586   2.140  0.03257 *
sqft          8.6006     0.3673  23.418  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.88 on 997 degrees of freedom
Multiple R-squared:  0.3566,    Adjusted R-squared:  0.3553
F-statistic: 276.3 on 2 and 997 DF,  p-value: < 2.2e-16
```

V. Dummy Variables (R-Example 2)

Model 2: Interaction Term

The following house price model allows the **slope** to vary with the presence or absence of a pool: $PRICE_i = \beta_1 + \beta_2 SQFT_i + \gamma \boldsymbol{POOL_i} \times SQFT_i + e_i$

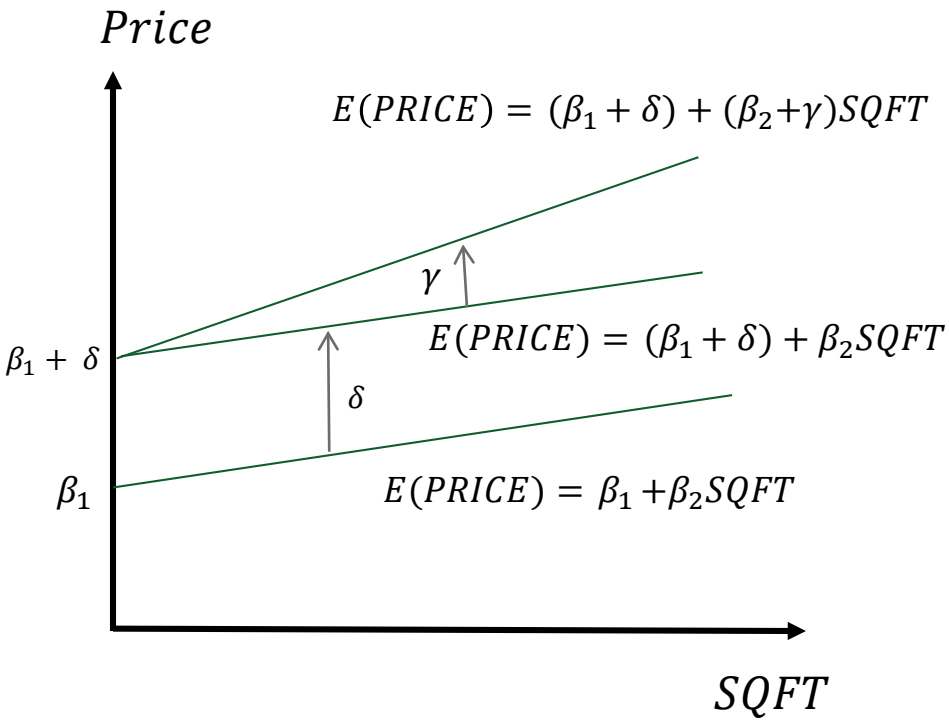



```
Call:
lm(formula = price ~ sqft + poolsqft, data = utown)

Residuals:
    Min      1Q  Median      3Q     Max
-77.239 -30.384   3.583  30.422  74.518

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.7784     9.3215   3.302 0.000995 ***
sqft          8.5583     0.3678  23.270  < 2e-16 ***
poolsqft      0.2191     0.1049   2.089 0.036963 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.88 on 997 degrees of freedom
Multiple R-squared:  0.3565,    Adjusted R-squared:  0.3552
F-statistic: 276.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

V. Dummy Variables (R-Example 2)

Model 3: Intercept Dummy and Interaction Term

A model that allows the **intercept and slope** to vary with the presence or absence of a pool: $PRICE_i = \beta_1 + \delta \boldsymbol{POOL_i} + \beta_2 SQFT_i + \gamma \boldsymbol{POOL_i} \times SQFT_i + e_i$




```
Call:
lm(formula = price ~ sqft + pool + poolsqft, data = utown)

Residuals:
    Min      1Q  Median      3Q     Max
-76.574 -30.046   3.619  30.324  74.497

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1672    10.3811   2.713  0.00678 **
sqft          8.6605     0.4089  21.179  < 2e-16 ***
pool         13.5182    23.6200   0.572  0.56724
poolsqft     -0.3107     0.9317  -0.334  0.73881
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.89 on 996 degrees of freedom
Multiple R-squared:  0.3567,    Adjusted R-squared:  0.3547
F-statistic: 184.1 on 3 and 996 DF,  p-value: < 2.2e-16
```

