ECON 3710, ECON 3720

SHORT INTRODUCTION TO REGRESSION ANALYSIS

Nico Keilman
January 2008

Outline

- Scattergram
- Various forms for the association between two variables
- Perfect linear association
- Imperfect linear association
- Estimation: the method of least squares
- Standard error of the estimates
- t-value
- R^2
- Multivariate regression
- Dummy variables (ordinal and nominal variables)
- Transformations
- The main idea behind Maximum Likelihood Estimation
- Individual versus aggregate data

Recommended literature: M.S. Lewis-Beck: "Applied Regression: An Introduction". Series Quantitative Applications in the Social Sciences nr. 22. Beverly Hills: Sage Publications 1980

SHORT INTRODUCTION TO REGRESSION ANALYSIS

Regression analysis is a statistical technique that attempts to answer the following question: "What is the association between variable X and variable Y?"

Examples:

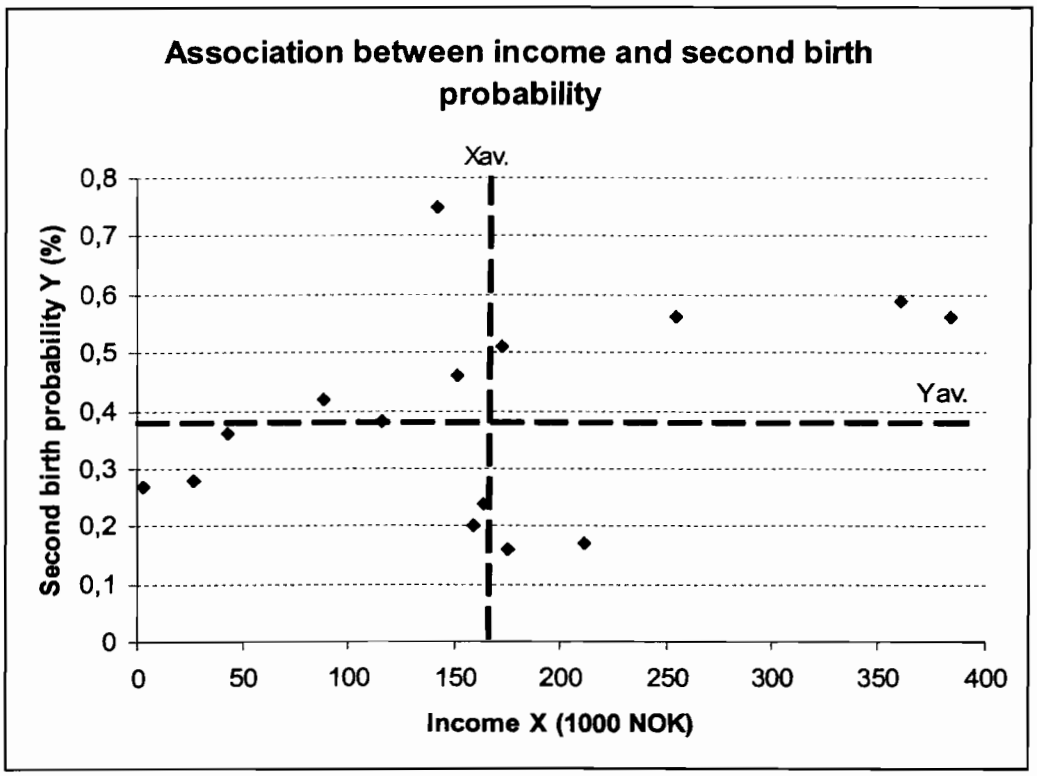
- Does the probability for a two-child mother to have an additional child increase with her income?
- Are death risks among smokers higher than those for non-smokers?
- Do cohabiting couples who converted their consensual union into a marriage, have lower divorce rates than married couples who married without prior cohabitation?

In this kind of analysis we have an independent variable (often written as X), which affects the dependent variable (Y).

