Counterfactual Impact Assessment

Day 2

Klara Major

- Understanding correlation

- Understanding why variables are correlated

- Understanding correlation through XY plots

- Correlation between several variables

- Relationship between two variables numerically

- Notation: correlation between X and Y is r_{XY}

- Square of correlation (r_{XY}^2): what percentage of Y's variation is explained by X = what percentage of X's variation is explained by Y

- Excel: CORREL() function

- Value between -1 and 1

- Positive value – positive relationship. r=0: no correlation between the variables

- Correlation between X and Y = correlation between Y and X

- Correlation of a variable with itself = 1

Example

5

- Correlation between unemployment rate and GDP/capita = -0,62

- Negative relationship

- Higher GDP/capita – lower unemployment

- The standard deviation of county level GDP/capita explains 38% of the standard deviation of unemployment rate (0.62*0.62=0.384)

- Using the data in HPRICE.XLS, calculate and interpret the correlation between Y = house price (labeled "sale price" in HPRICE.XLS), X = lot size and Z = number of bedrooms (labeled "n_bedroom") and W = number of bathrooms (labeled "n_bath") variables.

- Discuss to what extent it can be said that W causes Y.

- Does one variable „cause" the other?

- Correlation does not reveal the direction of causality

- There might be no causality at all

- Previous examples? (GDP – unemployment, GDP – number of enterprises)

- M variables – M(M-1)/2 correlations
- Correlation matrix of 3 variables (X, Y, Z):

	X	Y	Z
X	1		
Y	r_{XY}	1	
Z	r_{XZ}	r_{ZY}	1

- XY-plot (scatter plot) diagram between two variables

- „How difficult is to draw a line fitting the points?"
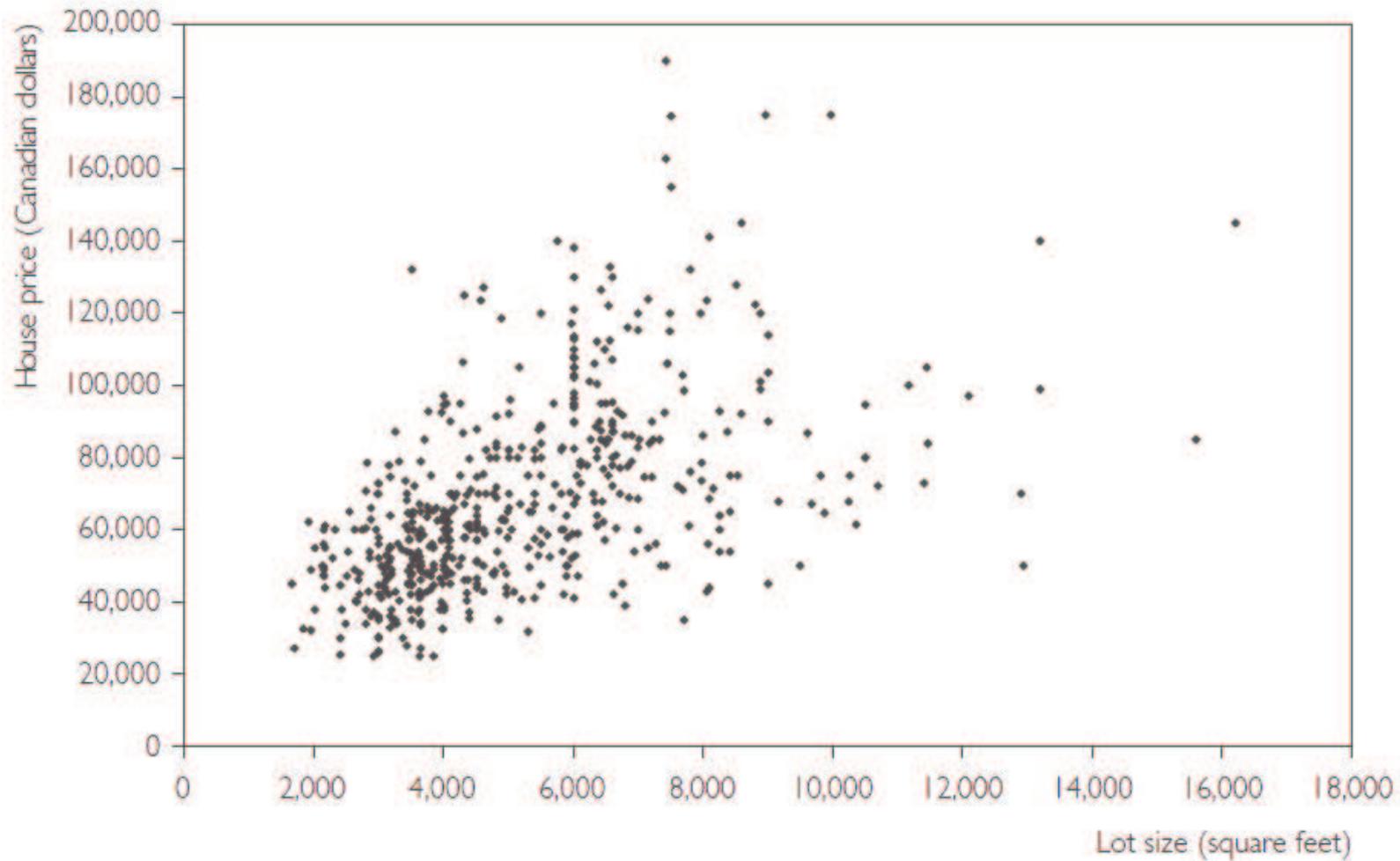
- „How scattered are the points?"
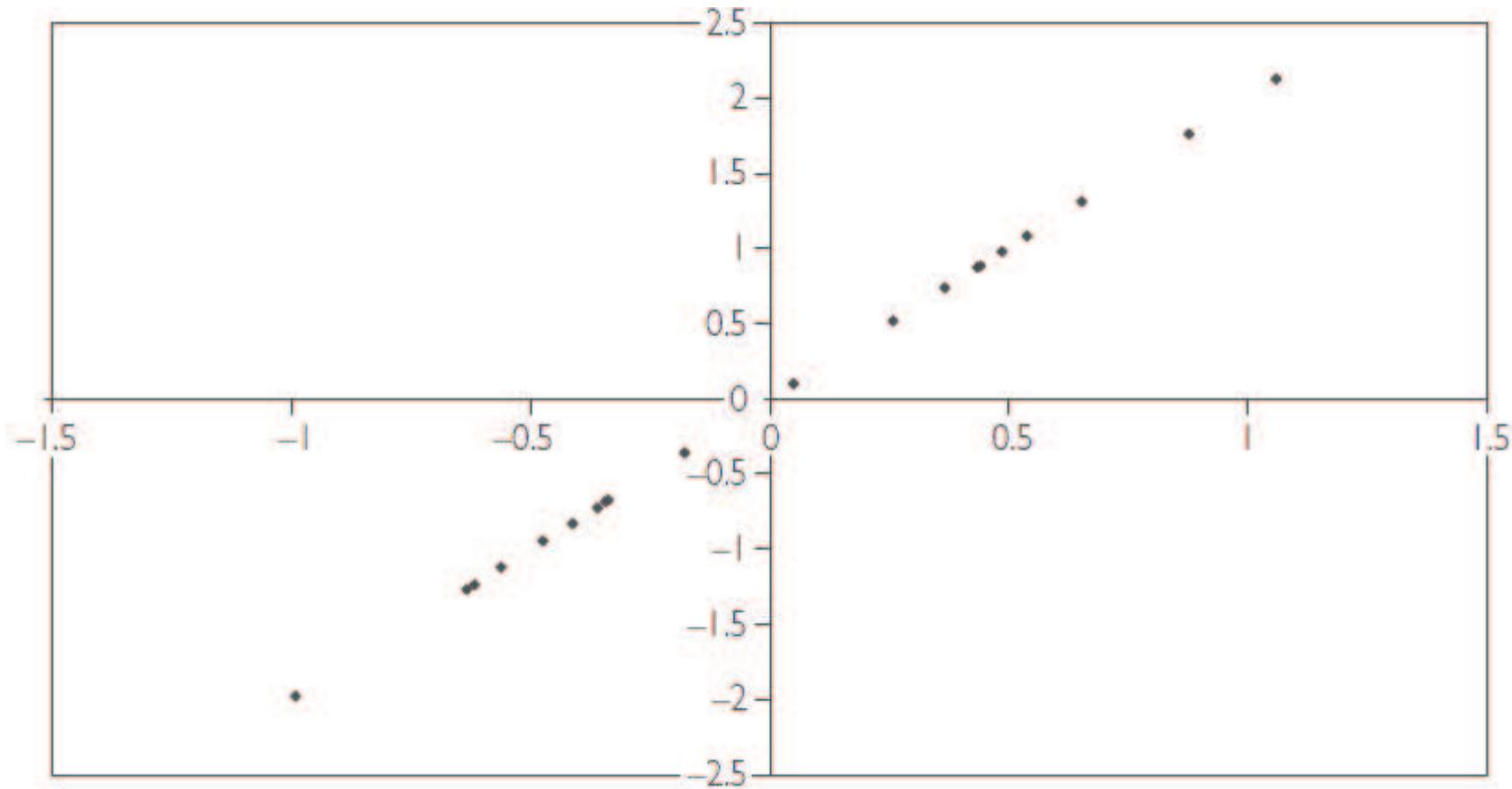
Fig. 3.1 XY-plot of price versus lot size.