V. Dummy Variables (R-Example 2)

When interpreting regression models with dummy variables, the **reference group** is particularly important. The reference group is corresponds to $D_i = 0$. For example, in the house price model:

$$\widehat{PRICE}_i = 29.68 + 5.69\boldsymbol{POOL_i} + 8.60SQFT_i + e_i$$

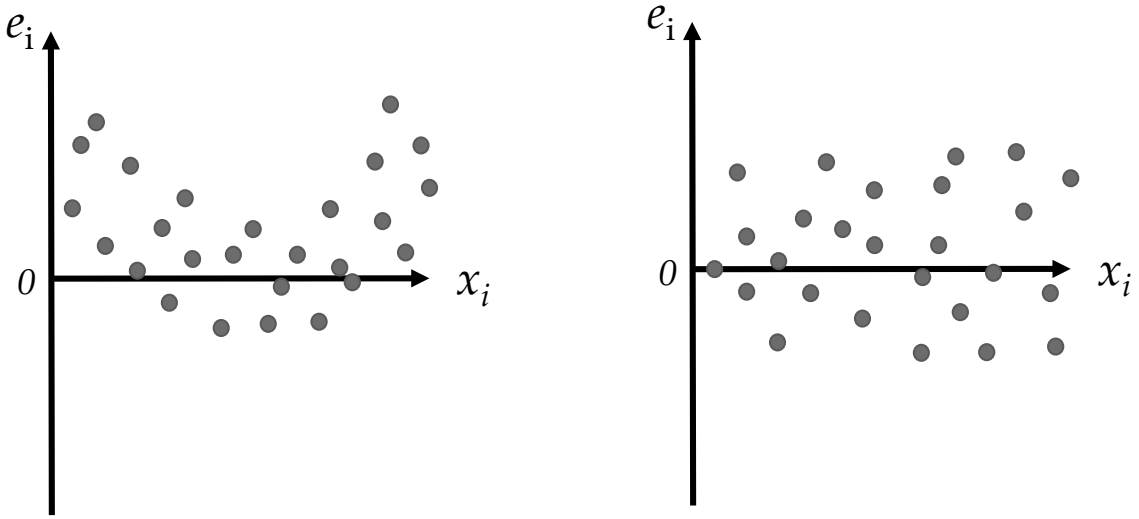
the reference group are houses with **no** pool (*POOL*$_i$ = *0*). Two alternative models:

$$\widehat{PRICE}_i = 35.37 - 5.69\boldsymbol{NOPOOL_i} + 8.60SQFT_i + e_i$$
$$\widehat{PRICE}_i = 29.68\boldsymbol{NOPOOL_i} + 35.37\boldsymbol{POOL_i} + 8.60SQFT_i + e_i$$

In these two models it holds that *NOPOOL*$_i$ = *1 - POOL*$_i$. **Note:** We cannot include both *POOL*$_i$ and *NOPOOL*$_i$ in a model containing a constant term because the two variables are perfectly collinear. This error is known as the **dummy variable trap**.

