

SOS3003
**Applied data analysis for social
science**
Lecture note 01-2009

Erling Berge
Department of sociology and political science
NTNU

Fall 2009

© Erling Berge 2009

1

History

- In the history of civilization there are 2 unrivalled accelerators:
 - The invention of writing about 5-6000 years ago
 - The invention of the scientific method for separating facts from fantasy about 5-600 years ago
- There is no topic more important to learn than the basics of the scientific method
- That does not mean that it is not – at times – rather boring

Fall 2009

© Erling Berge 2009

2

Basics of causal beliefs

- First: doubt what you believe is a causal link until you can give good valid reasons justifying your belief
- Second: there are many types of good valid reasons for believing in a particular causal link
 - For example if the overwhelming majority of certified scientists says that human activities contribute to global warming, then we are justified believing that by changing our activities we could contribute less to global warming
- Third: random conjunctures (“correlation”) are not good valid reasons for believing in a causal link

Fall 2009

© Erling Berge 2009

3

Causal correlations

- This class will focus on how to distinguish between random conjunctures and that which might be a valid causal correlation
- That which might be a valid causal correlation will need a *causal mechanism* explaining how the cause can produce the effect before we have a valid reason to believe in the causal link

Fall 2009

© Erling Berge 2009

4

Causal mechanism

- Elster 2007 *Explaining Social Behaviour*:
- "mechanisms are frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences" (page 36)
- Also sometimes limited to "causal chains"

Fall 2009

© Erling Berge 2009

5

Primacy of theory

- To say it more bluntly: If you do not have a believable theory (and this may well start as a fantasy) then regression techniques will tell you nothing even if you find a seemingly non-random correlation
- But without a valid and believable regression analysis any believable fantasy will remain just that: a fantasy (assuming you cannot find other valid empirical verification)

Fall 2009

© Erling Berge 2009

6

Preliminaries

- Prerequisite: SOS1002 or equivalent
- Goal: to read critically research articles from our field of interest
- Required reading
- Term paper: this is part of the examination and evaluation procedure

Fall 2009

© Erling Berge 2009

7

Goals for the class

- The goal is that each of you shall be able to read critically research articles discussing quantitative data. This means
 - You are to know the pitfalls
- You are to learn how to perform straightforward analyses of co-variation in "quantitative" and "qualitative" data (nominal scale data in regression analysis), and in particular:
 - Also here you have to demonstrate that you know the pitfalls

Fall 2009

© Erling Berge 2009

8

Required reading SOS3003

- Hamilton, Lawrence C. 1992. *Regression with graphics*. Belmont: Duxbury. Ch 1-8
- Hamilton, Lawrence C. 2008. *A Low-Tech Guide to Causal Modelling*.
<http://pubpages.unh.edu/~lch/causal2.pdf>
- Allison, Paul D. 2002. *Missing Data*. Sage University Paper: QASS 136. London: Sage.

Fall 2009

© Erling Berge 2009

9

Term paper

- **Deadline for paper: 23 November 12:00**
- The term paper shall be an independent work demonstrating how multiple regression can be used to analyze a social science problem. The paper should be written as a journal article, but with more detailed documentation of data and analysis, for example by means of appendices.
- Based on information about the dependent variable a short theoretical discussion of possible causal mechanisms explaining some of the variation in the dependent variable is presented. This leads up to a model formulation and operationalisation of possible causal variables taken from the data set. If missing data on one or more variables causes one or more cases to be dropped from the analysis, the selection problem must be discussed.
- By means of multiple regression (OLS or Logistic) the model should be estimated and the results discussed in relation to the initial theoretical discussion

More details will be available in a separate paper

10

Lecture I

Basics of what you are assumed to know

- The following is basically repeating known stuff
- Variable distributions
 - Ringdal Ch 12 p251-270
 - Hamilton Ch 1 p1-23
- Bivariat regression
 - Ringdal Ch 17-18 p361-387
 - Hamilton Ch 2 p29-59

Fall 2009

© Erling Berge 2009

11

Some basic concepts

- Cause
- Model
- Population
- Sample
- Variable: level of measurement
- Variable: measure of centralization
- Variable: measure of dispersion

Fall 2009

© Erling Berge 2009

12

Data analysis

- Descriptive use of data
 - Developing classifications
- Analytical use of data
 - Describe phenomena that cannot be observed directly (inference)
 - Causal links between directly eller indirectly observable phenomena (theory or model development)

Fall 2009

© Erling Berge 2009

13

Causal analysis: from co-variation to causal connection

- From colloquial speech to theory
 - Fantasy and intuition, established science tradition
- From theory to model
 - Operationalisation
- From observation to generalisation
 - Causal analysis

Fall 2009

© Erling Berge 2009

14

THREE BASIC DIVISIONS

<u>Observed</u>		<u>Real interest</u>
THEORY/ MODEL	-	REALITY
SAMPLE	-	POPULATION
CO-VARIATION	-	CAUSE

On the one hand we have what we are able to observe and record, on the other hand, we have what we would like to discuss and know more about

Basic sources of error

- Errors in theory / model
 - Model specification: valid tests require a correct (true) model
- Errors in the sample
 - Selection bias
- Measurement problems
 - Missing cases and measurement errors
 - Validity og reliability
- Multiple comparisons
 - Conclusions are valid only for the sample

From population to sample

- POPULATION (all units)

Simple random sampling

- SAMPLE (selected units)

Fall 2009

© Erling Berge 2009

17

Unit and variable

- A unit, as a carrier of data, will be contextually defined
 - SUPER - UNIT: e.g. the local community
 - UNIT: e.g. household
 - SUB - UNIT: e.g. person
- Variable: empirical concept used to characterize units under investigation. Each unit is characterized by being given a variable value

Fall 2009

© Erling Berge 2009

18

Data matrix and level of measurement

- Matrix defined by Units * Variables
 - A table presenting the characteristics of all investigated units ordered so that all variable values are listed in the same sequence for all units
- Level of measurement for a variable
 - Nominal scale *classification
 - Ordinal scale *classification and rank
 - Interval scale *classification, rank and distance
 - Ratio scale *classification, rank, distance and absolute zero

Fall 2009

© Erling Berge 2009

19

Variable analysis

- Description
 - Central tendency and dispersion
 - Form of distribution
 - Frequency distributions and histograms
- Comparing distributions
 - Quantile plots
 - Box plots

Fall 2009

© Erling Berge 2009

20

VARIABLE: central tendency

- **Mean**
sum of all values of the variable for all units divided by the number of units
- **MEDIAN**
The variable value in an ordered distribution that has half the units on each side
- **MODUS**
The typical value. The value in a distribution that has the highest frequency

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Fall 2009

© Erling Berge 2009

21

VARIABLE: measure of dispersion I

- **MODAL PERCENTAGE**
The percentage of units with value like the mode
- **RANGE OF VARIATION**
The difference between highest and lowest value in an ordered distribution
- **QUARTILE DIFFERENCE**
Range of variation of the 50% of units closest to the median ($Q_3 - Q_1$)
- **MAD - Median Absolute Deviation**
Median of the absolute value of the difference between median and observed value:
– $MAD(x_i) = \text{median } |x_i - \text{median}(x_i)|$

Fall 2009

© Erling Berge 2009

22

VARIABLE: measure of dispersion II

- **STANDARD DEVIATION**
 - Square root of mean squared deviation from the mean
 - $s_y = \sqrt{[(\sum_i(Y_i - \tilde{Y})^2)/(n - 1)]}$
 - **MEAN DEVIATION**
 - Mean of the absolute value of the deviation from the mean
 - **VARIANCE**
 - Standard deviation squared:
 - $s_y^2 = (\sum_i(Y_i - \tilde{Y})^2)/(n - 1)$
- (ps: here \tilde{Y} is the mean of Y)

