L17

Basic Statistics for Economists

Spring 2020

Department of Statistics

Stockholm University

Today

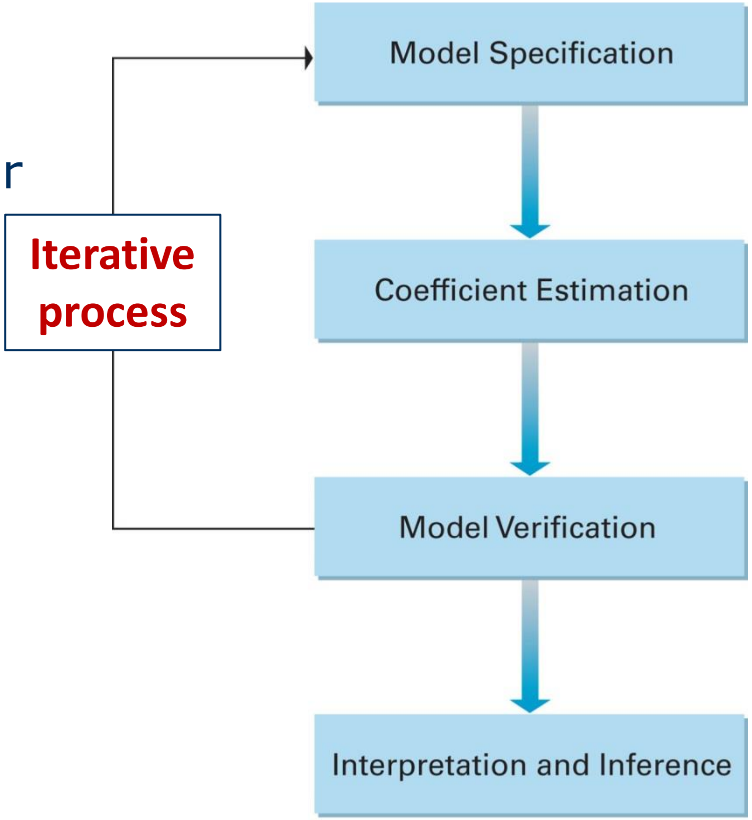
Selected topics from **12.8-9** + **13.1** + **13.4-6 NCT** + review

- Modelling
- Interpretation of t-test in multiple regression
  - correct statement of hypotheses
- Stepwise regression
- Dummy variables
- Specification bias, multicollinearity, heteroscedasticity
- Residual analysis

Maybe in a different order than that of NCT

Analysis and the model process

- This is true any study of relationships, developing explanatory models, models for prognosis, and so on
  - Regression analysis
  - Time series analysis
  - longitudinal models (panel data)
  - ARCH/GARCH-models
  - VAR-models
  - …



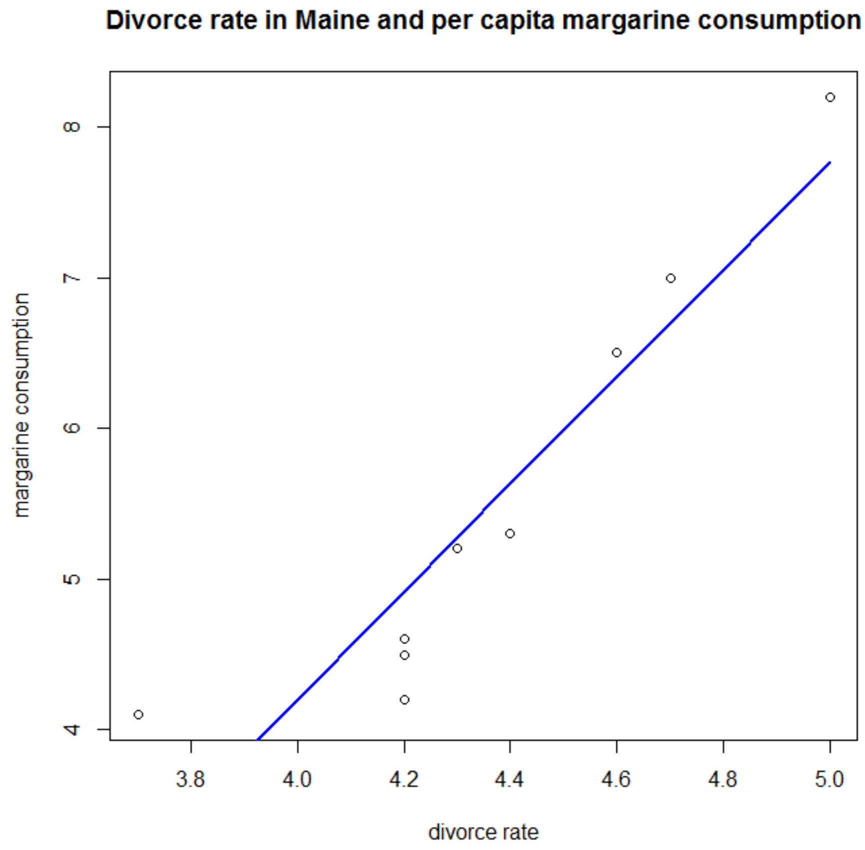
Stockholm University

Model building - method

- **Model specification**
  - Choice of variables **grounded in area specific theory**
  - linear – non-linear (we have not talk about this, read NCT 12.7)
- **Estimation of the model (we have covered this)**
  - **Data quality** should be always be considered and assessed
- **Model verification**
  - Does it match the theory: e.g. do the estimates have correct sign?
  - Are the assumptions correct – **residual analysis**
  - A prognosis should be compared to outcomes over time
- **Interpretation, inference, prediction (usage)**
  - Confidence interval, hypothesis test, prognosis