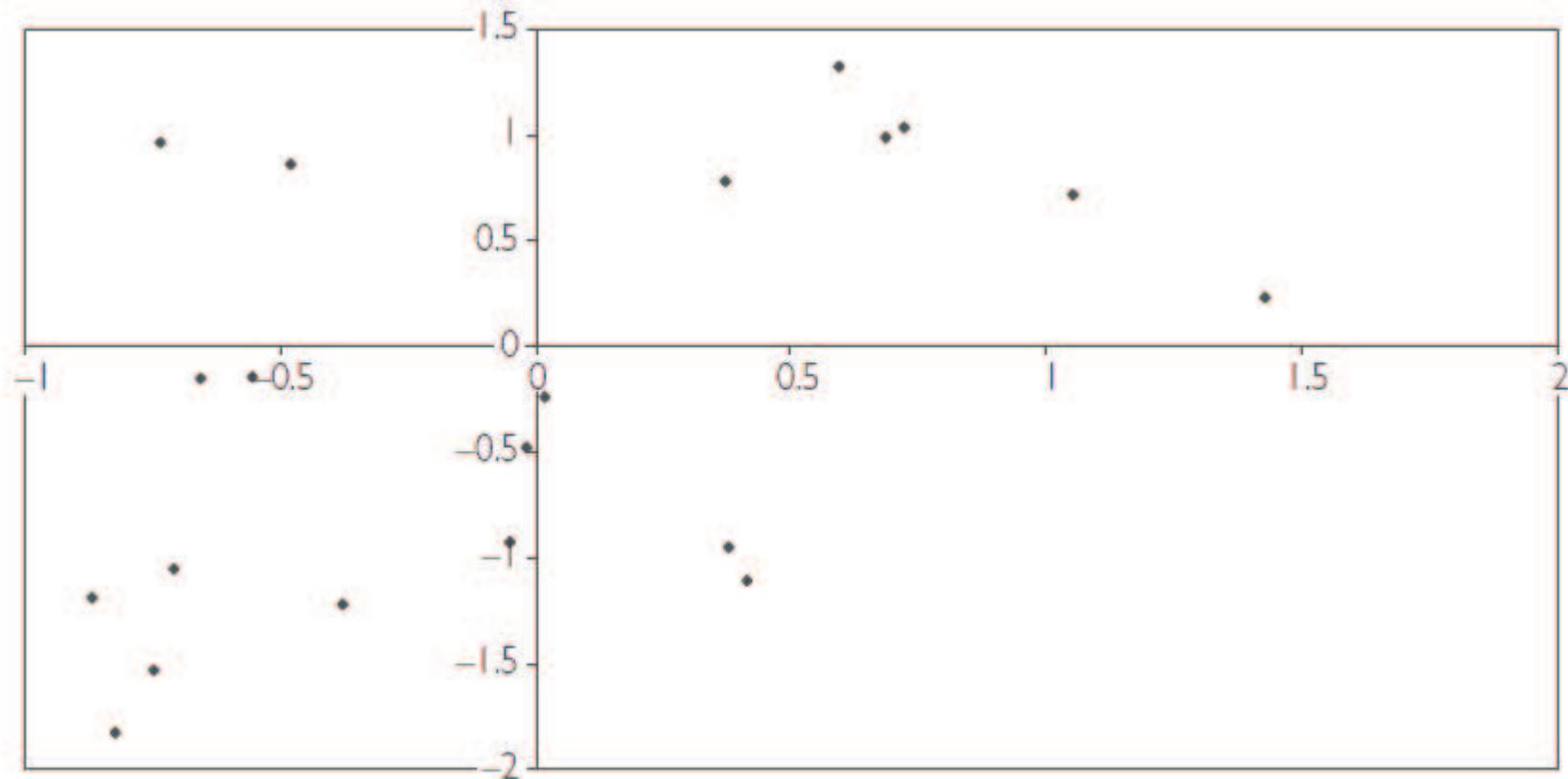Fig. 3.2 *XY*-plot of two perfectly correlated variables (*r* = 1).

Fig. 3.3 XY-plot of two positively correlated variables ($r = 0.51$).
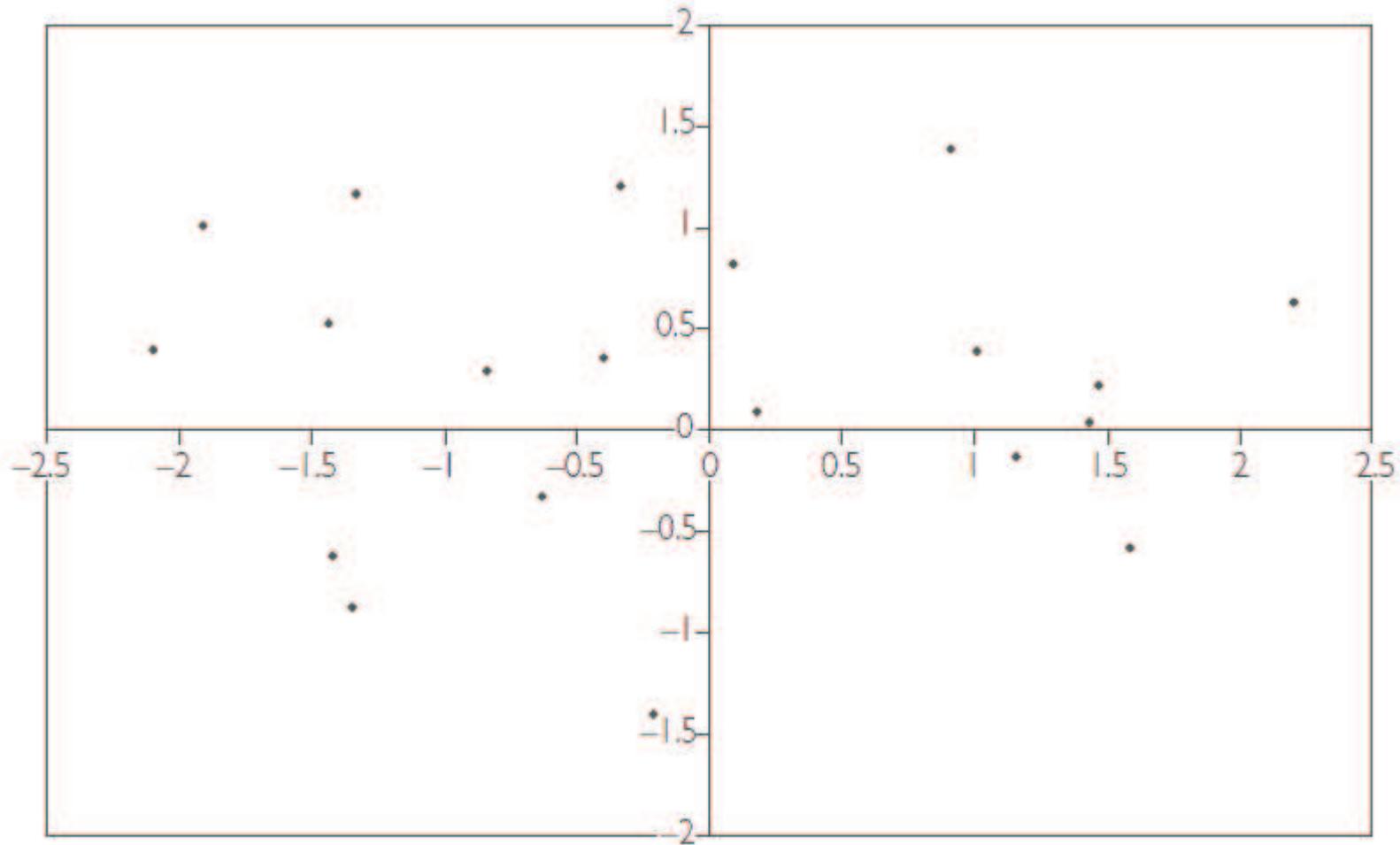
Fig. 3.4 *XY*-plot of two uncorrelated variables ($r = 0$).