Formally: $X \longrightarrow Y$

How can we find out what the association between X and Y looks like? Measure both X-values and Y-values among a group of individuals (the whole population, or just a sample). You obtain a series of measured values: (X_1,Y_1) for person nr. 1, (X_2,Y_2) for person nr. 2, (X_3,Y_3) for person nr. 3, ... etc. In general we write (X_i, Y_i) for person nr. i. These values may be plotted in a graph with horizontal X-axis and vertical Y-axis. Each person is represented by one point. All points are "scattered" over the graph – hence the name "scattergram" (Norwegian: "spredningsdiagram").

Example 1: the association between income and the probability to have a second child.



Each dot represents measurements for one woman: her annual income, and the probability to have a second child. We see a tendency that a high income goes together with a high second birth probability: women with an income above the average (X_i > Xav.) have also a higher than average probability for a second birth (Y_i > Yav.).

Example 2: the association between income and satisfaction with the current marriage. 100 women, married for at least 5 years.

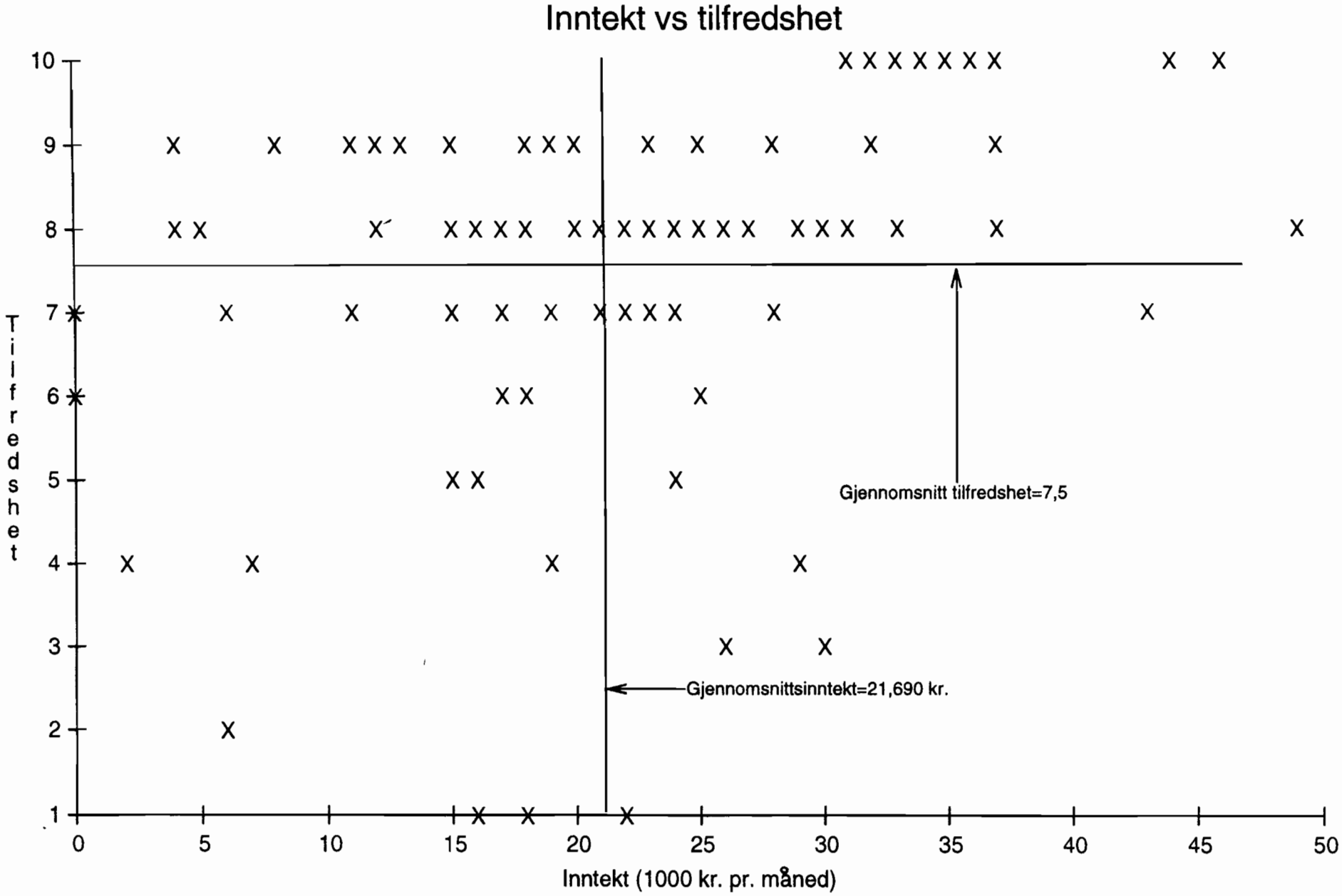
X: monthly household income (1000 NOK)