Stockholm University

Example: what not to do



source of data: http://www.tylervigen.com/spurious-correlations

Stockholm University

t-test for one of the coefficients, β_j

- Hypotheses to test if a coefficient is significant (adds something to the model) ***given that all the other variables are in the model***

- Suppose that $Y = \beta_0^{\circ} + \beta_1^{\circ} X_1 + \varepsilon$ is significant, a "good" model
- Test the extended model with an additional variable X_2

- Re-estimate the model with both X_1 and X_2 and test:

$$H_0: \beta_2 = 0 \mid \beta_1 \neq 0 \quad \text{vs.} \quad H_1: \beta_2 \neq 0 \mid \beta_1 \neq 0$$

- This is a clearer way of stating what is really being tested: does including X_2 significantly improve the model compared to just X_1
  - remember: R^2 and often also R^2_{adj} <u>increase</u> with every extra predictor

Idea: Stepwise regression

- Start with the variable that has the ***strongest correlation*** with $\boldsymbol{Y}$, lets say $\boldsymbol{X_1}$
  - estimate $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ and test $\boldsymbol{H_0: \beta_1 = 0}$

- Given that $\boldsymbol{X_1}$ is in the model, choose the one that ***adds the most***, e.g. $\boldsymbol{X_2}$
  - **NOTE**! This is not necessarily the second most one that correlates $\boldsymbol{Y}$ – it will depend on the how the variable correlates with $\boldsymbol{X_1}$
  - estimate $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ and test $\boldsymbol{H_0: \beta_2 = 0 \mid \beta_1 \neq 0}$

- Continue with more variables $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, …, choose the one that ***adds the most***
  - Estimate $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1} - \boldsymbol{\beta_{k+1}}$ and test $\boldsymbol{H_0: \beta_{k+1} = 0 \mid \beta_1 \neq 0, \dots, \beta_k \neq 0}$

- Stop when no available variable gives a significant improvement

Stepwise regression

Important to check at every step:

- That the $\boldsymbol{t}$-test is significant
- Changes in coefficients of the $\boldsymbol{X}$ that already are in the model
  - Large changes of the values?
  - Still significant? It happens that something that was significant now is non-significant!
  - Should you exclude a previously included variable?
- Study changes in $\boldsymbol{R^2}$, $\boldsymbol{R^2_{adj}}$, and $\boldsymbol{MSE = s_e^2}$
- Check if the model can be **grounded in area specific theory**
  - You can always find the ”best” model, but is it meaningful?

Stockholm University

Model building

- **Stepwise regression** is often built into statistical software packages. These functions carry out the steps **automatically**

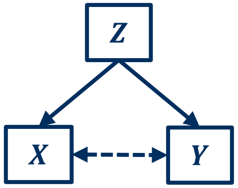
- Different strategies:
  - **Bottom-Up**: start with an ”empty” model and add $\boldsymbol{X}$ one by one.
  - **Top-Down**: start with a ”full” model and reduce it by removing the worst $\boldsymbol{X}$ at every step
  - **Add-Delete strategi**: at every step, the program will do what is “best,” by adding or removing variables

- You have to be careful!

Spurious (fake) relationships

When two variables covary stongly, but there is no explanation

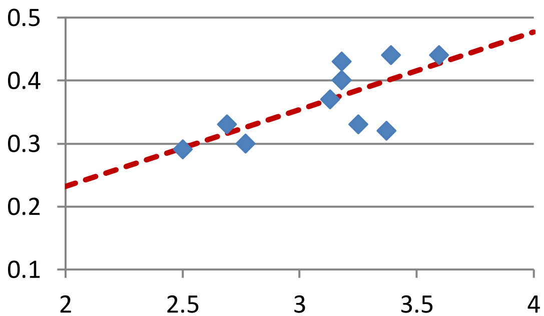
- A third variable influences



... or several...