- Numerical analysis of the relationship between variables:

- Correlation:

  - Between 2 variables

  - Causality?

- Regression:

  - Complex relationships (more variables)

  - There might be an underlying economic model – causality

  - Examples: wage regression (education, …), inflation regression (interest rate, …)

- *An introduction to simple regression*

- *Statistical aspects of regression*

.

- Y dependent variable, X explanatory variable (regressor)

- Assumption: linear relationship

- Regression line:

$$Y = \alpha + \beta X$$

- Reality: the data do not fit a line
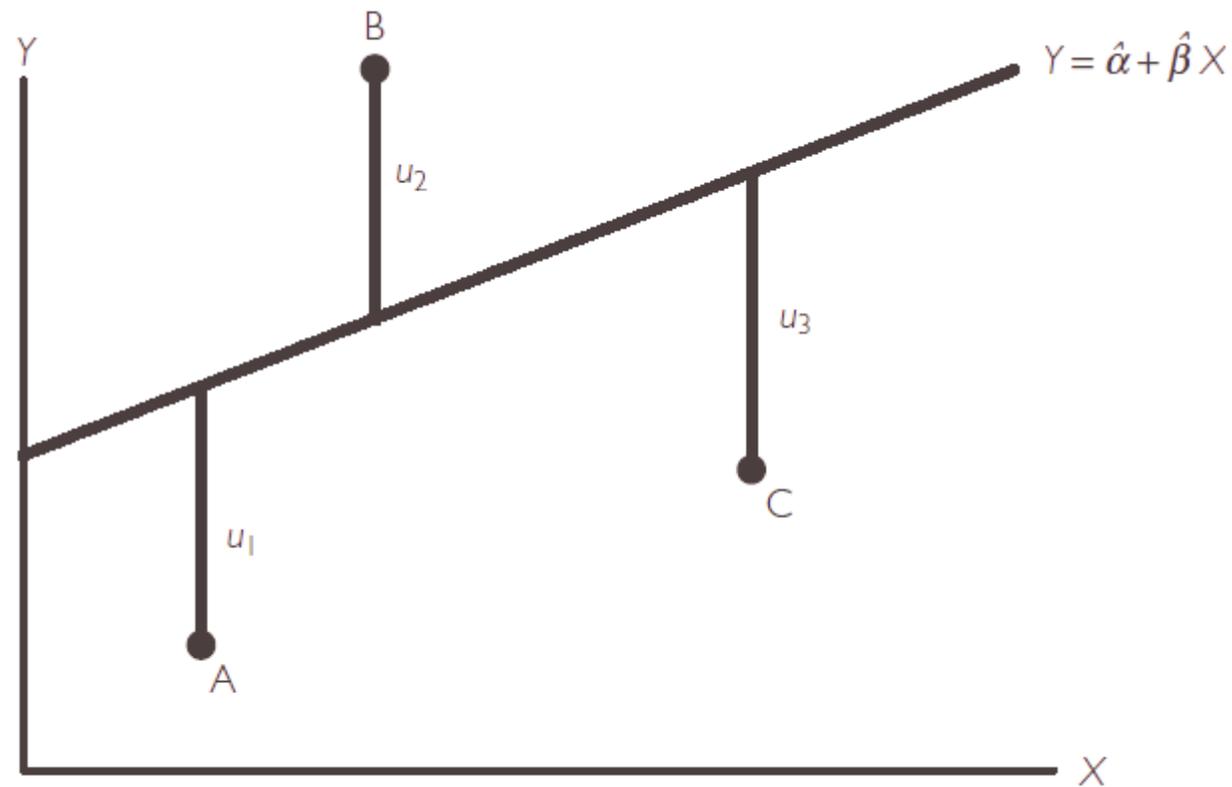
Fig. 4.1 Best fitting line for three data points.

Example: advetisement expenditures
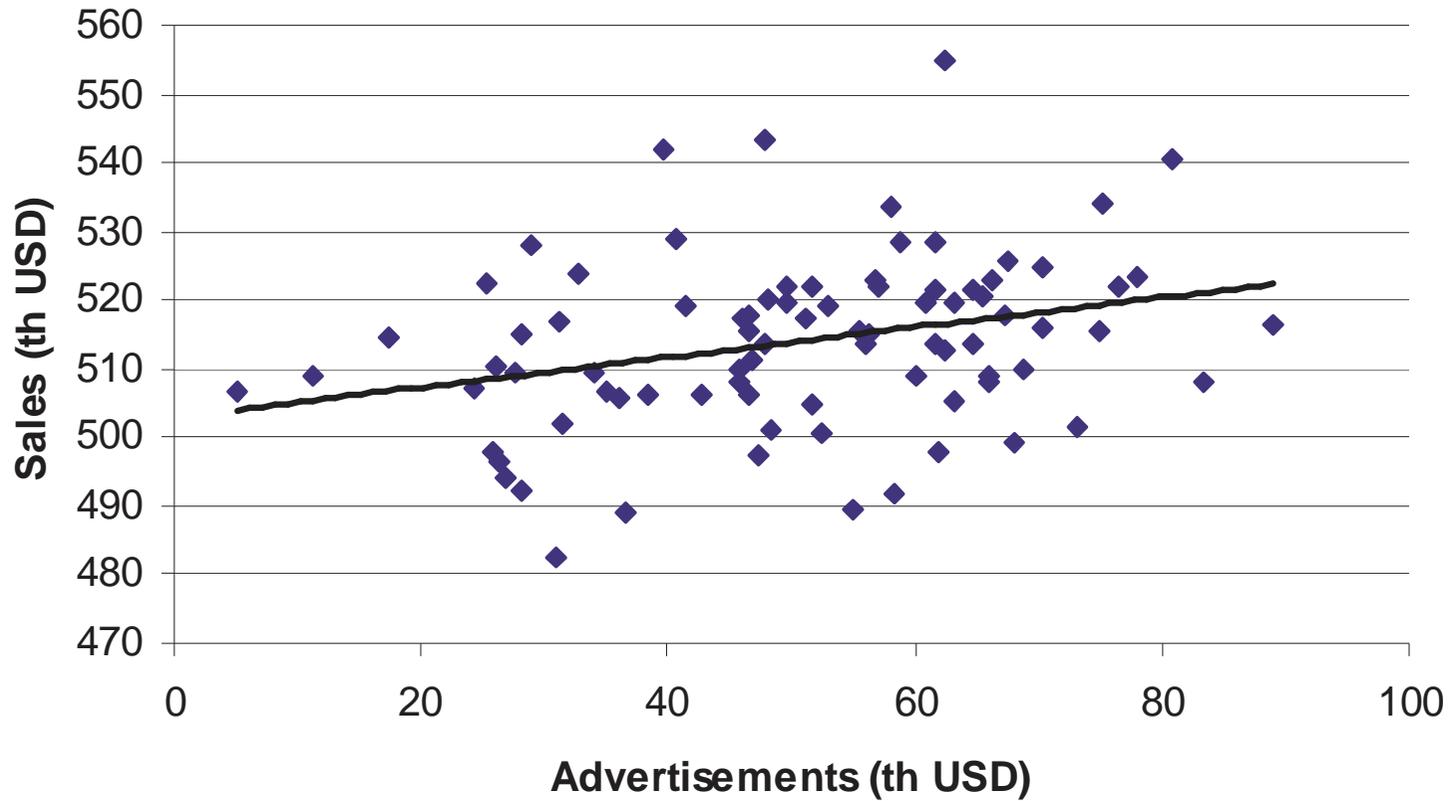
- Linear regression: no perfect fit

  - Omitted, unobservable variables

  - Non-linear relationship

- Regression model with error term:

$$Y = \alpha + \beta X + e$$

- Error term (disturbance): distance between the data point and the regression line

- Causality (model)? Generalization of correlation?