Fall 2009

© Erling Berge 2009

23

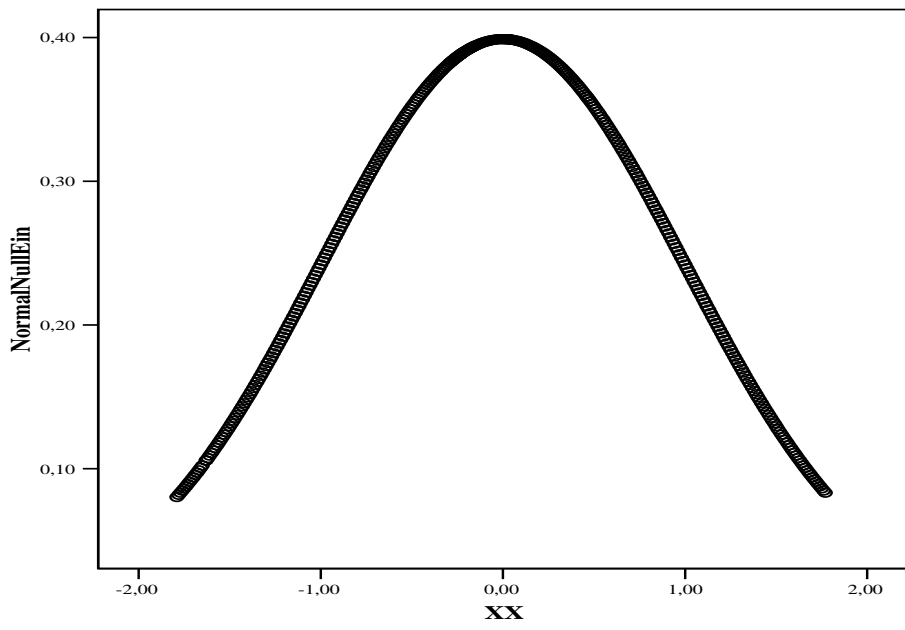
Variable: form of distribution I

- Symmetrical distributions
- Skewed distributions
 - "Heavy" and "Light" tails
- Normal distributions
 - Are not "normal"
 - Are unambiguously determined by mean and variance (μ og σ^2)

Fall 2009

© Erling Berge 2009

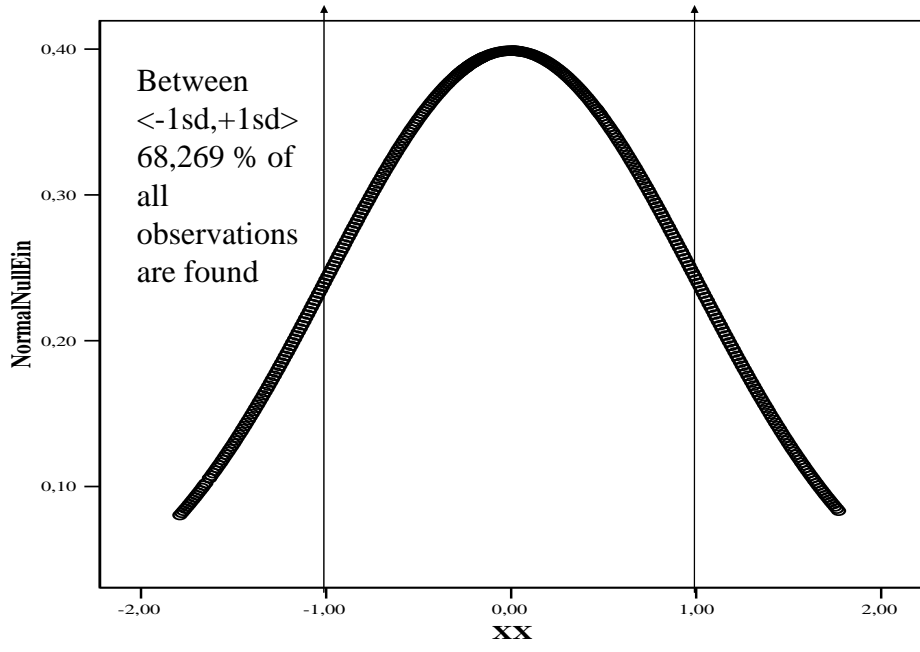
24



Fall 2009

© Erling Berge 2009

25



Fall 2009

© Erling Berge 2009

26

Skewed distributions

- Positively skewed has $\tilde{Y} > Md$
- Negatively skewed has $\tilde{Y} < Md$
- Symmetric distributions has $\tilde{Y} \approx Md$

- Ps: $\tilde{Y} = \text{mean of } Y$

Fall 2009

© Erling Berge 2009

27

Symmetric distributions

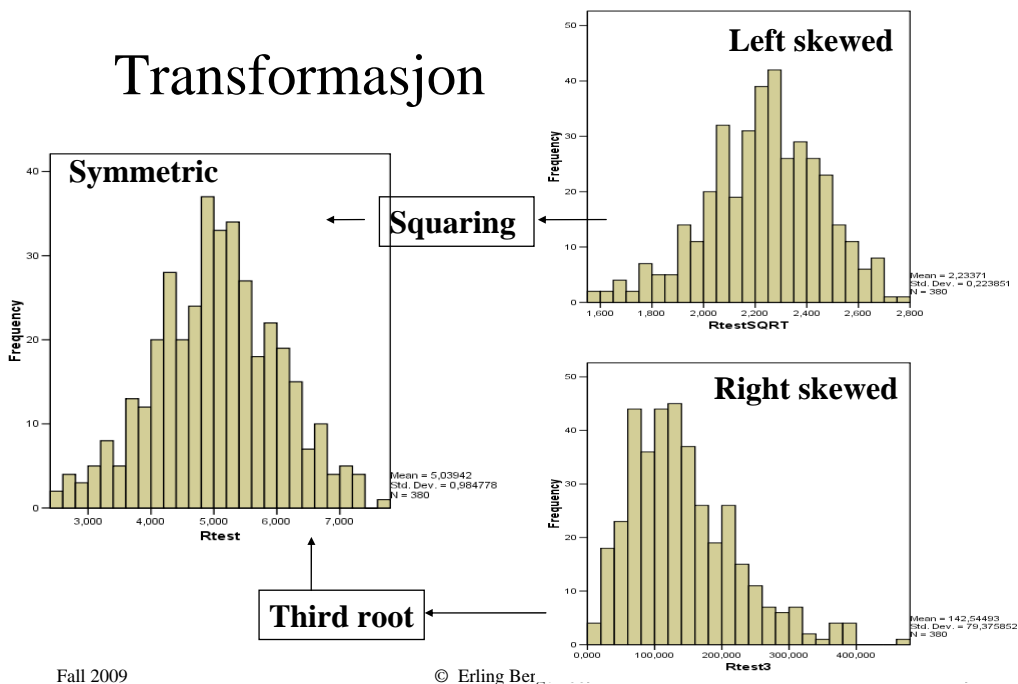
- Median and IQR are resistant against the impact of extreme values
- Mean and standard deviation are not
- In the normal distribution (ND) $s_y \approx \text{IQR}/1.35$
- If we in a symmetric distribution find
 - $s_y > \text{IQR}/1.35$ then the tails are heavier than in the ND
 - $s_y < \text{IQR}/1.35$ then the tails are lighter than in the ND
 - $s_y \approx \text{IQR}/1.35$ then the tails are about similar to the ND

Fall 2009

© Erling Berge 2009

28

Transformasjon



Variable: analyzing distributions I

- Boxplot
 - The box is constructed based on the quartile values Q_1 og Q_3 . Observations within $< Q_1, Q_3 >$ are in the box-
 - Adjacent large values are defined as those outside the box but inside $Q_3 + 1.5 \cdot IQR$ or $Q_1 - 1.5 \cdot IQR$
 - Outliers (seriously extreme values) are those outside of $Q_3 + 1.5 \cdot IQR$ or $Q_1 - 1.5 \cdot IQR$

Variables: analyzing distributions II

- Quantiles is a generalisation of quartiles and percentiles
- Quantile values are variable values that correspond to particular fractions of the total sample or observed data, e.g.
 - Median is 0.5 quantile (or 50% percentile)
 - Lower quartile is 0.25 quantile
 - 10% percentile is 0.1 quantile ...