- E.g. observations are sampled over **time**; both show a **trend**

Y = **Illnesses among public employees,** X = **Price of household electricity, kr**

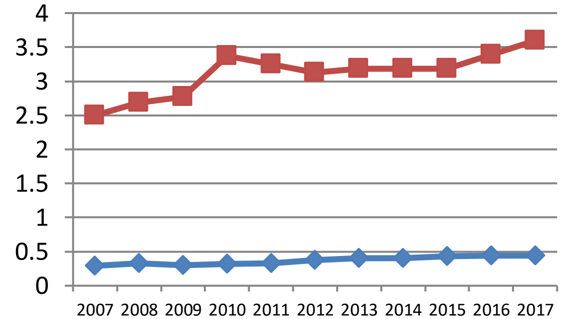


Source: SCB, www.scb.se

$r_{xy} = 0,712$

$R^2 = 0,508$

$p = 0,014$



Hard to see at this scale, but similar trends

Stockholm University

Dummy variables

NCT 12.8

- So far, $\boldsymbol{X}$ has mostly been **continuous** variables, but $\boldsymbol{X}$ can be **discreet**
- No problem if $\boldsymbol{X}$ is **categorical…**
- Special case: $\boldsymbol{X}$ is ***binary*** (***dichotomous***) with $\mathbf{0}$ and $\mathbf{1}$ as possible values
- If you for example want to compare group A to group B

$$\boldsymbol{X} = \begin{cases} \mathbf{0} & \text{if group A} \\ \mathbf{1} & \text{if group B} \end{cases}$$

- $\boldsymbol{X}$ is called ***dummy variable*** or ***indicator variable***

Dummy variable, cont.

- Model with one numerical predictor $\boldsymbol{X_1}$ and one dummy $\boldsymbol{X_2}$:

$$\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

Case $\boldsymbol{X_2 = 0}$: $\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 = \beta_0 + \beta_1 X_1}$

Case $\boldsymbol{X_2 = 1}$: $\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 = (\beta_0 + \beta_2) + \beta_1 X_1}$

- Same slope $\boldsymbol{\beta_1}$ but **different intercept** for the two groups

- Test for different intercept: $\boldsymbol{H_0: \beta_2 = 0 \mid \beta_1 \neq 0}$

Stockholm University

Dummy variable, cont.

- Extend the model with a ”new” variable = $\boldsymbol{X_1 \cdot X_2}$

$$\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2)}$$

Case $\boldsymbol{X_2 = 0}$: $\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 X_1}$

Case $\boldsymbol{X_2 = 1}$: $\boldsymbol{\mu_{Y|X} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1}$

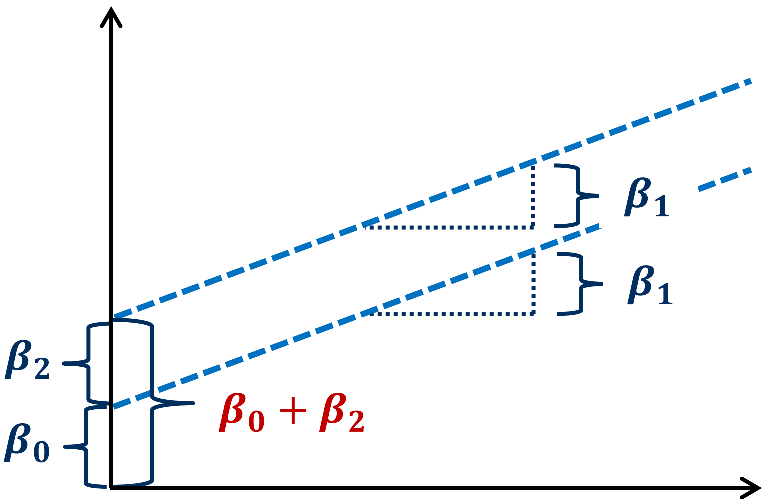
- Results in **different intercept <u>and</u> slope** for the two groups

- Test for different slope: $\boldsymbol{H_0 : \beta_3 = 0 \mid \beta_1 \neq 0,\ \beta_2 \neq 0}$

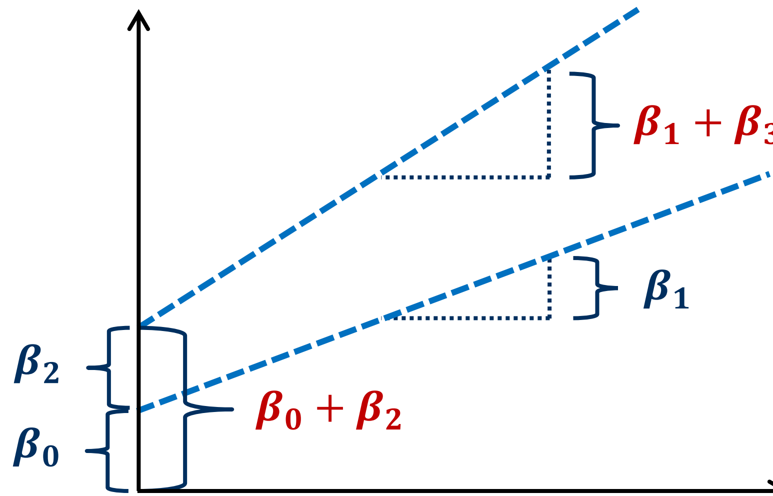
Stockholm University

Dummy variable, graphically

- Model with different intercepts, same slope



- Model with different intercepts and slope



Stockholm University

Short discussion of potential problems

NCT 13.4-6

- **Specification bias**
  - Something important is missing from the model
- **Multicollinearity**
  - Predictor variables correlates
- **heteroscedasticity**
  - not constant variance

Specification bias

Section 13.4

Occurs if important (significant) predictor variables are left out of the model.