Y: the woman's satisfaction with her marriage (1 = very unsatisfied, …., 10 = very satisfied).

Source: the 1988 Family and Occupation Survey for Norway.

Average monthly income = 21 690 NOK. Average satisfaction = 7.5.

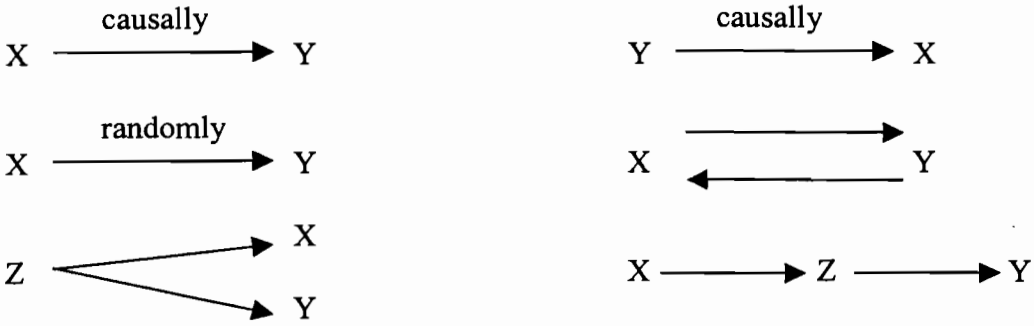
See plot next page.


Inntekt vs tilfredshet
Tilfredshet
10
9
8
7
6
5
4
3
2
1
0
5
10
15
20
25
30
35
40
45
50
Inntekt (1000 kr. pr. måned)
Gjennomsnitt tilfredshet=7,5
Gjennomsnittsinntekt=21,690 kr.

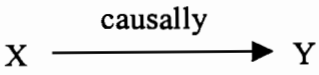
Important

When we analyse the association between two variables X and Y, we should start off from a theory, or at least a hypothesis (based on plausible reasoning) for that association.

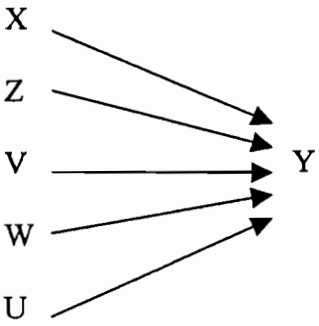
Possible associations:



In a bivariate analysis, we study the association between two variables



In a multivariate analysis, we study the association between three or more variables – only one of these is the dependent variable.

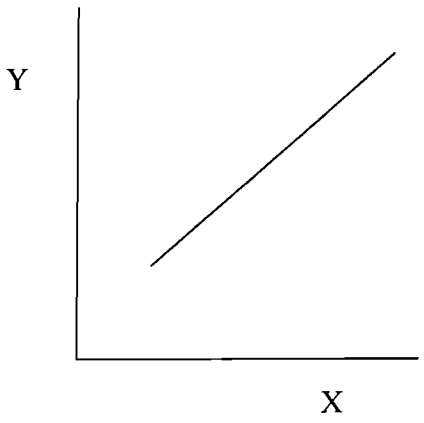


In this note, we first present bivariate regression analysis, next multivariate regression analysis.