Fall 2009

© Erling Berge 2009

31

Variables: analyzing distributions III

- Quantile plots
 - Quantile values against value of variable
 - The Lorentz curve is a special case of this (it gives us the Gini-index)
- Quantile-Normal plot
 - Plot of quantil values on one vairable against quantil values of Normal distribution with the same mean and standard deviation

Fall 2009

© Erling Berge 2009

32

Example: Randaberg 1985

- Questionnaire: (the number of decares land you own / 10 da = 1 ha)

Q: ANTALL DEKAR GRUNN DU

eier: _____

(Number of decares you own: ____)

Fall 2009

© Erling Berge 2009

33

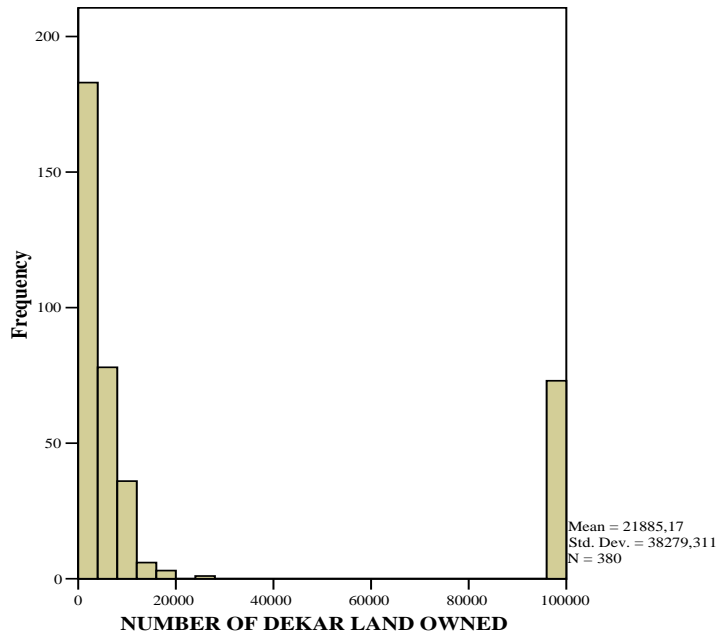
NUMBER OF DEKARE LAND OWNED

	NUMBER OF DEKARE LAND OWNED	Valid N (listwise)
N	380	380
Minimum	0	
Maximum	99900	
Mean	21885.17	
Std. Deviation	38279.311	

Fall 2009

© Erling Berge 2009

34



Fall 2009

© Erling Berge 2009

35

XAreaOwned (NUMBER OF DEKARE LAND OWNED)

	XAreaOwned	Valid N (listwise)
N	307	307
Minimum	.00	
Maximum	25000.00	
Mean	3334.4104	
Std. Deviation	4201.54943	

Fall 2009

© Erling Berge 2009

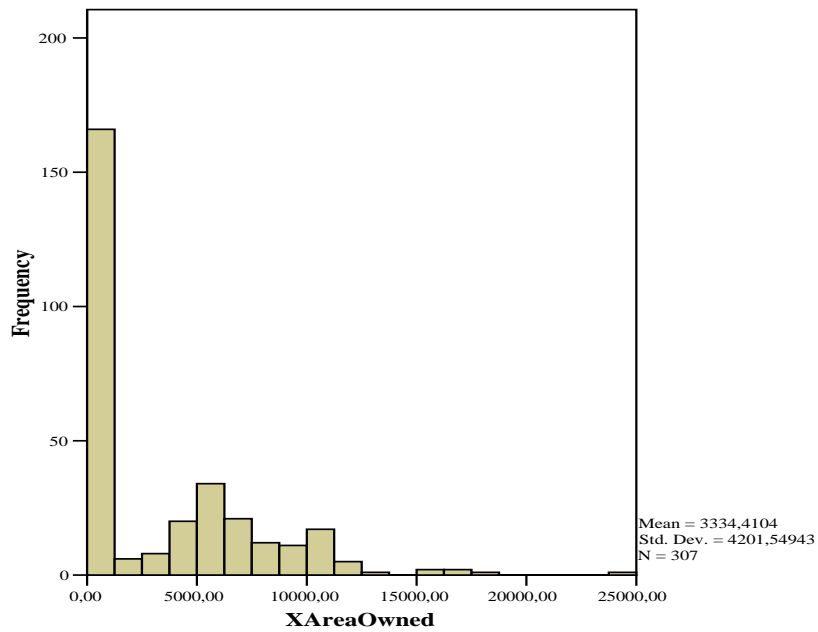
36

		XAreaOwned	Valid N (listwise)
N	Statistic	307	307
Range	Statistic	25000.00	
Minimum	Statistic	.00	
Maximum	Statistic	25000.00	
Sum	Statistic	1023664.00	
Mean	Statistic	3334.4104	
	Std. Error	239.79509	
Std. Deviation	Statistic	4201.54943	
Variance	Statistic	17653017.596	
Skewness	Statistic	1.352	
	Std. Error	.139	
Kurtosis	Statistic	2.194	
	Std. Error	.277	

Fall 2009

© Erling Berge 2009

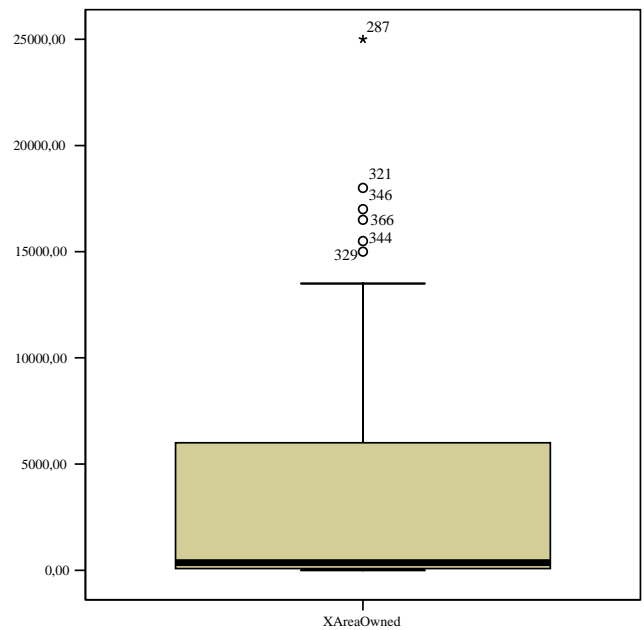
37



Fall 2009

© Erling Berge 2009

38



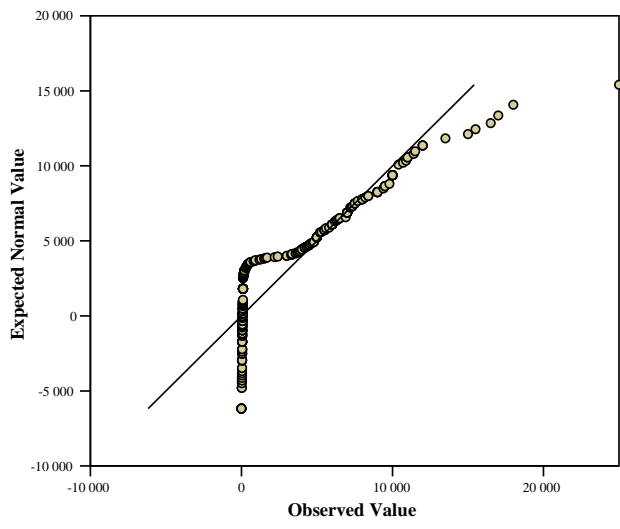
Fall 2009

© Erling Berge 2009

39

Normal Q-Q Plot of XAreaOwned

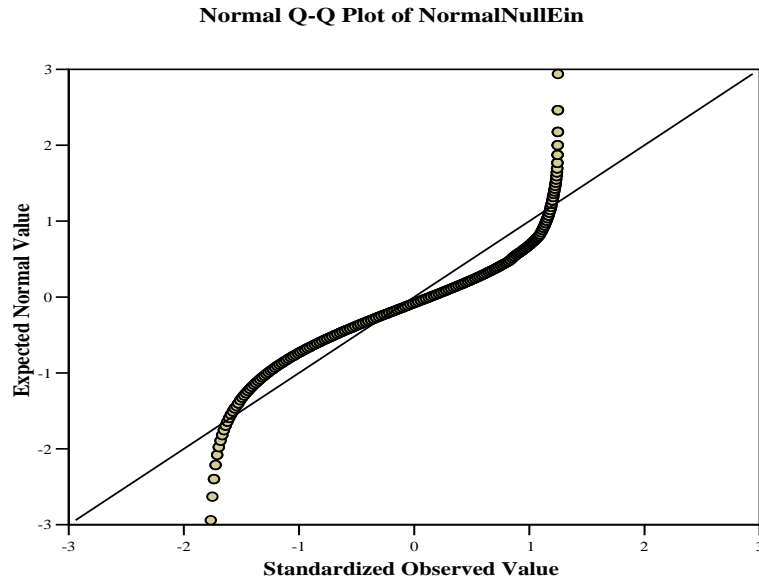
NB
Figures
from SPSS
are mirrors
of figures
in
Hamilton



Fall 2009

© Erling Berge 2009

40



Fall 2009

© Erling Berge 2009

41

Questionnaire:

- **Hvor viktig er det at myndighetene kontrollerer og regulerer bruken av arealer gjennom for eksempel kontroll av**
 - av tomtetildelinger (kommunal formidl.)