- This will make the **OLS estimates** of the coefficients of the included variables **biased**.

Shown on 572, but not part of the course!

- The **statistical inference** (hypothesis test and CI) can in this case be **misleading**.
- The effect of the missing variables are instead captured by the residual variance (s_e^2) which typically becomes much greater, i.e. **worse precision**.

Stockholm University

Multicollinearity

Section 13.5

- When two or more of the explanatory variables are strongly correlated with each other ⇒ serious problems for the entire model!
- Illustration (constructed example):

Variables

- Y: **Sales (mkr)**
- X_1: **Print advertising (100 tkr)**
- X_2: **Other advertising (100 tkr)**
- $n = 7$

Y	X_1	X_2	X_3
7	4	1	5
9	7	2	9
16	9	5	14
19	12	8	20
25	15	10	25
26	17	14	31
33	20	17	37

Multicollinearity, cont.

```
Regression Analysis: Y versus X1
The regression equation is
Y = - 0,31 + 1,63 X1

Predictor     Coef   SE Coef      T      P
Constant    -0,306     1,388  -0,22  0,834
X1          1,6327    0,1058  15,42  0,000

S = 1,48186   R-Sq = 97,9%   R-Sq(adj) = 97,5%
```

$r_{x_1,y} = 0,990$

NOTE: R^2 close to 1 and significant coefficients ($p = 0$) for both models

```
Regression Analysis: Y versus X2
The regression equation is
Y = 6,68 + 1,55 X2

Predictor     Coef   SE Coef      T      P
Constant     6,676     1,282   5,21  0,003
X2          1,5485    0,1302  11,89  0,000

S = 1,90830   R-Sq = 96,6%   R-Sq(adj) = 95,9%
```

$r_{x_2,y} = 0,983$

Stockholm University

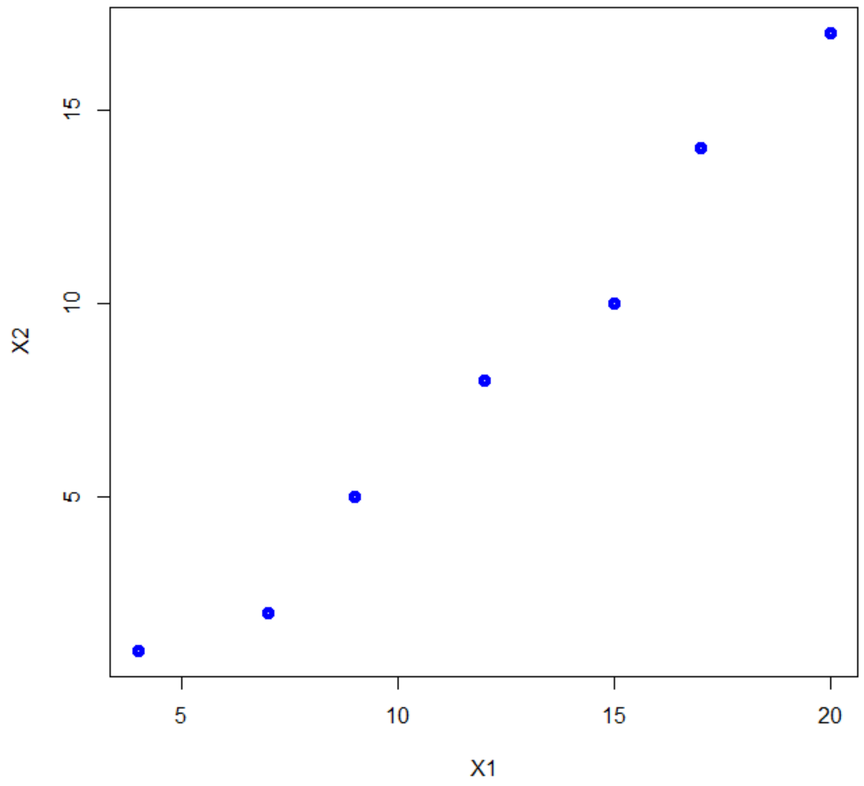
Multicollinearity, cont.

```
Regression Analysis: Y versus X1; X2
The regression equation is
Y = 0,89 + 1,34 X1 + 0,279 X2

Predictor     Coef  SE Coef     T      P
Constant     0,890    3,594  0,25  0,817
X1          1,3436   0,7948  1,69  0,166
X2          0,2791   0,7592  0,37  0,732

S = 1,62948   R-Sq = 98,0%   R-Sq(adj) = 97,0%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  522,81  261,40  98,45  0,000
Residual Error   4   10,62    2,66
Total            6  533,43
```

Comments: see next page

Stockholm University

Plot: Print ads vs. Other ads



Stockholm University

Multicollinearity, cont.