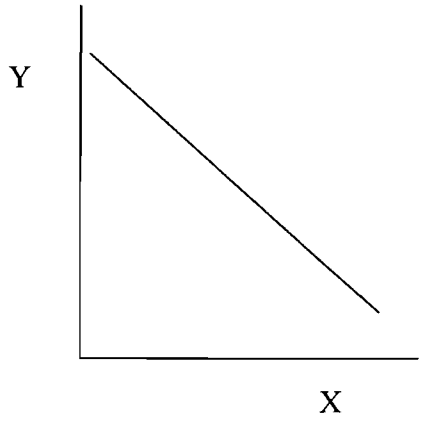
The purpose of a regression analysis is to find a mathematical relationship between the two variables X and Y.

Ideally, we would like to find a perfect linear relationship between X and Y

For instance,



or



Let us assume that we have measurements for n persons. The relationship between values X and Y for person nr. i (assumed linear) can be written as

$$Y_i = a + bX_i \qquad i = 1, 2, 3, \ldots, n$$

All points (X_i, Y_i) lie on a straight line. a and b are parameters, the value of which we wish to compute on the basis of the data set.

a: constant

b: coefficient, slope

a represents that value the dependent variable Y takes, in case the independent variable X is zero. For instance, the satisfaction with marriage when household income is zero. Thus, in this simple case with a perfect linear relationship, we can find the value of a by looking for a person who has X = 0; a equals this person's Y-value.

b equals the increase in Y for a unit increase in X. "How large is the increase in marriage satisfaction when household income increases by one unit, i.e. 1000 NOK?" When we have two persons, i=1 and i=2, with measurements (X_1,Y_1) and (X_2,Y_2), we can compute the slope b as $(Y_2 - Y_1)/(X_2 - X_1)$.

In practice, such a perfect linear relationship never occurs. Frequently we encounter an imperfect linear relationship (I), or an imperfect non-linear relationship (II, III).

I II III

To start with, we shall look at *linear* relationships → linear regression

Towards the end, we shall briefly look at *non-linear* relationships

In Figure I above, all points lie near the straight line – some very close to it, others further away.

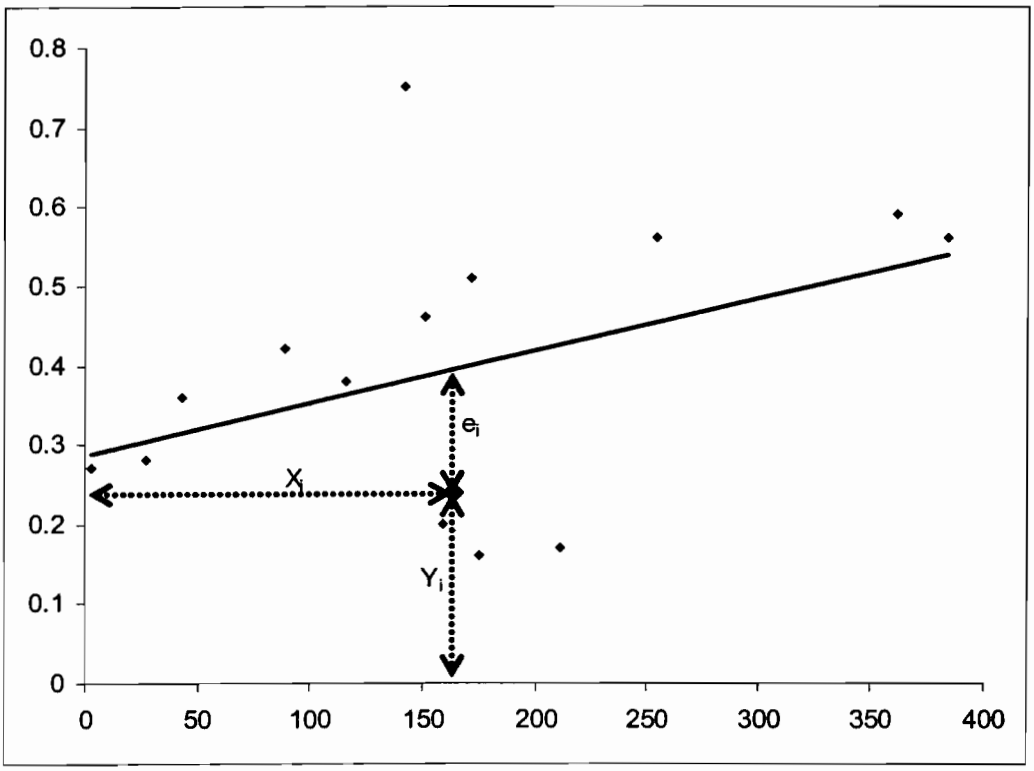
Thus means that the relationship $Y_i = a + bX_i$ no longer holds. Instead we write