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---
 - avkjørsler fra hus til vei

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---
 - kjøp og salg av landbrukseiendommer

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---

Fall 2009

© Erling Berge 2009

42

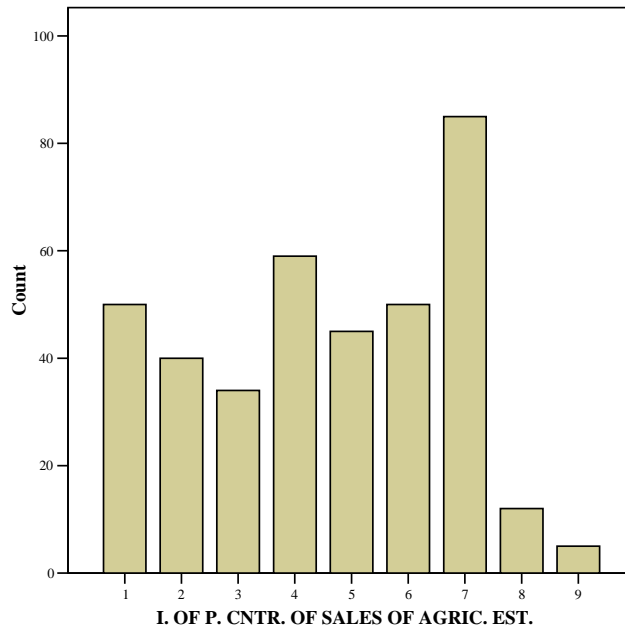
Importance of public control of sales of agric. estates

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1	50	13.2	13.2	13.2
2	40	10.5	10.5	23.7
3	34	8.9	8.9	32.6
4	59	15.5	15.5	48.2
5	45	11.8	11.8	60.0
6	50	13.2	13.2	73.2
7	85	22.4	22.4	95.5
8	12	3.2	3.2	98.7
9	5	1.3	1.3	100.0
Total	380	100.0	100.0	

Fall 2009

© Erling Berge 2009

43



Fall 2009

© Erling Berge 2009

44

Questionnaire: coding

Ved utfylling: sett ring rundt et tall som synes å gi passelig uttrykk for viktigheten når 1 betyr svært lite viktig og 7 særdeles viktig, eller sett et kryss inne i parentes () som står bak svaret du velger

På noen spørsmål kan du krysse av flere svar

	lykkes dårlig/ lite viktig							lykkes godt/ svært viktig	vet ikke
Kodeverdi	1	2	3	4	5	6	7		8

Dei som ikkje kryssar av noko svar vert koda 9 (ie. missing)

Fall 2009

© Erling Berge 2009

45

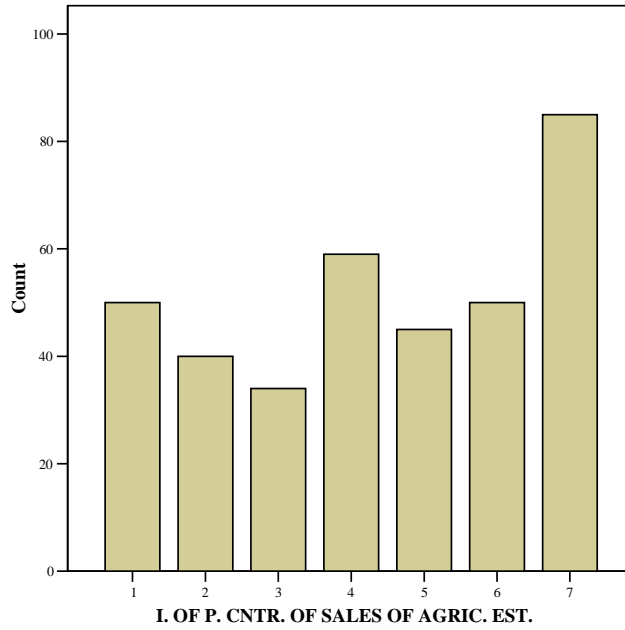
I. OF P. CNTR. OF SALES OF AGRIC. EST.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	50	13.2	13.8	13.8
	2	40	10.5	11.0	24.8
	3	34	8.9	9.4	34.2
	4	59	15.5	16.3	50.4
	5	45	11.8	12.4	62.8
	6	50	13.2	13.8	76.6
	7	85	22.4	23.4	100.0
	Total	363	95.5	100.0	
Missing	8	12	3.2		
	9	5	1.3		
	Total	17	4.5		
Total		380	100.0		

Fall 2009

© Erling Berge 2009

46



Fall 2009

© Erling Berge 2009

47

		I. OF P. CNTR. OF SALES OF AGRIC. EST.	YControlSalesAgricEstate Valid N (listwise)
N	Statistic	380	363
Range	Statistic	8	6.00
Minimum	Statistic	1	1.00
Maximum	Statistic	9	7.00
Sum	Statistic	1729	1588.00
Mean	Statistic	4.55	4.3747
	Std. Error	.114	.11045
Std. Deviation	Statistic	2.213	2.10435
Variance	Statistic	4.897	4.428
Skewness	Statistic	-.171	-.234
	Std. Error	.125	.128
Kurtosis	Statistic	-1.148	-1.267
	Std. Error	.250	.255

Fall 2009

© Erling Berge 2009

48

Distributions with or without missing?

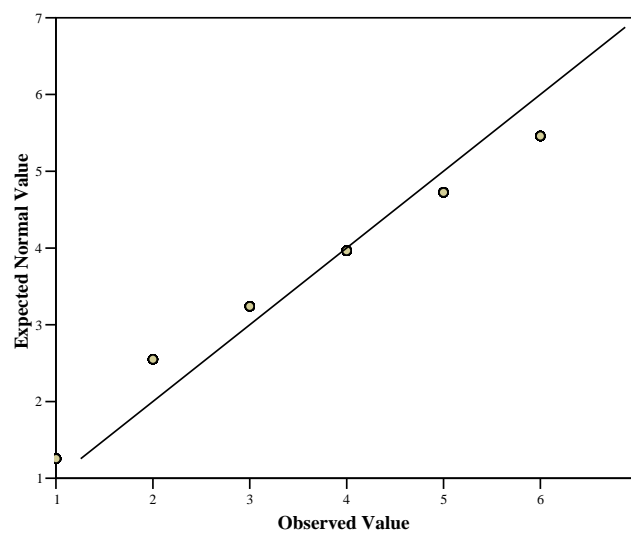
- What difference do the 17 missing observations make in the
 - Quantile-Normal plot?
 - Box plot?