- Large $\boldsymbol{R^2}$ but marginally greater than the previous models
  - but $\boldsymbol{R^2_{adj}}$ has decreased compared to model with just $\boldsymbol{X_1}$
- Test $\boldsymbol{H_0: \beta_1 = \beta_2 = 0}$ gives significant results ($\boldsymbol{F}$-test, $\boldsymbol{p = 0}$)
  - But neither $\boldsymbol{b_1}$ or $\boldsymbol{b_1}$ are significantly different from zero with $\boldsymbol{p = 0,166}$ and $\boldsymbol{0,732}$ respectively
- Large change in the estimated $\boldsymbol{X_2}$ coefficient, slightly smaller change of the $\boldsymbol{X_1}$ coefficient
- Standard errors $\boldsymbol{s_{b_1}}$ and $\boldsymbol{s_{b_2}}$ are approximately 7,5 and 6 times greater compared to the model with only $\boldsymbol{X_1}$, the residual standard error $\boldsymbol{s_e}$ has also increased
  - We have also lost a degree of freedom

Stockholm University

Multicollinearity, cont.

- X_1 and X_2 covary so strongly that we cannot discern their respective effects on Y
  - Their correlation is $r_{x_1,x_2} = 0,989$

- Do try a model with all three explanatory variables X_1, X_2, and X_3 and see what happens!
  - The problems will be even more serious. Do you see why?

What can we do?

- Specify a different model!

Absolute multicollinearity

- Anta $\boldsymbol{Y}$ = shipping cost, $\boldsymbol{X_1}$ = distance in kilometers, $\boldsymbol{X_2}$ = distance in Swedish mil (10 km = 1 mil).
- Model including both:

$$\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ = \ \beta_0 + \beta_1 X_1 + \beta_2 \cdot 10 X_1}$$
$$\boldsymbol{= \ \beta_0 + (\beta_1 + 10\beta_2) X_1 \ = \ \beta_0 + \alpha_1 X_1}$$

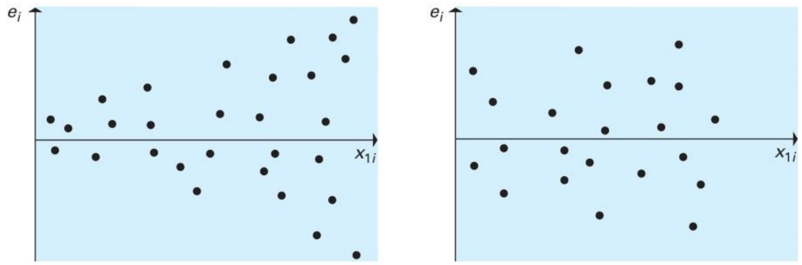
- This is really two variables, $\boldsymbol{X_1}$ and $\boldsymbol{Y}$, but we are trying to estimate this using 3 coefficients
- Mathematically, this will not work!

The relationships between the independent variables do not need to this obvious. Strong correlations can also occur in more complex situations. Be careful!

heteroscedasticity

Section 13.6

- <u>Model assumptions</u>: the error terms are normally distributed with expected value = 0 and **constant variance σ_ε^2**
- This is often not a good assumption



(a) Heterosecedasticity (b) No Apparent Heterosecedasticity

Plot the residuals (e_i) against each of the explanatory variables (x_{ki})

- E.g. Larger companies are subject to more factors with may affect variation of $\boldsymbol{Y}$, compared to smaller companies.

Stockholm University

heteroscedasticity, cont.

Problem:

- OLS estimates are **not efficient** (not best precision)
- Hypothesis test and confidence interval for the coefficients are **not correctly defined** – e.g. you can not be sure that your CI or p-values are even approximately correct.
- **NCT** describes a test for homoscedasticity – **skip it!**
- Do plot the residuals $\boldsymbol{e_i}$ against the explanatory variables $\boldsymbol{x_{ki}}$ and also against the predicted values $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$!

Verification of the model: Residual analysis

NCT 12.9 p. 534-537

- The model assumption must be verified!

1. $\boldsymbol{Y}$ is a ***linjear*** function of $\boldsymbol{X_1, \dots, X_K}$ + error term $\boldsymbol{\varepsilon_i}$
2. $\boldsymbol{X_1, \dots, X_K}$ are all ***independent*** of the error terms $\boldsymbol{\varepsilon_i}$.
3. The error terms are $\boldsymbol{\varepsilon_i}$ ***normally distributed*** r.v. with expected value $\boldsymbol{E(\varepsilon_i) = 0}$ and ***constant variance*** $\boldsymbol{\sigma_\varepsilon^2}$
4. The error terms are pairwise ***independent*** of each other