$$Y_i = a + bX_i + e_i \qquad i = 1, 2, 3, \ldots, n \qquad (1)$$

a and b are as before, constant and coefficient. e_i is a random term, "residual"

The residual is different for different persons, hence index i.

e_i has a statistical distribution. Its expected value ("mean") is zero.



How can we compute ("estimate") a and b?

Several methods. Most often used are:

- Least Squares
- Maximum likelihood – ML

This note will present the main principles for the method of Least Squares, and explain briefly the idea behind Maximum Likelihood. Both Least Squares and Maximum Likelihood require some calculus and statistics.

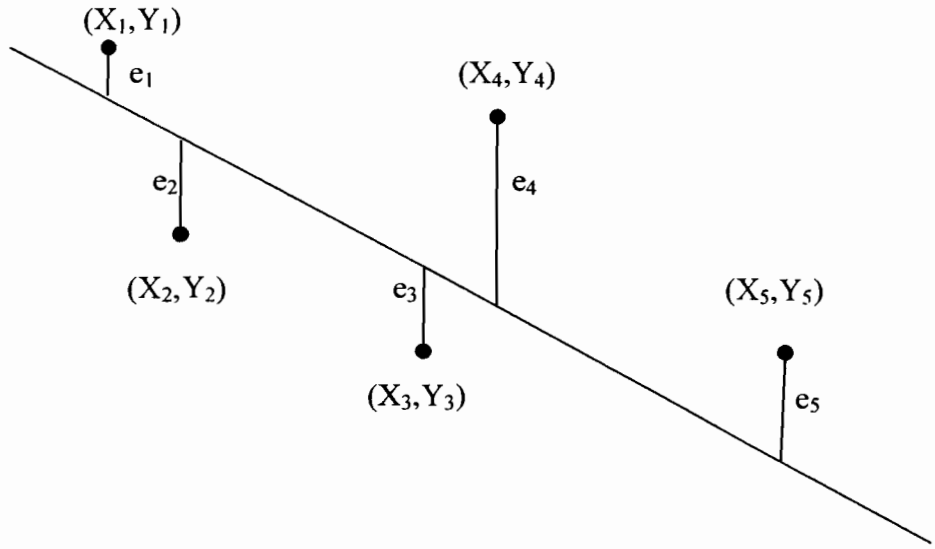
The method of Least Squares

The purpose is to find a straight line which fits the data best, according to some criterion. The values of a and b determine the exact position of the straight line in the scattergram.

Increase/decrease a → line shifts upwards/downwards.

Increase/decrease b → line rotates counter clockwise/clockwise

Main idea behind Least Squares (LS)-regression: Select a and b such that the vertical distance between the line and the data points (X_i, Y_i) is minimal. This distance is represented by the residuals $e_1, e_2, \ldots, e_n$. More exactly: minimize the sum $(e_1)^2 + (e_2)^2 + (e_3)^2 + \ldots + (e_n)^2$..



We take the squared value of the residuals, because positive residuals (e_1, e_4, e_5) should get as much weight as negative residuals. From formula (1) we know that

$e_i = Y_i - a - bX_i$. Hence we have to choose a and b such that $\sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(Y_i - a - bX_i)^2$ is minimal.