- Unknown value of coefficients

- Estimated coefficients: coefficients of the best fitting line

  - Notation:

$$\hat{\alpha}, \hat{\beta}$$

- Residual:

$$Y = \hat{\alpha} + \hat{\beta}X + u$$

$$u \neq e$$

- Best fitting line – minimal sum of square of residuals

$$SSR = \sum_{i=1}^{N} u_i^2$$

- Ordinary least squares (OLS) estimation

- Regression line with Excel: Diagram/Trend line

- Excel: Data analysis/Regression – estimate and interpret coefficients of the advertisement examples

- Sensitivity of the coefficients to scaling (unit of measurement)?

- Estimated coefficients:
    - 502.92 – intercept parameter (constant);
    - 0.22 – coefficient of advertisements (slope)

- Interpretation?

- Slope:
    - Average change in Y if X increases by one unit
    - Marginal effect

- The Excel data set FOREST.XLS contains data on Y = deforestation, X = population density, W = change in cropland and Z = change in pasture land. Run a regression of Y on X and interpret the results.

- OLS: finding the best fitting line

- How good is the fit?

    - Measure: R^2

    - Simple (univariate) regression:

        square of correlation= R^2

- Regression line: $Y = \alpha + \beta X + e$
- Estimated/fitted/forecasted value: $\hat{Y} = \hat{\alpha} + \hat{\beta} X$
- Comparison of the two – how good is the fit
- Residual: $u = Y - \hat{Y}$ vs. error term.
- Advertisement example:
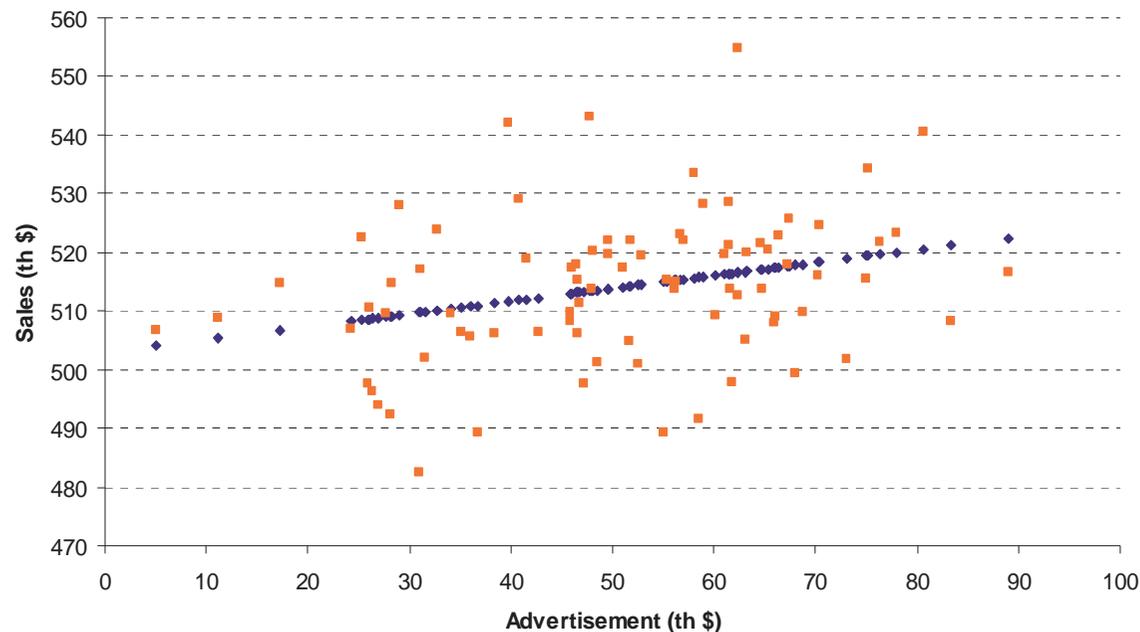
Total sum of squares : $TSS = \sum (Y_i - \bar{Y})^2$

Variance : TSS/(N -1)

Regression sum of squares : $RSS = \sum (\hat{Y}_i - \bar{Y})^2$

Sum of squared residuals : $SSR = \sum (Y_i - \hat{Y}_i)^2 = \sum u_i^2$

$$TSS = RSS + SSR$$

$$R^2 = 1 - \frac{SSR}{TSS} = \frac{RSS}{TSS}$$

- What percentage of the variance of Y is explained by X

$$0 \leq R^2 \leq 1$$

- perfect fit: $R^2=1$

Regression statistics		
r-squared	0,434	
ANALYSIS OF VARIANCE		
	df	SS
Regression	1	25,828
Residual	68	33,618
Total	69	59,446

- Nonlinear relationship between X and Y

- Common examples:
  - Quadratic
  - Logarithmic
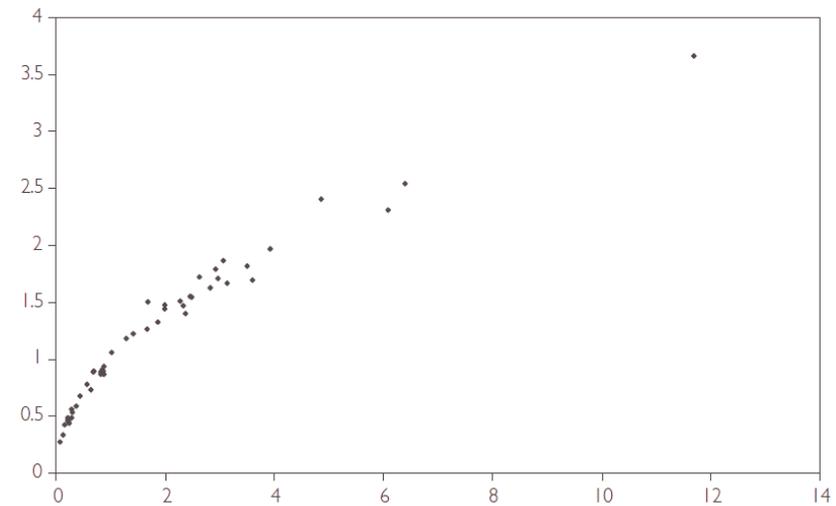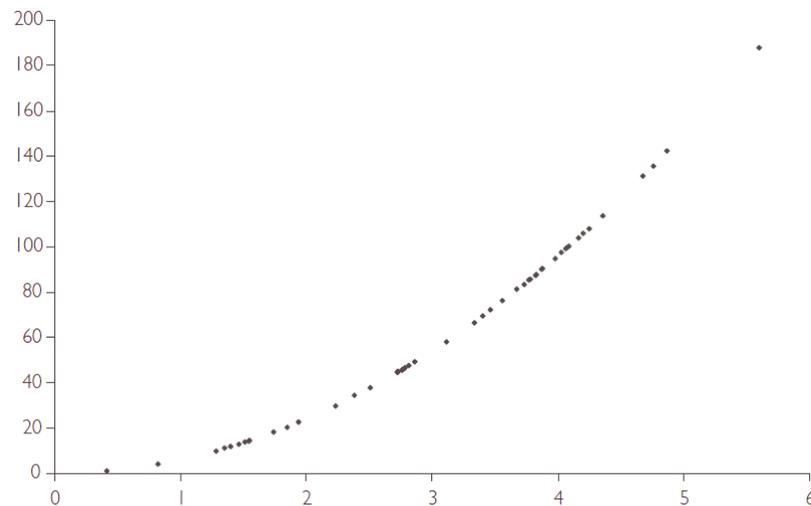
- Can ensure linear relationship
- Easy to interpret – elasticity:

$$\ln Y = \alpha + \beta \ln X$$

$$\beta = \frac{d \ln Y}{d \ln X}$$

- If X increases by one %, Y increases by beta % on average
  - Unit of measurement does not matter
  - Approximation of % change:

$$100 \cdot d \ln Y$$

- Use the data in the example related to costs of production in the electric utility industry (ELECTRIC.XLS), where Y = cost of production and X = output.