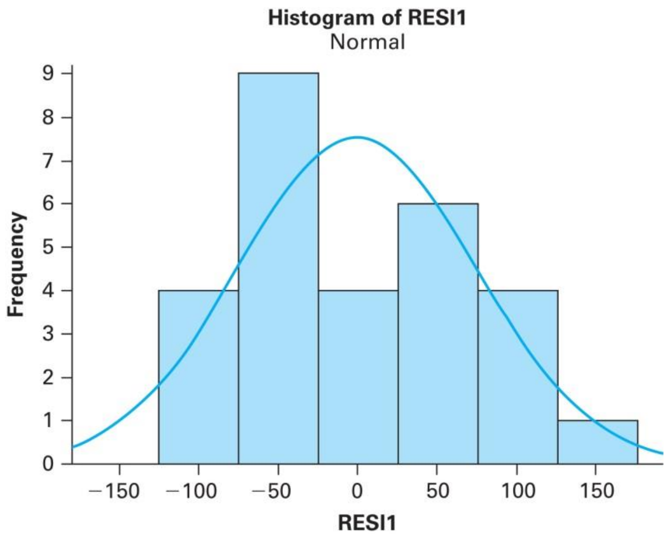
- The residuals $\boldsymbol{e_i}$ are proxies of $\boldsymbol{\varepsilon_i}$ $\Rightarrow$ **analyze the residuals**
  - We have just talked about the assumption of constant variance and how we can check this

Stockholm University

Analysis of residuals, cont.

Normally distributed error terms $\boldsymbol{\varepsilon_i}$

- There are test, e.g. Jarque-Bera test. (not covered)
- **Histogram** of $\boldsymbol{e_i}$ or ”Probability plots” (not covered)



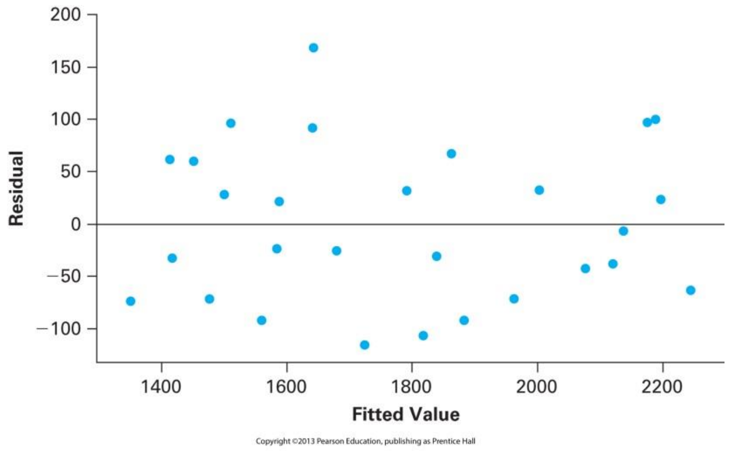
Does not have to be very normal like. If it looks ”almost” normal, then that is ok, too.

The other assumptions are more critical: independence, constant variance, linearity

Analysis of residuals, cont.

$X_1, \dots, X_K$ **independent** of error terms ε_i

- Plot residuals e_i against each of $X_1, \dots, X_K$
- Also plot against the predicted values $\hat{y}$ (= linj.komb. av $X_1, \dots, X_K$)
- Should be void of clear patterns and should look random

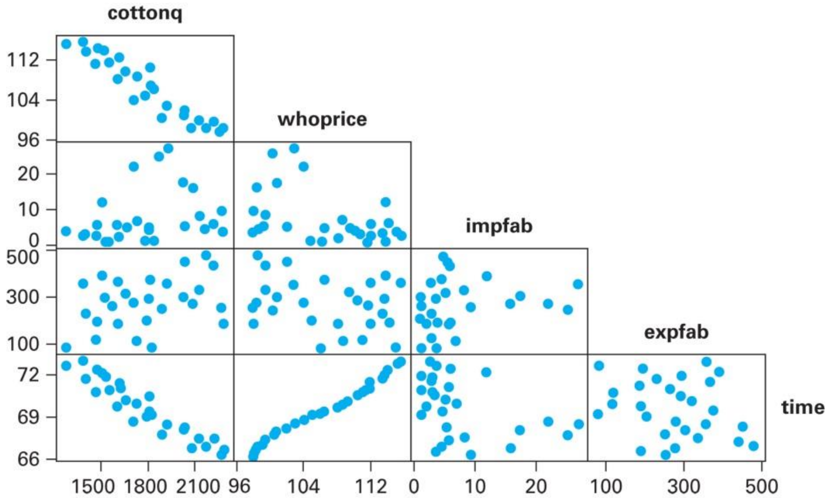


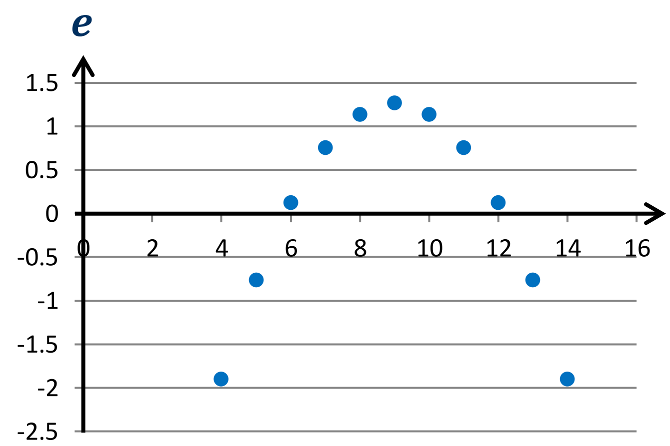
A ”clear pattern” could be any kind of pattern. To understand what causes the pattern requires more thinking and discussion. But dependence, non-linearity causes typical patterns (see next page).