Some calculus results in the following expressions:

$b = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, and $a = \bar{Y} - b\bar{X}$ .Here $\bar{X}$ and $\bar{Y}$ represent the average values of the observed X- and Y-values, respectively.

Take a look at these expressions, remember where you can find them, and forget them. Statistical programs such as Excel compute them, in addition to many other variables that you need to carry out a regression analysis.

Example

Satisfaction with marriage versus income

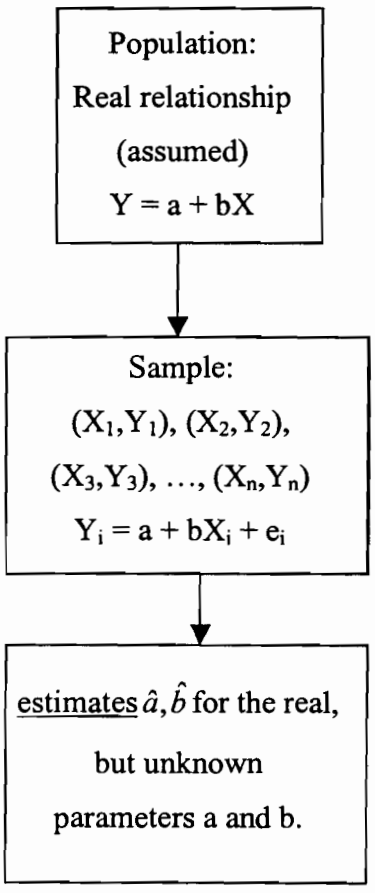
Excel gave the following results: a = 6.045 and b = 0.067.

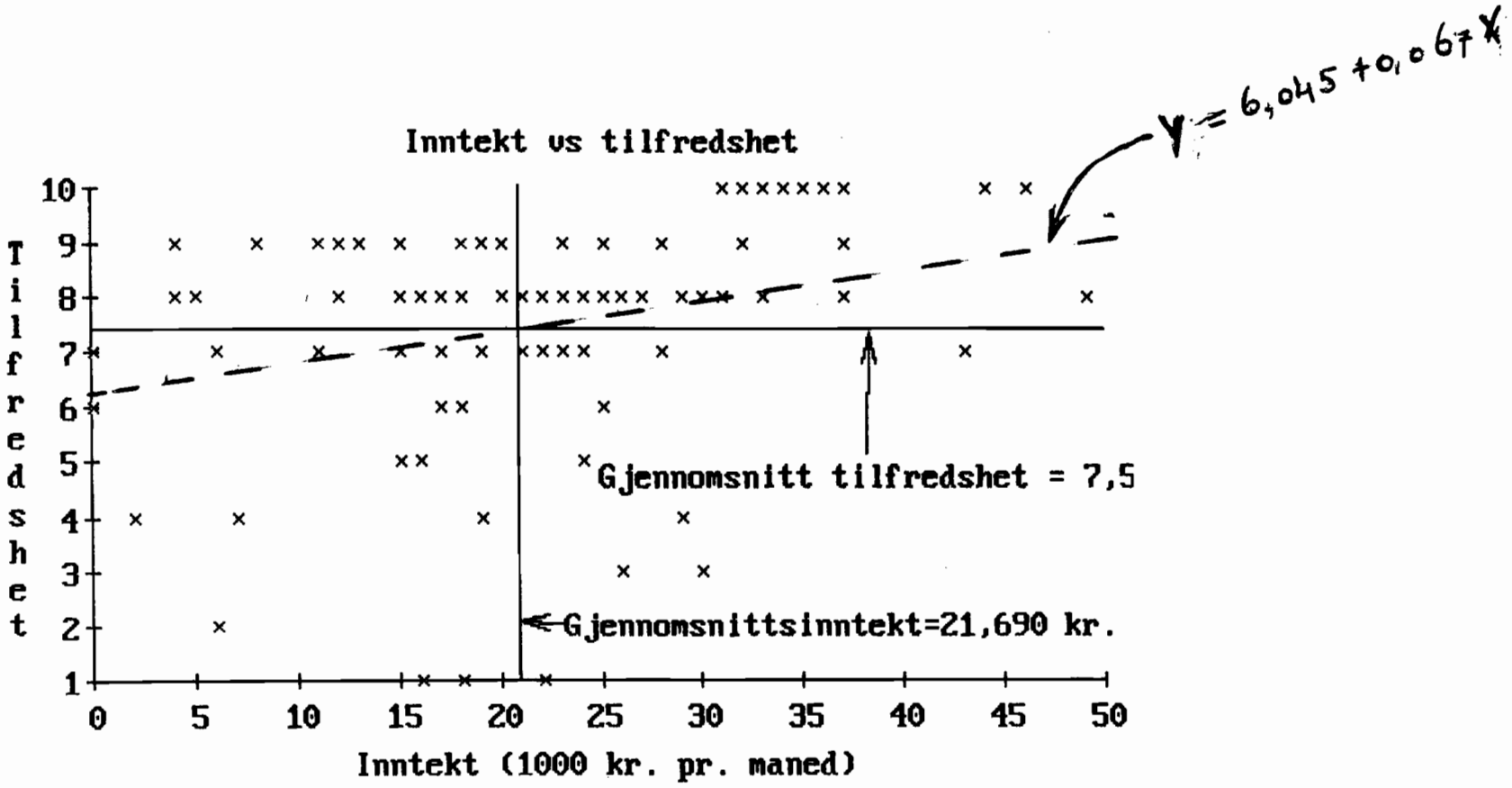
This means that a person with no income has satisfaction a little higher than 6. For every additional 10 000 crowns earned will satisfaction be 0.67 points higher.