Fall 2009

© Erling Berge 2009

49

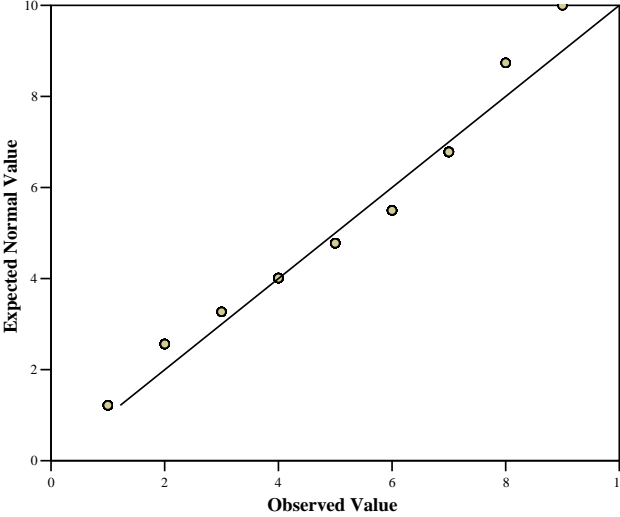
Normal Q-Q Plot of I. OF P. CNTR. OF SALES OF AGRIC. EST.



Fall 2009

50

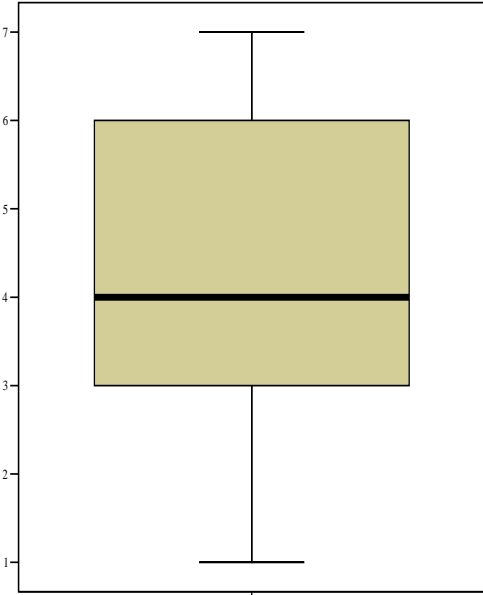
Normal Q-Q Plot of I. OF P. CNTR. OF SALES OF AGRIC. EST.



Fall 2009

© Erling Berge 2009

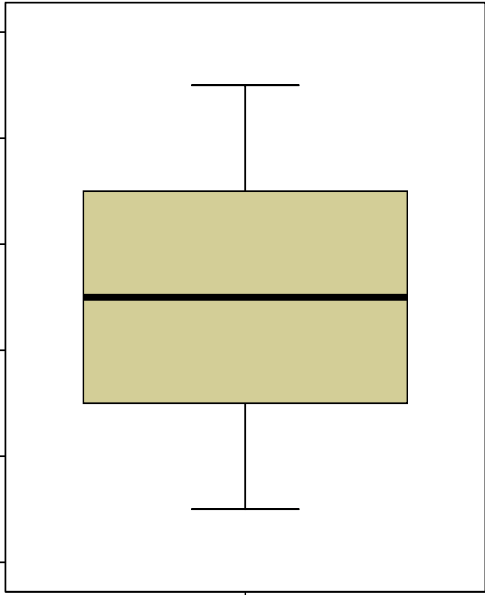
51



I. OF P. CNTR. OF SALES OF AGRIC. EST.

Fall 2009

© Erling Berge 2009



YControlSalesAgricEstatesORD

52

Data collection and data quality

- Questions – techniques for asking questions will not be discussed
- Sample
 - From sampling to final data matrix: selection of cases, refusing to participate, and missing answers on questions
- What is important for the quality of the data?
 - A possible causal link between missing observations and the topic studied
- What can be done if data are faulty?

Fall 2009

© Erling Berge 2009

53

Writing up a model

- Defining the elements of the model
 - Variables, error term, population, and sample
- Defining the relations among the elements of the model
 - Sampling procedure, time sequence of the events and observations, the functions that links the elements into an equation
- Specification of the assumptions stipulated to be true in order to use a particular method of estimation
 - Relationship to substance theory (specification requirement)
 - Distributional characteristics of the error term

Fall 2009

© Erling Berge 2009

54

Elements of a model

- Population
- Sample
- Variables
- Error terms

Fall 2009

© Erling Berge 2009

55

Relations among elements of a model

- Sampling: biased sample?
- Time sequence of events and observations
(important to aid causal modelling)
- Co-variation (genuine vs spurious co-variation)
 - Conclusions about causal impacts require genuine co-variation
- Equations and functions

Fall 2009

© Erling Berge 2009

56

Bivariat Regresjon: Modelling a population

- $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
- $i=1, \dots, n$ $n = \#$ cases in the population
- Y and X must be defined unambiguously, and Y must be interval scale (or ratio scale) in ordinary regression (OLS regression)

Fall 2009

© Erling Berge 2009

57

Bivariat Regresjon: Modelling a sample

- $Y_i = b_0 + b_1 x_{1i} + e_i$
- $i=1, \dots, n$ $n = \#$ cases in the sample
- e_i is usually called the residual (not the error term as in the population model)
- Y and X must be defined unambiguously, and Y must be interval scale (or ratio scale) in ordinary regression (OLS regression)

Fall 2009

© Erling Berge 2009

58

An example of a bad regression

- The example following contains a series of errors. If you present such a regression in your term paper you will fail
- Your task is to identify the errors as quickly as possible and then never do the same
- Clue: look again at the distributions of the variables above

Fall 2009

© Erling Berge 2009

59

Importance of public control of sales of agric. Estates

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.047(a)	.002	.000	2.213

a Predictors: (Constant), NUMBER OF DEKAR LAND OWNED

Fall 2009

© Erling Berge 2009

60

Importance of public control of sales of agric. Estates ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.145	1	4.145	.846	.358(a)
	Residual	1851.905	378	4.899		
	Total	1856.050	379			

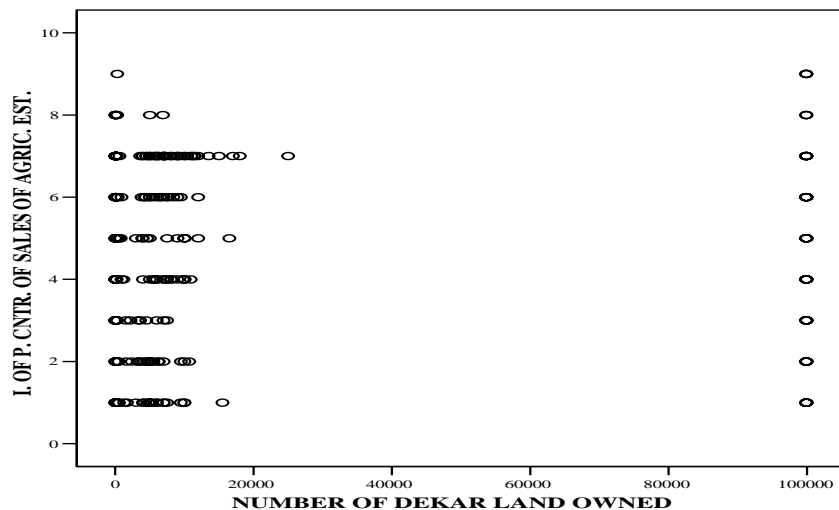
a Predictors: (Constant), NUMBER OF DEKAR LAND OWNED
 b Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Importance of public control of sales of agric. Estates Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.610	.131		35.233	.000
	NUMBER OF DEKAR LAND OWNED	.000	.000	-.047	-.920	.358

a Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Scatterplot

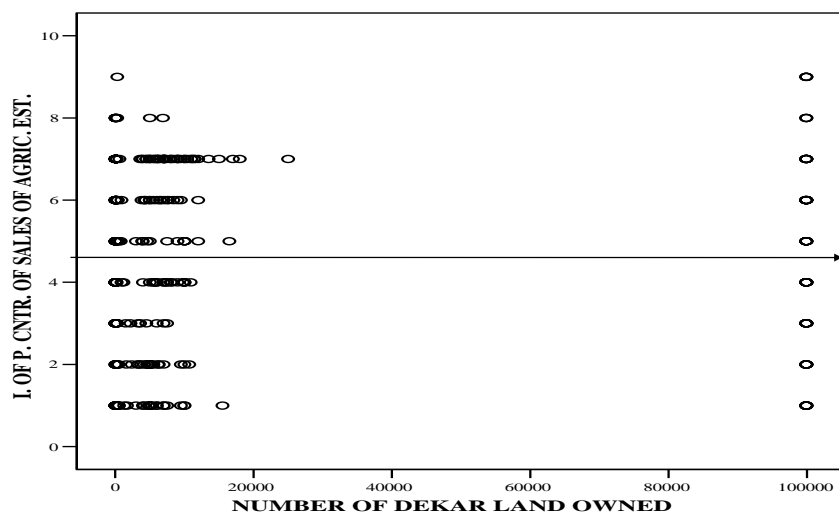


Fall 2009

© Erling Berge 2009

63

Scatterplot with regression line



Fall 2009

© Erling Berge 2009

64

Assumptions needed for the use of OLS to estimate a regression model

OLS: ordinary least squares (minste kvadrat metoden)