Analysis of residuals, cont.

Linear correlation

- Plot Y against each of $X_1, \dots, X_K$; matrix plots
- Also plots with e_i against each of $X_1, \dots, X_K$ and $\hat{y}$ can indicate non-linearity

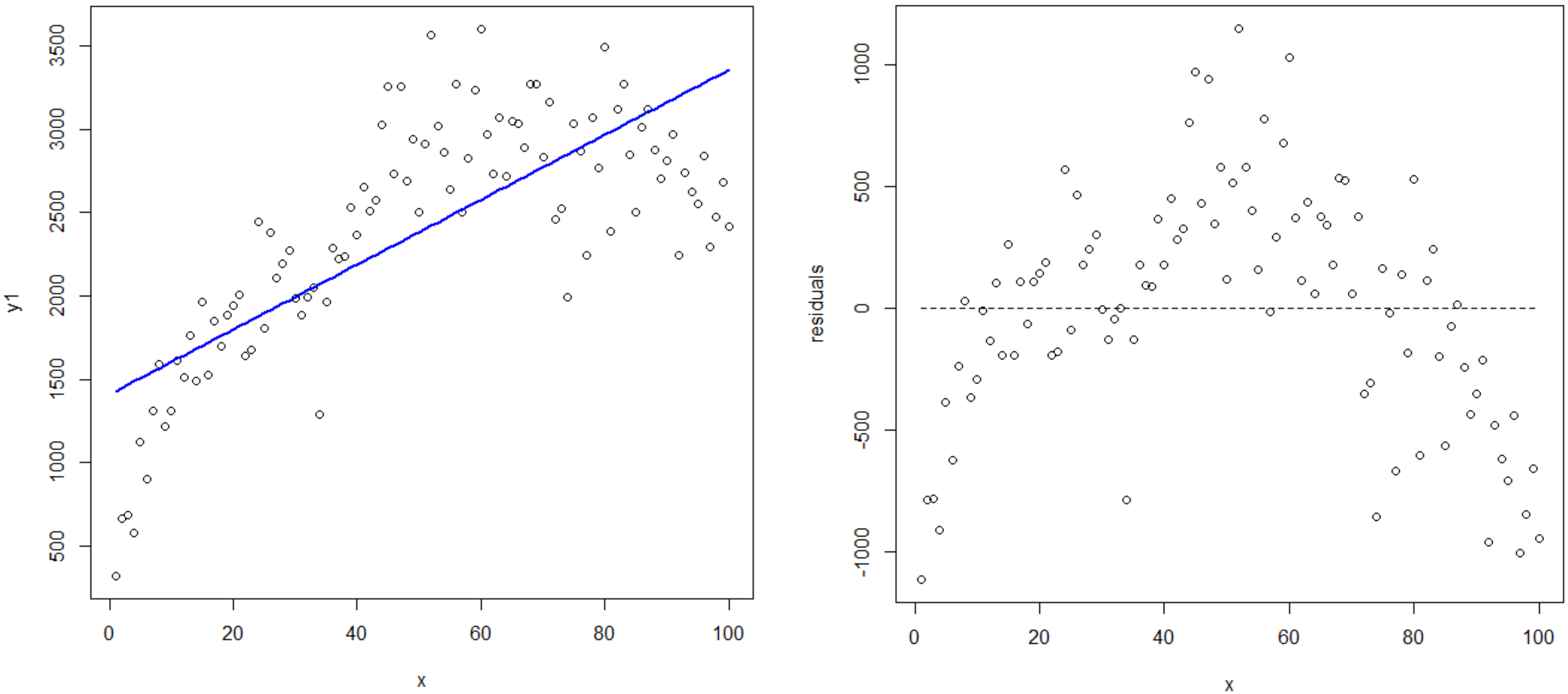




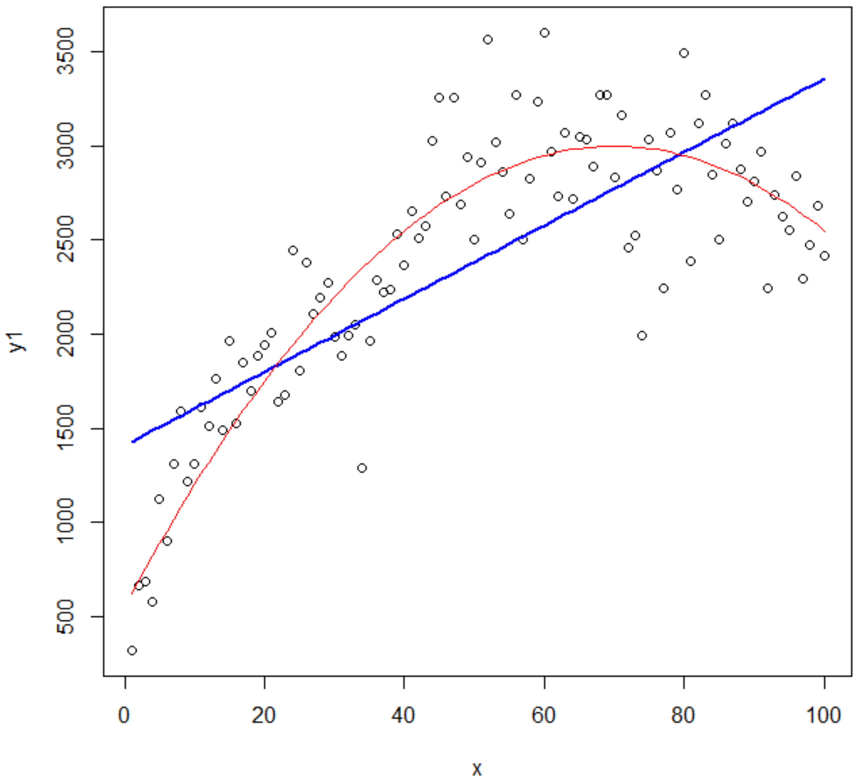
Clear pattern because of non-linearity!