VI. Non-Linearity

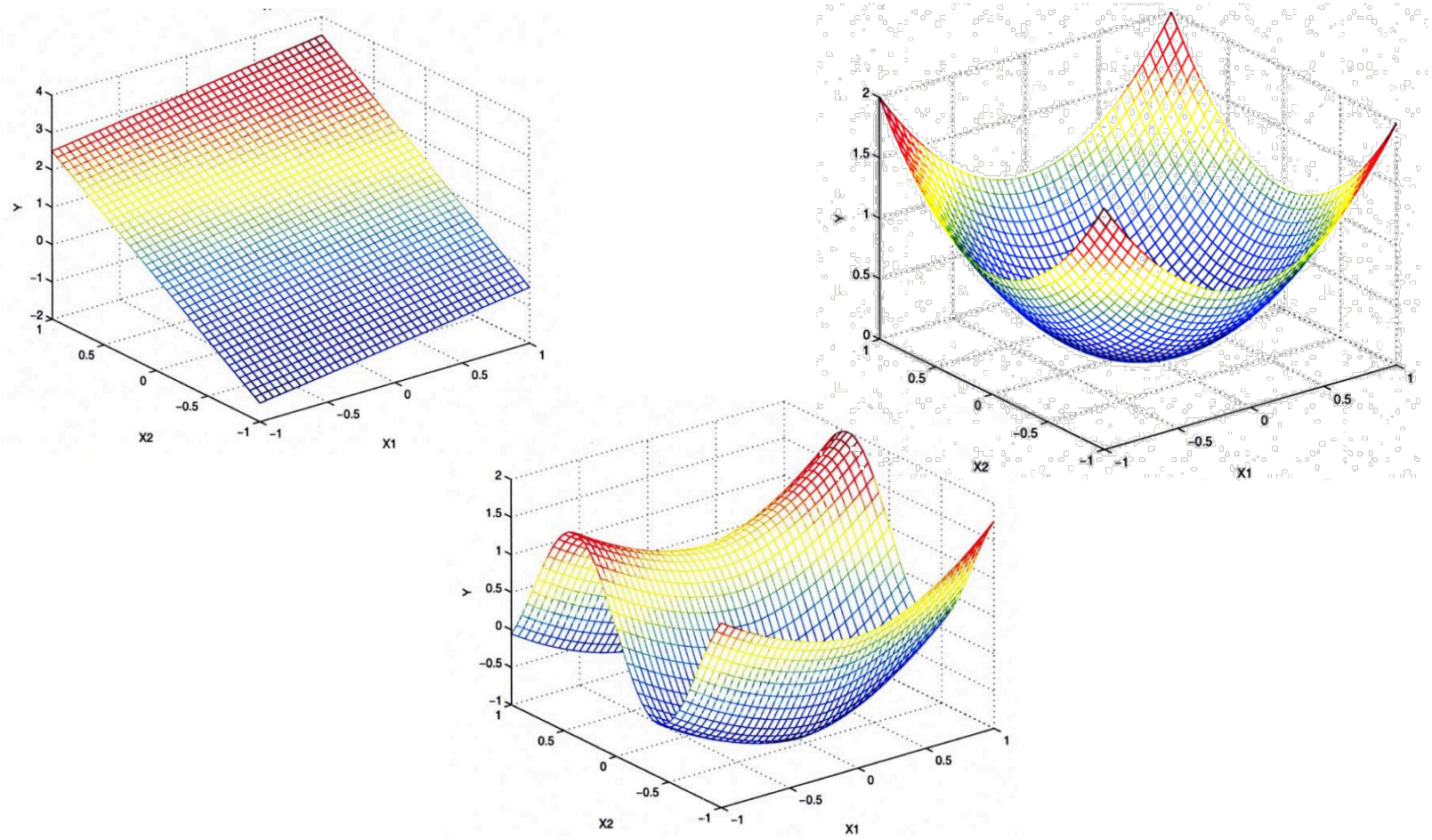
Not all relationships between variables are linear. However, if these non-linear relationships are described by a linear regression model, violations of **assumptions 2** ($E[e_i|x_i] = 0$) and **3** ($\mathrm{Cov}(e_i, e_j|x_i, x_j) = 0$) are the result. The expected value and the variance of the residuals (left) do not correspond to the assumed case (right).



Non-linearities can only be solved with a new specification of the regression model (e.g., logarithmic and quadratic transformations).

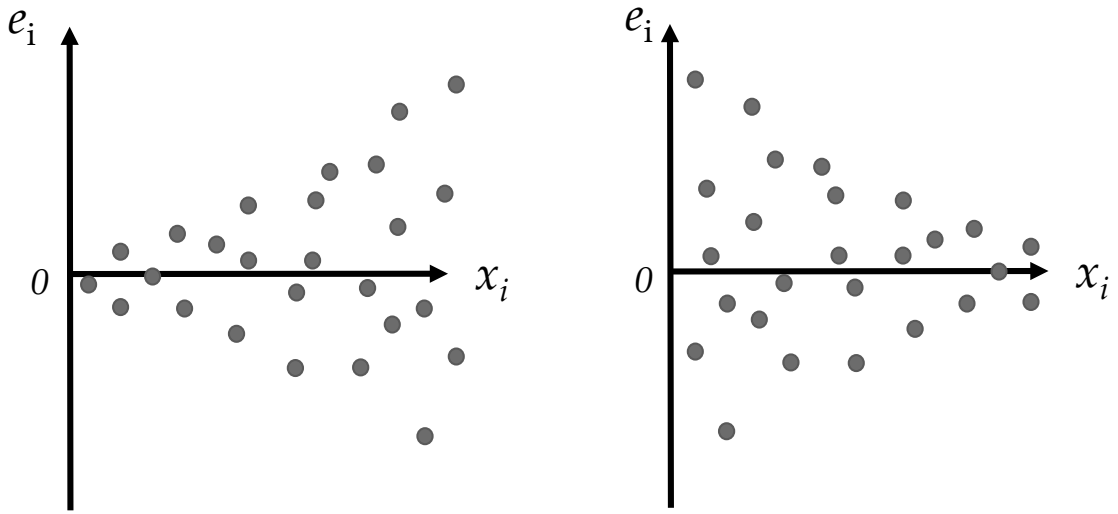
VI. Non-Linearity

The following graphs show a perfect linear relationship and two perfect non-linear relationships in a regression model with two independent variables.