  - Run a regression of Y on X.

  - Take log transformations of both variables.

  - Run a regression of ln(Y) on ln(X) and interpret your results verbally.

- Real values of the coefficients are unknown

- Estimated based on a sample

  - Estimated value is not exactly equal to the true value

- Point estimation: does not reveal the uncertainty

- sample size

- size of error

- range of X variable

Very small sample size
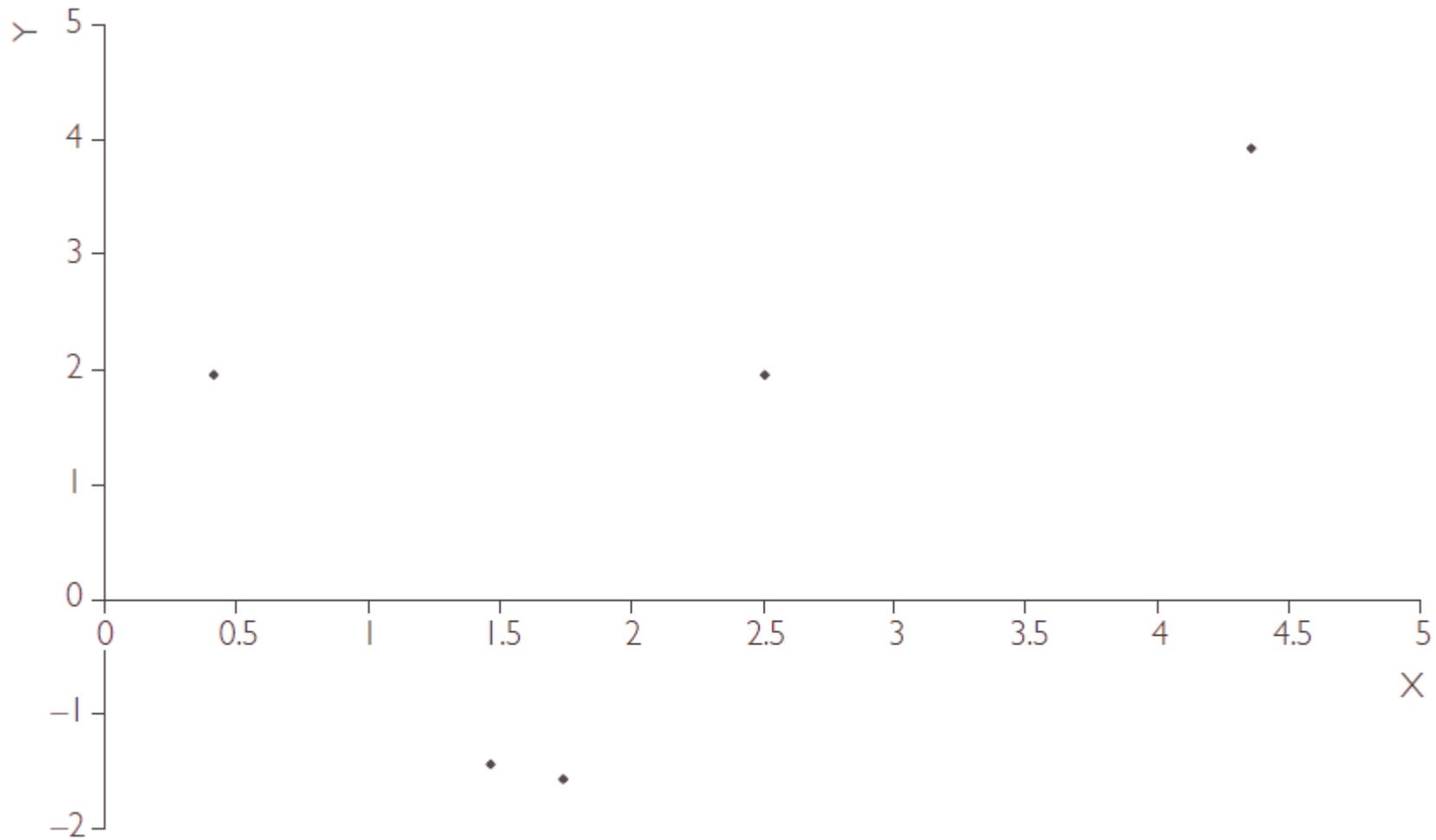
Large sample size, large error variances
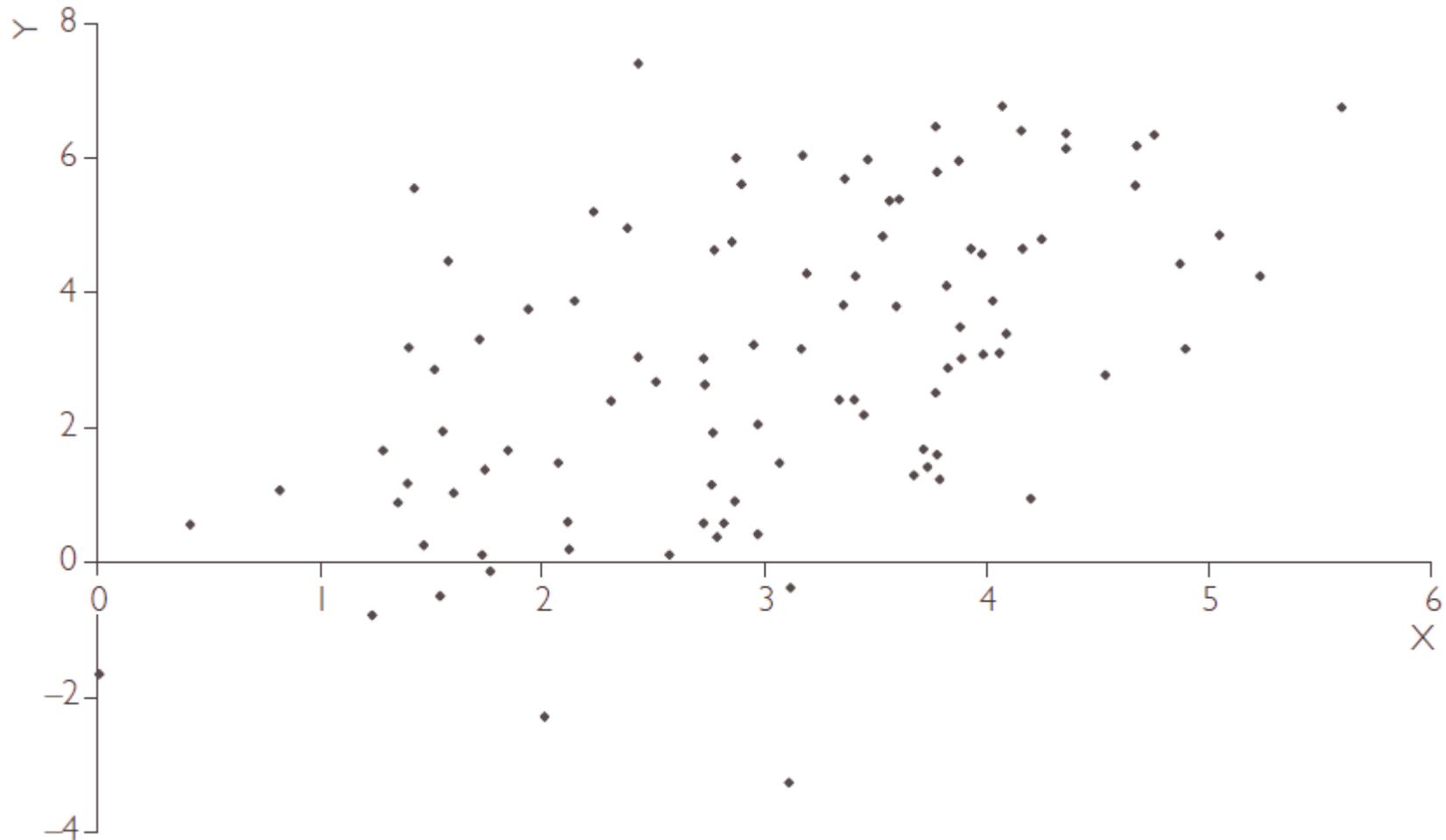
Large sample size, small error variance
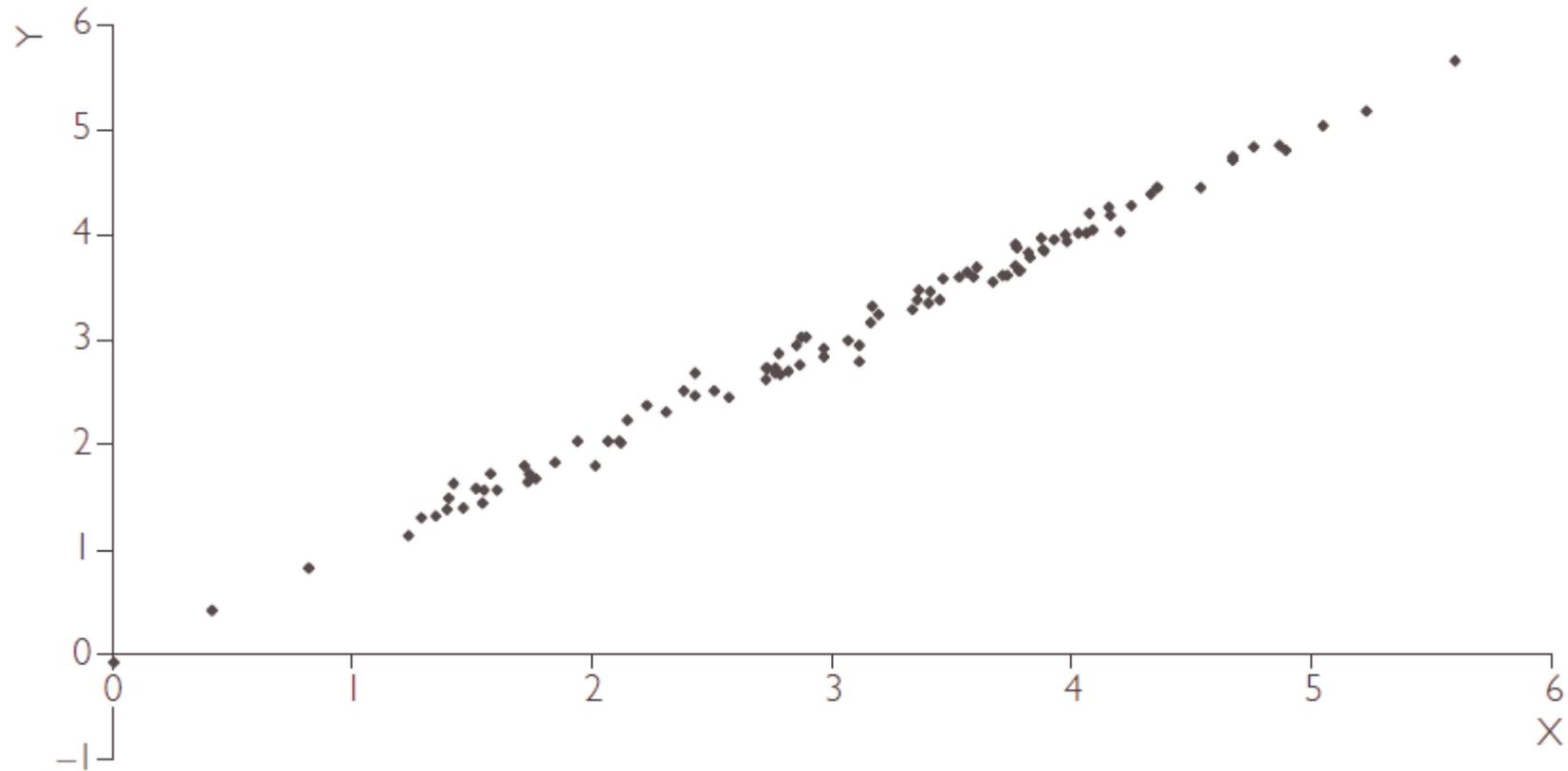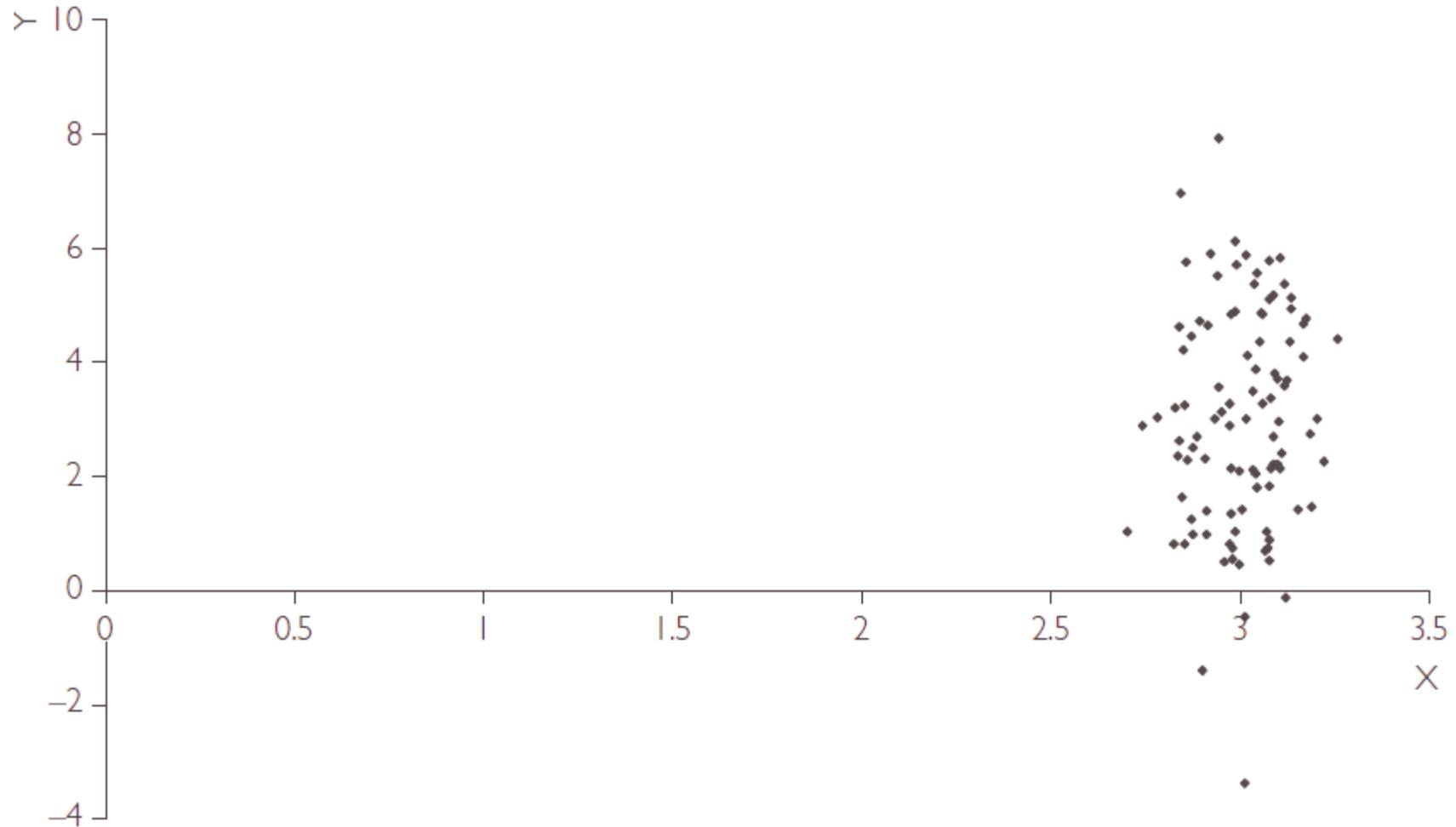
- Confidence interval for the estimated beta coefficient:

$$\left[\hat{\beta} - t_b s_b , \hat{\beta} + t_b s_b \right]$$

- where

  - s_b: standard deviaton of βhat

  - t_b: Students' t-distribution

- Larger confidence level → larger t_b

- More observations → smaller s_b

- Most common: 95% confidence interval
  - „There is 95% chance that the true value of the coefficient lies in the given interval"
- Large N, 95%: t=1.96
- Table of t-distribution

- Excel: confidence level can be chosen

	Coefficients	Standard deviation	Bottom 95%	Top 95%
Intercept	0,6000	0,1123	0,3758	0,8241
X variable	0,0008	0,0001	0,0006	0,0011

- Examples:
  - Does advertising influence the sales?
  - Does education level influence the earnings?
- Null hypothesis vs. Alternative hypothesis

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

- Two-sided hypothesis
- Significance of intercept can also be tested – interpretation?