Requirements for OLS estimation of a regression model can shortly be summed up as

- We assume that the linear model is correct (true) with independent, and identical normally distributed error terms ("normal i.i.d. errors")

Fall 2009

© Erling Berge 2009

65

Estimation method: OLS

- Model $Y_i = b_0 + b_1 x_{1i} + e_i$

The observed error (the residual) is

- $e_i = (Y_i - b_0 - b_1 x_{1i})$

Squared and summed residual

- $\sum_i (e_i)^2 = \sum_i (Y_i - b_0 - b_1 x_{1i})^2$

Find b_0 and b_1 that minimizes the squared sum

Fall 2009

© Erling Berge 2009

66

Relationship sample - population (1)

- A new mathematical operator: $E[\varpi]$ meaning the expected value of $[\varpi]$ where ϖ stands for some expression containing at least one variable or unknown parameter, e.g.
- $E[Y_i] = E[b_0 + b_1 x_{1i} + e_i]$

$$= \beta_0 + \beta_1 x_{1i}$$
- Note in particular that in our model
 - $E[b_0] = \beta_0$;
 - $E[b_1] = \beta_1$;
 - $E[e_i] = \varepsilon_i$

Fall 2009

© Erling Berge 2009

67

Relationship sample – population (2)

- Relationship sample - population is determined by the characteristics that the error term has been given in the sampling and observation procedure
- In a simple random sample with complete observation
 $E[\varepsilon_i] = 0$ for all i , and
 $\text{var}[\varepsilon_i] = \sigma^2$ for all i

NB: $\text{var}(\varpi)$ is a new mathematical operator meaning "the procedure that will find the variance of some algebraic expression " ϖ "

Fall 2009

© Erling Berge 2009

68

Complete observation

- Make it possible to make a completely specified model. This means that all variables that causally affects the phenomenon we study (Y) have been observed, and are included in the model equation
- This is practically impossible. Therefore the error term will include also unobserved factors affecting (Y)

Fall 2009

© Erling Berge 2009

69

Testing hypotheses I

	In reality H_0 is true	In reality H_0 is untrue
We conclude that H_0 is true	Our method gives the correct answer with probability $1 - \alpha$	<u>Error of type II</u> (probability $1 - \beta$)
We conclude that H_0 is untrue	<u>Error of type I</u> The test level α is the probability of errors of type I	Our method gives the correct answer with probability β (= power of the test)

Fall 2009

© Erling Berge 2009

70

Testing hypotheses II

- A test is always constructed based on the assumption that H_0 is true
- The construction leads to a
 - **Test statistic**
- The test statistic is constructed so that it has a known probability distribution, usually called a
 - **sampling distribution**

Fall 2009

© Erling Berge 2009

71

T-test and F-test

- Sums of squares
 - $TSS = ESS + RSS$
 - $RSS = \sum_i (e_i)^2 = \sum_i (Y_i - \hat{Y}_i)^2$ distance observed- estimated value
 - $ESS = \sum_i (\hat{Y}_i - \bar{Y})^2$ distance estimated value - mean
 - $TSS = \sum_i (Y_i - \bar{Y})^2$ distance observed value – mean
- Test statistic
 - $t = (\mathbf{b} - \boldsymbol{\beta}) / \mathbf{SE}_b$ SE = standard error
 - $F = [ESS/(K-1)]/[RSS/(n-K)]$ K = number of model parameters

Fall 2009

© Erling Berge 2009

72

The p-value of a test

- The p-value of a test gives the estimated probability for observing the values we have in our sample or values that are even more in accord with a conclusion that H_0 is untrue; assuming that our sample is a simple random sample from the population where H_0 in reality is true
- Very low p-values suggest that we cannot believe that H_0 is true

Fall 2009

© Erling Berge 2009

73

Confidence interval for β

- Pick a t_α -value from the table of the t-distribution with $n-K$ degrees of freedom so that the interval
$$\langle b - t_\alpha(SE_b), b + t_\alpha(SE_b) \rangle$$
in a two-tailed test gives a probability of α for committing error of type I
- This means that $b - t_\alpha(SE_b) \leq \beta \leq b + t_\alpha(SE_b)$ with probability $1 - \alpha$

Fall 2009

© Erling Berge 2009

74

Coefficient of determination

Coefficient of determination:

- $R^2 = ESS/TSS = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$
 - Tells us how large a fraction of the variation around the mean we can "explain by" (attribute to) the variables included in the regression (\hat{Y}_i = predicted y)
- In bi-variate regression the coefficient of determination equals the coefficient of correlation:
 $r_{yu}^2 = s_{yu} / s_y s_u$
- Co-variance: $s_{yu} = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})(U_i - \bar{U})$

Fall 2009

© Erling Berge 2009

75

Detecting problems in a regression

- Take a second look at the example presented above where
 - Y = IMPORTANCE OF PUBLIC CONTROL OF SALES OF AGRICULURAL ESTATES
 - X = NUMBER OF DEKAR LAND OWNED
 - $Y_i = b_0 + b_1 x_{1i} + e_i$

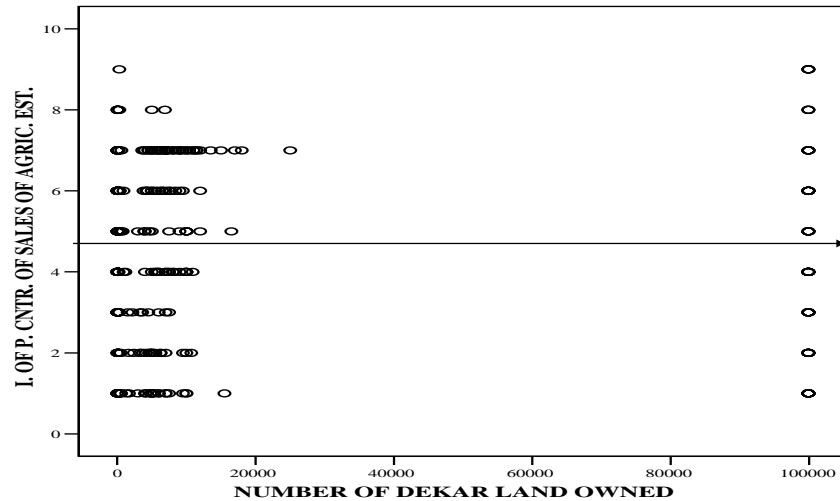
What was the problem in this example?

Fall 2009

© Erling Berge 2009

76

What is wrong in this scatterplot with regression line?



Fall 2009

© Erling Berge 2009

77

In general: what can possibly cause problems?