VI. Heteroscedasticity

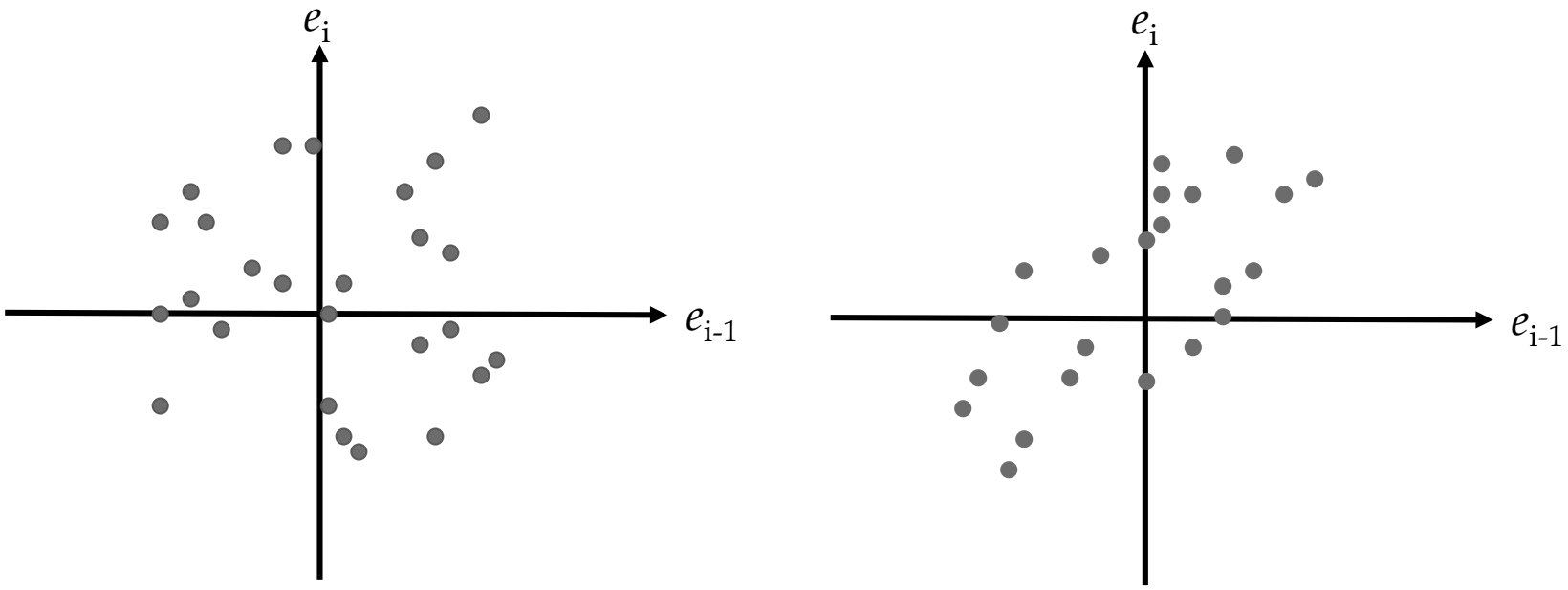
In case of a violation of **assumption 3** ($\mathrm{Var}(e_i|x_i) = \sigma^2$), heteroscedasticity is present. One refers to heteroscedasticity (vs. homoscedasticity) if the variance of the residuals increases or decreases over x_i.



Heteroskeedasticity is problematic as it leads to a **wrong estimation of the standard errors** of the coefficients. Hence, the statistical significance of the estimators is distorted and as a result hypothesis tests and confidence intervals are biased.

VI. Autocorrelation

In the case of autocorrelation, **assumption 4** ($\mathrm{Cov}(e_i, e_j | x_i, x_j) = 0$) is **violated**. The residuals in one period are correlated with the residuals in another period. This problem occurs mainly with **time series data**.



The graph on the right illustrates autocorrelation. In contrast to the graph on the left, the residuals are correlated with each other (linear pattern). Hence, similar to heteroskedasticity, the significance of OLS estimators is incorrectly assessed.

VI. Multicollinearity

In the case of multiple regressions one has to ensure that no perfect **multicollinearity** is present, which would violate **assumption 5**.

Perfect multicollinearity occurs when there is an exact linear relationship between two or more independent variables. **Weak multicollinearity** occurs when two independent variables are highly correlated but do not determine each other (e.g., dummy variables for lipstick users and a dummy variable for gender).

However, multicollinearity is not a major problem. In the case of perfect multicollinearity, the superfluous independent variables are simply dropped from the analysis. In the case of weak multicollinearity, the standard errors increase, which could be compensated by a larger sample size.