- Relationship to confidence interval

  - Does it include the zero?

- Significance level: 100%-confidence level

  - „Probability of mistake"

- t-test:

$$t = \frac{\hat{\beta}}{s_b}$$

- „Large" absolute value – significantly different from zero

- P-value: „probability that the coefficient equals zero"

- Hypothesis to test

- Statistical test

- Decision

- Regression table of Excel: t-value, P-value presented

- If P-value<5%: $\beta=0$ hypothesis is rejected at 5% significance level

- If P-value<1%: $\beta=0$ hypothesis is rejected at 1% significance level

	Coeff	Standard deviation	t-stat	p-value	Bottom 95%	Top 95%
Intercept	11,543	0,133	86,905	0,000	11,281	11,804
X-variable	0,013	0,003	4,768	0,000	0,008	0,0018

- Estimated coefficient and its standard deviation

- t-statistic (large – significant)

- P-value (small – significant)

- Confidence interval (does it include zero?)

- The ADVERT.XLS file contains data on the sales and advertising expenditures of 84 companies. Set up and run a regression using this data and discuss your results verbally as you would in a report.

- Include a discussion of the marginal effect of advertising on sales and a discussion of whether this marginal effect is statistically significant.

- Test R^2=0 hypothesis
- Does the regression have explanatory power?
  - Simple regression: equivalent to testing β=0
- F-test:

$$F = \frac{(N-2)R^2}{1-R^2}$$

- Accept or reject null hypothesis based on the P-value („significance of F")

Regression statistics						
r-squared	0.09					
ANALYSIS OF VARIANCE						
	df	SS	MS	F	F significance	
Regression	1	4.50	4.50	22.74	3.27E-06	
Residual	234	46.32	0.20			
Total	235	50.82				
	Coeff.	Standard dev.	t-stat.	p-value	Bottom 95%	Top 95%
Intercept	11.54	0.13	86.91	0.00	11.28	11.80
X variable	0.01	0.00	4.77	0.00	0.01	0.02

- OLS estimation of the multiple regression model

- Statistical aspects of the multiple regression

- Interpreting OLS estimates

- Omitted variables bias

- Multicollinearity

- Dummy variables

.

- County level unemployment: number of enterprises, geographical location, average education level, ...

- Sales: advertisement expenditures, hours worked, quality of the product, ...

- Real estate prices: lot size, number of rooms, location, ...

- Regression with k regressors:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + e_i$$

$$SSR = \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - ... - \hat{\beta}_k X_{ik})^2$$

- OLS: minimal sum of squared residuals
- Interpretation of the coefficients:
  - Marginal effect
  - Other explanatory variables held constant
  - „Ceteris paribus"

- Confidence interval: analogously to univariate case

- Significance of coefficients: t-test, p-value

- R^2 = 1-SSR/TSS

  - Measure of fit

  - What % of the variation of the dependent variable is explained by the explanatory variables

  - Testing R^2=0: F-test

$$F = \frac{(N-k-1)R^2}{1-R^2}$$

Regression statistics						
r-squared	0.46					
ANALYSIS OF VARIANCE						
	df	*SS*	*MS*	*F*	*F sign.*	
Regression	2	2.87E+13	1.44E+13	2170.7	0	
Residual	4997	3.30E+13	6.61E+09			
Total	4999	6.18E+13				
	Coeff.	*St. dev.*	*t stat.*	*p-value*	*Bottom 95%*	*Top 95%*
Intercept	-328321.34	8040.13	-40.84	0.00	-344083.52	-312559.16
Education	27250.22	452.97	60.16	0.00	26362.20	28138.24
Age	3171.29	109.05	29.08	0.00	2957.52	3385.07

Regression statistics					
r-squared	0.54				
ANALYSIS OF VARIANCE					
	df	*SS*	*MS*	*F*	*F sign.*
Regression	4	2.08E+11	5.2E+10	155.95	0.00
Residual	541	1.80E+11	3.34E+08		
Total	545	3.89E+11			
	Coeff.	*St. dev.*	*t stat.*	*p-value*	*Bottom 95%*
Intercept	-4009.55	3603.11	-1.11	0.27	-11087.35
Lot size	5.43	0.37	14.70	0.00	4.70
#bedrooms	2824.61	1214.81	2.33	0.02	438.30
#bathrooms	17105.17	1734.43	9.86	0.00	13698.12
#stories	7634.90	1007.97	7.57	0.00	5654.87

- Example: housing prices
- Multivariate:

$$\hat{P} = -4009.6 + 5.4\text{lot} + 2824.6\text{bedroom} +$$
$$+ 17105.2\text{bathroom} + 7634.9\text{stories}$$

- Univariate:

$$\hat{P} = 32794.0 + 27477.0\text{bathroom}$$

- Larger estimated coefficient

- Explanation for different coefficients:

- Influence of several factors

- Correlation with the number of bathrooms

  - E.g. positive correlation between lot size – number of bathroom

- Univariate regression: cannot separate the effects

- Production costs depend upon the volume of output and the prices of inputs aswell. Estimate the cost function of electric firms using data from ELECTRIC.XLS.

- The cost of production depends upon output, price of labor, price of capital and price of fuel.

- Interpret the estimated coefficients and the goodness of fit.

- Bias due to omitted variables:

  - Estimation is not correct if we omit such a variable which is correlated with the included explanatory variables

- Include those variables which have explanatory power!

- But: redundant variables – estimation precision decreases

  - General practice: omit the insignificant ones

- Simple:

	Coeff.	Standard dev.	t stat.	P-value
Intercept	-161796.32	9514.04	-17.01	0.00
Education	24855.33	707.51	35.13	0.00

- Multiple, corr(educ.,age)=-0.04

	Coeff.	Standard dev.	t stat.	P-value
Intercept	-328321.34	8040.13	-40.84	0.00
Education	27250.22	452.97	60.16	0.00
Age	3171.29	109.05	29.08	0.00

- Some of the explanatory variables are strongly correlated

- The effects of the regressors are difficult to separate

- Solution: omit some of the regressors – not always desirable!

- „Symptoms":

  - Low t-, high P-values

  - At the same time, R-squared is high

  - Coefficients are very sensitive to the inclusion of additional (collinear) variables

  - Estimated coefficients are very different from the expected values (clearly unreasonable coefficients)

- Earnings regressions, corr(age, experience)=0.97

r-squared	0.468			
	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	-1.7E+11	3.05E+10	-5.647	1.72E-08
Education	-2.9E+10	5.08E+09	-5.647	1.72E-08
Age	2.87E+10	5.08E+09	5.647	1.72E-08
Experience	-2.9E+10	5.08E+09	-5.647	1.72E-08

r-squared	0.465			
	Coeff.	*Standard dev.*	*t stat.*	*P-value*
Intercept	-328321	8040.126	-40.835	0
Education	27250.22	452.9723	60.159	0
Age	3171.293	109.0451	29.082	6.3E-172

- Qualitative, coding: 0 – 1

- Binary = dummy = dichotomous variable

- Examples:

  - Housing prices: is there garage, air conditioning, …

  - Wages: male – female

  - Medical expenditures: if insured or not

  Etc.