- Omitted variables
- Non-linear relationships
- Non-constant error term
(heteroskedastisitet)
- Correlation among error terms
(autocorrelation)
- Non-normal error terms
- Influential cases

Fall 2009

© Erling Berge 2009

78

Non-normal errors:

- Regression **DO NOT need assumptions about the distribution of variables**
- But to test hypotheses about the parameters we need to assume that the **error terms are normally distributed** with the same mean and variance
- **If the model is correct** (true) and n (number of cases) is large the central limit theorem demonstrates that the error terms approach the normal distribution
- **But usually a model will be erroneously or incompletely specified.** Hence we need to inspect and test residuals (observed error term) to see if they actually are normally distributed

Fall 2009

© Erling Berge 2009

79

Residual analysis

- This is the most important starting point for diagnosing a regression analysis

Useful tools:

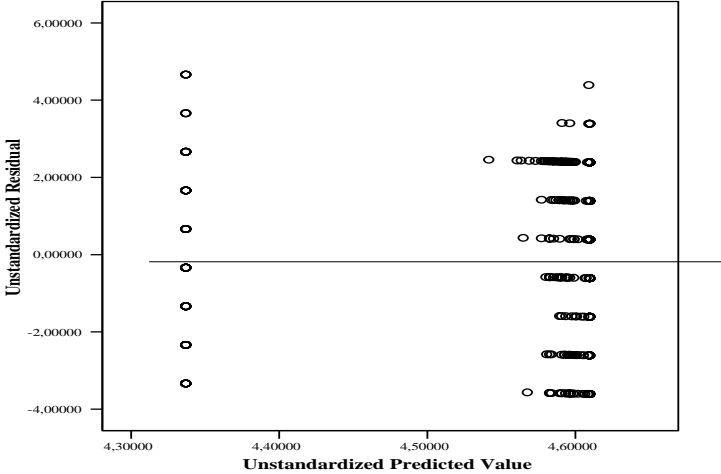
- Scatterplot
- Plot of residual against predicted value
- Histogram
- Boxplot
- Symmetry plot
- Quantil-normal plot

Fall 2009

© Erling Berge 2009

80

What went wrong? (1) residual-predicted value plot



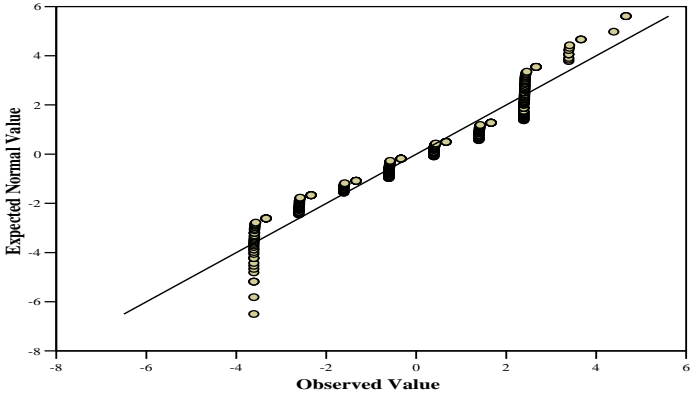
Fall 2009

© Erling Berge 2009

81

What went wrong? (1) normal-quantile plot

Normal Q-Q Plot of Unstandardized Residual



Fall 2009

© Erling Berge 2009

82

Power transformations

May solve problems related to

- Curvilinearity in the model
- Outliers
- Influential cases
- Non-constant variance of the error term (heteroskedasticity)
- Non-normal error term

NB: Power transformations are used to solve a problem. If you do not have a problem do not solve it.

Fall 2009

© Erling Berge 2009

83

Power transformations (see H:17-22)

Y^* : read

“transformed Y ”

(transforming Y to Y^*)

- $Y^* = Y^q \quad q > 0$
- $Y^* = \ln[Y] \quad q = 0$
- $Y^* = - [Y^q] \quad q < 0$

Inverse

transformation

(transforming Y^* to Y)

- $Y = [Y^*]^{1/q} \quad q > 0$
- $Y = \exp[Y^*] \quad q = 0$
- $Y = [- Y^*]^{1/q} \quad q < 0$

Fall 2009

© Erling Berge 2009

84

Power transformations: consequences

- $X^* = X^q$
 - $q > 1$ increases the weight of the right hand tail relative to the left hand tail
 - $q = 1$ produces identity
 - $q < 1$ reduces the weight of the right hand tail relative to the left hand tail
- If $Y^* = \ln(Y)$ the regression coefficient of an interval scale variable X can be interpreted as % change in Y per unit change in X

E.g. if $\ln(Y) = b_0 + b_1 x + e$

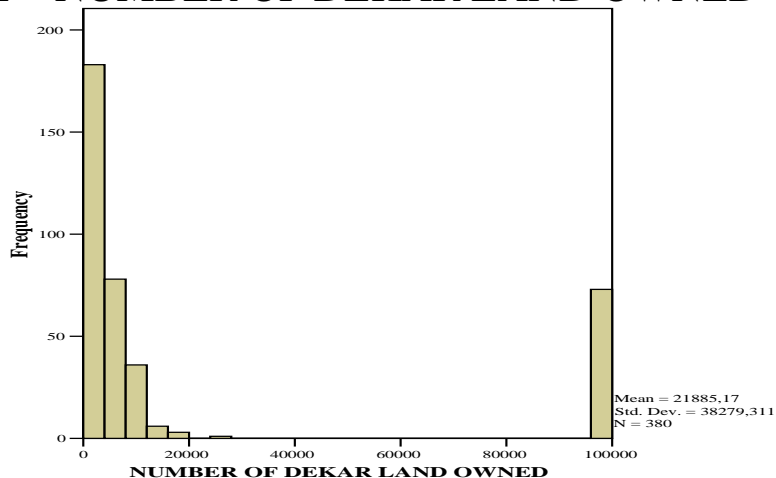
b_1 can be interpreted as % change in Y pr unit change in X

Fall 2009

© Erling Berge 2009

85

Point of departure $X = \text{NUMBER OF DEKAR LAND OWNED}$

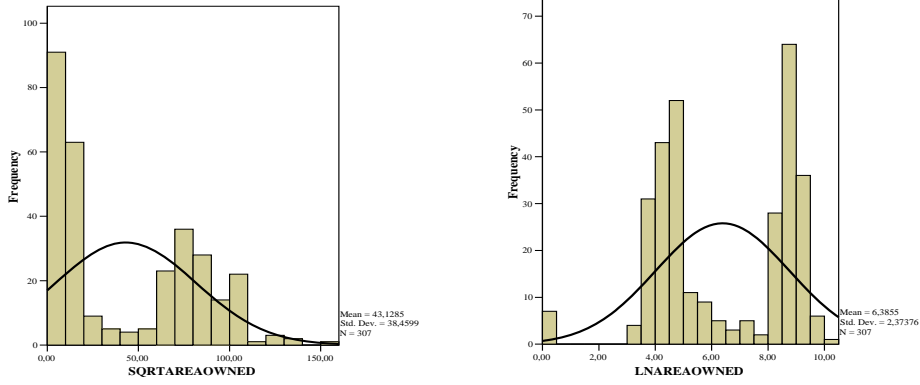


Fall 2009

© Erling Berge 2009

86

Power transformed
X = NUMBER OF DEKAR LAND OWNED



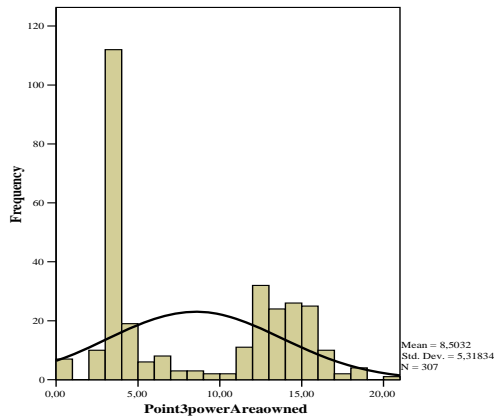
SQRT=square root of areaowned – LN= natural logarithm of (areaowned+1)

Fall 2009

© Erling Berge 2009

87

Power transformed
X = NUMBER OF DEKAR LAND OWNED



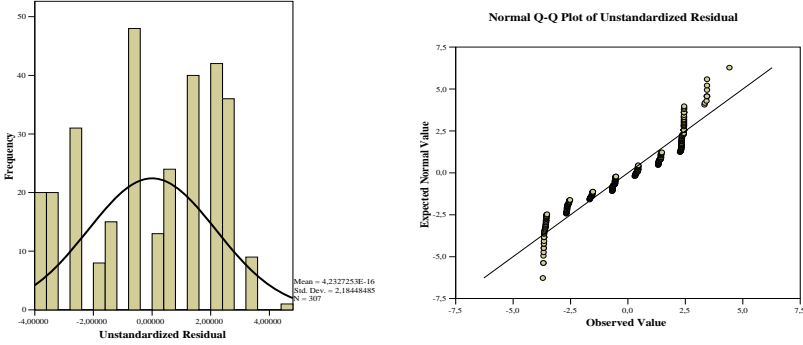
Point3power = 0,3 power of areaowned

Fall 2009

© Erling Berge 2009

88

Does power transformation help?