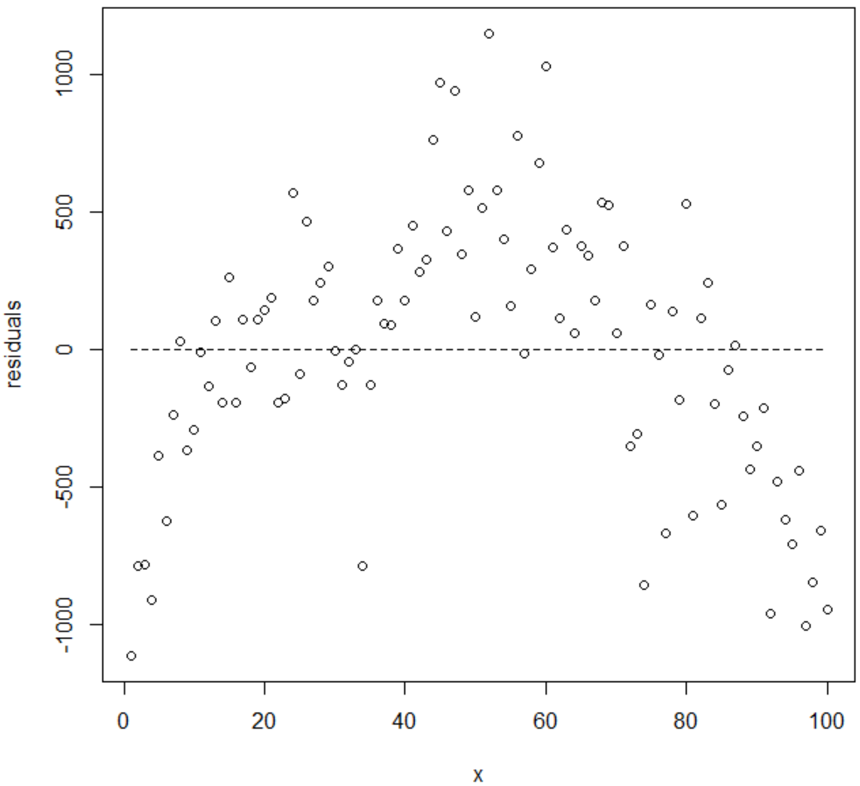
Stockholm University

Example: non-linear relationship



Example: non-linear relationship, cont.



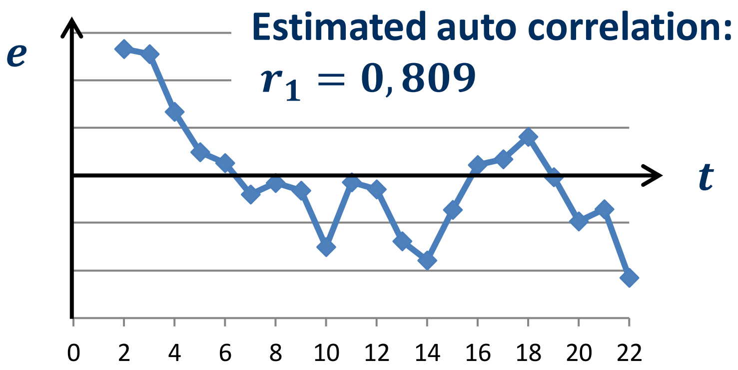


Stockholm University

Residual analysis, cont.

The error terms $\boldsymbol{\varepsilon_i}$ are **dependent** of each other

- Common vid repeated measures over time
  - The value of today influences the value of tomorrow
- Let $\boldsymbol{\varepsilon_t}$ denote the error term at time $\boldsymbol{t}$
- **Auto correlation**, serial correlation: $\boldsymbol{\rho_1 = Corr(\varepsilon_t, \varepsilon_{t-1})}$



Note! Described in section 12.9, but not part of reading instructions!

Can be difficult to notice if the error terms are dependent in some other way.

Area specific knowledge required!

Stockholm University

Next time

- Another kind of test: **χ^2-test**
- Used for **categorical data**
  - Goodness-of-fit test
  - Independence test/ Homogeneity test

Stockholm University