- OLS method unchanged, different interpretation of coefficients
- Simple regression:

$$Y = \alpha + \beta D + e$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta} D$$

$$\hat{Y} = \hat{\alpha}, \text{ if } D = 0$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}, \text{ if } D = 1$$

- Mean of two subgroups

1. Housing prices

$$\hat{P} = 59\,885 + 25\,996\,Cond$$

- Mean price with air conditioning: 85 881 CAD

2. Earnings (Wage tariff 2003 subsample)

$$\hat{W} = 159\,289 + 66\,854\,male$$

- Average earnings, males: 226 142 Ft
- Average earnings, females: 159 289 Ft

$$Y_i = \alpha + \beta_1 D_{i1} + ... + \beta_k D_{ik} + e_i$$

- Number of groups: 2^k

- Group means: sum of respective coefficients

- Interpretation of coefficients: partial effect

- Only binary: different means

- Binary and not binary: different intercept

- Simplest model:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + e_i$$

$$\text{Intercept}: \alpha \text{ or } \alpha + \beta_1$$

	Coeff.	Standard dev.	t stat.	P-value
Intercept	30555.75	2289.991	13.34317	2.59E-35
Air cond.	19268.8	1909.658	10.09018	4.72E-22
Recreation room	7395.032	2462.386	3.003198	0.002795
Basement	6187.162	1945.687	3.179937	0.001557
Lot size	5.433193	0.410367	13.23985	7.35E-35

- The effects of two variables are interdependent – multiplication

- Interaction between continuous X and binary D:
  - Different effect of X in two subgroups
  - Different slope and different intercept:

$$Y = \alpha + \beta_1 X + \beta_2 D + \beta_3 (DX) + e$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X \text{ if } D = 0$$

$$\hat{Y} = \left(\hat{\alpha} + \hat{\beta}_2\right) + \left(\hat{\beta}_1 + \hat{\beta}_3\right) X \text{ if } D = 1$$

4 possibilities:

- Same regression line for all subgroups

- Different intercept

- Different slope

- Intercept and slope are both different

- Housing prices (hprice.xls)

- Dependent variable: price (CAD)

- Regressors: air conditioning, recreation room, basement, lot size

- Interaction between recreation room and basement

- Interpretation of the coefficient of the interaction term?

	Coefficient	Standard dev.	t statistic	p-value
Intercept	33707.34	2575.85	13.09	0.00
Air cond.	5990.53	5412.63	1.11	0.27
Recr. room	7584.15	2450.21	3.10	0.00
Basement	6179.81	1935.23	3.19	0.00
Lot size	4.77	0.48	9.96	0.00
Recr. room*basement	2.38	0.91	2.62	0.01

- Excel file, WAGEDISC.XLS contains data on N = 100 employees in a particular occupation. Suppose that interest centers on investigating the factors which explain salary differences in this occupation with a view to addressing the issue of sex discrimination in this occupation. The data set contains the following variables:

- Y = salary (measured in thousands of dollars)

- X1 = education level (measured in years of schooling)

- X2 = experience level (measured in years of employment)

- D = dummy for sex (= 1 for male, = 0 for female).

- Calculate and discuss descriptive statistics for this data set.

- Calculate the mean salary for female employees and male employees separately.

- Run a simple regression of Y on D. Is the slope coefficient in this regression statistically significant? Compare your regression result with your finding in (b). Can you use these findings to conclude that women are discriminated against in this occupation?

- Run a multiple regression of Y on X1, X2 and D. Write a short report outlining your findings and addressing the issue of wage discrimination in this occupation. Are your results statistically significant?

- Compare your results in part (d) to part (c). Why do they differ? Hint: Calculate a correlation matrix for all the explanatory variables and think intuitively about what the correlations mean.

- Construct a new variable Z = DX2 and run a regression of Y on X1, X2, D and Z. Is Z statistically significant? How would the short report you wrote in part (d) change? Explain verbally what the coefficient on Z measures.

- Example:
  - Having own car?
  - Company profitable?

  Etc.

- OLS estimation feasible, but:
  - Estimated value is not 0-1!
  - Further problems – different estimation method is preferable

- Pooling Cross Sections

- Panel Data Methods

.

- panel data structure gives the following equation:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + u_{it}$$

- Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension

- More precisely it's only data following the same cross-section units over time

- Otherwise it's a pooled cross-section

- We may want to pool cross sections just to get bigger sample sizes

- We may want to pool cross sections to investigate the effect of time

- We may want to pool cross sections to investigate whether relationships have changed over time

- After controlling for education, has the pattern of fertility among women over age 35 changed between 1972 and 1984?

- answer: using multiple regression analysis with dummy variables

- Open FERTIL1.GDT in Gretl.

- Regress kids on educ and other explanatory variables, including dummy variables for year

- Changes in the return to education and the gender wage gap

- Regress the log(wage) on gender, other explanatory variables and year dummies using CPS78_85.GDT

- Which estimated coefficient is to show the change in the relationship between the return to education and the gender?

- interaction term: multiplication of the year dummy and the gender dummy

- How did the wage gap changed?

- Say random assignment to treatment and control groups, like in a medical experiment

- One can then simply compare the change in outcomes across the treatment and control groups to estimate the treatment effect

- For time 1,2, groups A, B $(y_{2,B} - y_{2,A}) - (y_{1,B} - y_{1,A})$, or equivalently $(y_{2,B} - y_{1,B}) - (y_{2,A} - y_{1,A})$, is the difference-in-differences

- A regression framework using time and treatment dummy variables can calculate this difference-in-difference as well

- Consider the model: $y_{it} = \beta_0 + \beta_1 treatment_{it} + \beta_2 after_{it} + \beta_3 treatment_{it}*after_{it} + u_{it}$

- The estimated β_3 will be the difference-in-differences in the group means

- When don't truly have random assignment, the regression form becomes very useful

- Additional x's can be added to the regression to control for differences across the treatment and control groups

- Sometimes referred to as a "natural experiment" especially when a policy change is being analyzed

- Effect of a Garbage Incinerator's Location on Housing Prices

- Open KIELMC.GDT

- naive analyst would regress:

$$rprice = beta0 + beta1 * nearinc + u$$

- Estimate this equation on 1981 data

- How to interpret the results? Does the ininerator *lower* the housing prices?

- Repeat the regression with 1978 data!

- Draw the conclusion!

- It's possible to use a panel just like pooled cross-sections, but can do more than that

- Panel data can be used to address some kinds of omitted variable bias

- If can think of the omitted variables as being fixed over time, then can model as having a composite error

- Suppose the population model is $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}$

- Here we have added a time-constant component to the error, $v_{it} = a_i + u_{it}$

- If a_i is correlated with the x's, OLS will be biased, since we a_i is part of the error term

- With panel data, we can difference-out the unobserved fixed effect

- We can subtract one period from the other, to obtain $\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + ... + \beta_k \Delta x_{ik} + \Delta u_i$

- This model has no correlation between the x's and the error term, so no bias

- Need to be careful about organization of the data to be sure compute correct change

- Is there relationship between city crime rates and unemployment rates?

- Open CRIME2.GDT

- Regress crmrte on unem! How to interpret the result?

- Repeat the regression using pooled OLS! How much did it improved the regression results?

- Suppose that unmeasured city factors, ai, that affect crime rate to also be correlated with the unemployment rate

- differencing the variables the unobserved variable drops out

- Then we have the fix effect estimator!

- Create the dataset in panel format (CRIME2P.GDT) and run the regression.

- Panel data set are very useful for poliy analysis, and, in particular, program evaluation.

- Simplest program evaluation setup: sample of units (individuals, firms, cities, etc.). Some of them take part in a program, become „treated". The rest forms the control group.

- Suppose, we wish to evaluate the effect of Michigen job training program on worker productivity of manufacturing firms.

- See JTRAIN.GDT file.

- Regress scrap on grant using fixed effect panel regression (that is, regression should be made on differences)

- Regress log(scrap) on grant using similar method.

- How to interpret the coefficients of the „grant" variable?

- Can extend this method to more periods

- Simply difference adjacent periods

- So if 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual

- Simply estimate by OLS, assuming the Δu_{it} are uncorrelated over time

Thank you for your attention!

AAM CONSULTING

Bucharest   30.10.2011

AAM Management Information Consulting Ltd.   www.aam.hu