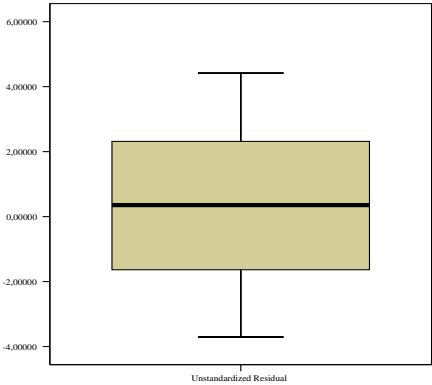
0.3 power-transformation gives lighter tails and no outliers

Fall 2009

© Erling Berge 2009

89

Box plot of the residual shows approximate symmetry and no outliers



Fall 2009

© Erling Berge 2009

90

Curvilinear regression

- The example above used the variable "Point3powerAreaowned", or 0.3 power of number of dekar land owned:

- $\text{Point3powerAreaowned} = (\text{NUMBER OF DEKAR LAND OWNED})^{0.3}$

The model estimated is thus

$$y_i = b_0 + b_1 (x_i) + e_i$$

$$y_i = b_0 + b_1 (\text{Point3powerAreaowned}_i) + e_i$$

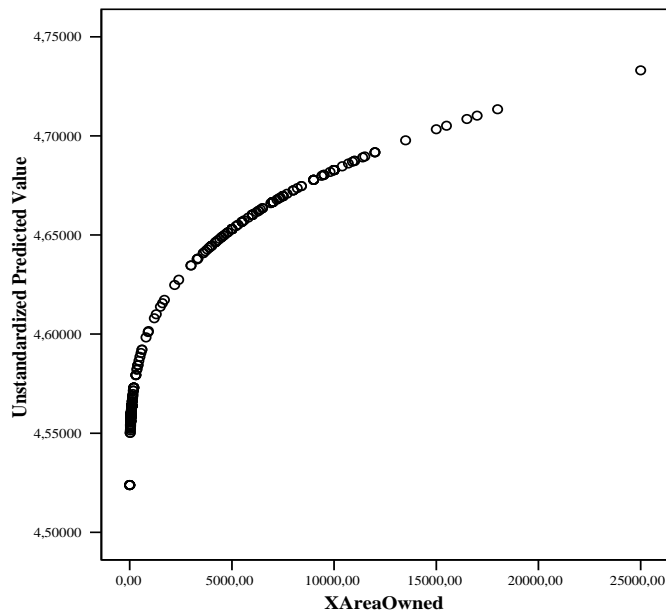
$$\hat{y}_i = 4.524 + 0.010 * (\text{NUMBER OF DEKAR LAND OWNED}_i)^{0.3}$$

Fall 2009

© Erling Berge 2009

91

Use of power transformed variables means that the regression is curvilinear



Fall 2009

© Erling Berge 2009

92

Summary

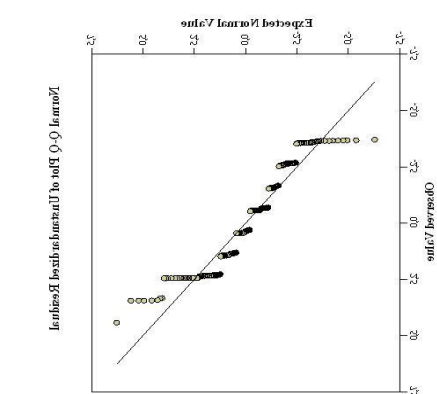
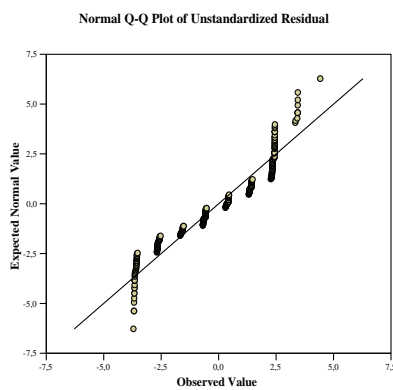
- In bi-variate regression the OLS method finds the "best" LINE or CURVE in a two dimensional scatter plot
- Scatter-plot and analysis of residuals are tools for diagnosing problems in the regression
- Transformations are a general tool helping to mitigate several types of problems, such as
 - Curvilinearity
 - Heteroscedasticity
 - Non-normal distributions of residuals
 - Case with too high influence
- Regression with transformed variables are always curvilinear. Results can most easily be interpreted by means of graphs

Fall 2009

© Erling Berge 2009

93

SPSS printout vs the book (see p16)



Fall 2009

© Erling Berge 2009

94

Reading printout from SPSS (1)

Descriptive Statistics					Mean	Std. Deviation ¹	N ²		
I. OF P. CNTR. OF SALES OF AGRIC. EST.					4.61	2.185	307		
Point3powerAreaowned					8.5032	5.31834	307		

Model	Change Statistics								
	R	R Square ³	Adjusted R Square ⁴	Std. Error of the Estimate ⁵	R Square Change	F Change	df1	df2	Sig. F Change
1	.024(a)	.001	-.003	2.188	.001	.182	1	305	.670

a Predictors: (Constant), Point3powerAreaowned

b Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Footnotes to the table above (1)

1. Standard deviation of the mean
2. Number of cases used in the analysis
3. Coefficient of determination
4. The adjusted coefficient of determination (see Hamilton page 41)
5. Standard deviation of the residual

$$s_e = \text{SQRT} (\text{RSS}/(n-K)),$$

where SQRT (*) = square root of (*)

Reading printout from SPSS (2)

Model		Sum of Squares ³	df	Mean Square	F ¹	Sig. ²
1	Regression	.870	1	.870	.182	.670(a)
	Residual	1460.224	305	4.788		
	Total	1461.094	306			

•Sums of squares: $TSS = ESS + RSS$

• $RSS = \sum_i (e_i)^2 = \sum_i (Y_i - \hat{Y}_i)^2$: sum of squared (distance observed – estimated value)

•Mean Square = RSS / df For RSS it is known that $df=n-K$

K equals number of parameters estimated in the model (b_0 og b_1)

Here we have $n=307$ and $K=2$, hence $Df = 305$

Fall 2009

© Erling Berge 2009

97

Footnotes to the table above (2)

1. F-statistic for the null hypothesis $\beta_1 = 0$ (see Hamilton p45)
2. p-value of the F-statistic: the probability of finding a F-value this large or larger assuming that the null hypothesis is correct
3. Sums of squares
 1. $TSS = ESS + RSS$
 2. $RSS = \sum_i (e_i)^2 = \sum_i (Y_i - \hat{Y}_i)^2$ distance observed value – estimated value
 3. $ESS = \sum_i (\hat{Y}_i - \bar{Y})^2$ distance estimated value – mean
 4. $TSS = \sum_i (Y_i - \bar{Y})^2$ distance observed value – mean

Fall 2009

© Erling Berge 2009

98

Reading printout from SPSS (3)

M o d e l		Unstandardized Coefficients		Standardized Coefficients	t ⁴	Sig. ⁵	95% Confidence Interval for B	
		B ¹	Std. Error ²	Beta ³			Lower Bound	Upper Bound
1	(Constant)	4.524	.236		19.187	.000	4.060	4.988
	Point3- powerA rea- owned	.010	.024	.024	.426	.670	-.036	.056

Fall 2009

© Erling Berge 2009

99

Footnotes to the table above (3)

1. Estimates of the regression coefficients b_0 og b_1
2. Standard error of the estimates of b_0 og b_1
3. Standardized regression coefficients: $b_1^{st} = b_1 * (s_x / s_y)$ see Hamilton pp38-40
4. t-statistic for the null hypothesis $\beta_1 = 0$ (see Hamilton p44)
5. p-value of the t-statistic: the probability of finding a t-value this large or larger assuming that the null hypothesis is correct

Fall 2009

© Erling Berge 2009

100