See plot on next page.

Note: the average person, that is, a person who has $X_i = \overline{X}$ and $Y_i = \overline{Y}$ has a residual e_i equal to zero!

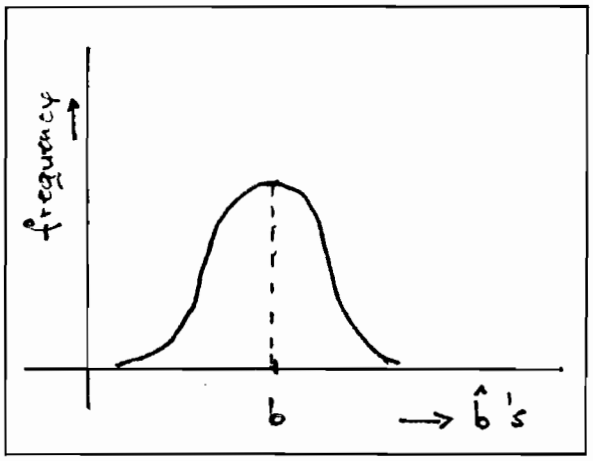
Summing up




Y = 6,045 + 0,067 X
Inntekt vs tilfredshet
Tilfredshet
10
9
8
7
6
5
4
3
2
1
Gjennomsnitt tilfredshet = 7,5
Gjennomsnittsinntekt=21,690 kr.
0
5
10
15
20
25
30
35
40
45
50
Inntekt (1000 kr. pr. maned)

Uncertainty in the estimates

The estimates that we computed (for instance $\hat{a}$=6.045, $\hat{b}$=0.067 in this example) are unique for this sample. A different (representative) sample would have given different estimates for a and b, (for example $\hat{a}$=6.5, $\hat{b}$=0.08. Many samples would have resulted in a whole series of estimates $\hat{a}$ and $\hat{b}$. We can prove statistically that the average of the $\hat{a}$-values over all (representative) samples equals the unknown value a. Similarly for all $\hat{b}$'s and b.



We have only one data set. The flatter the distribution of $\hat{b}$ is, the larger is the chance that we will estimate a value $\hat{b}$, which is far off from the real b.

In other words, a flat distribution for the $\hat{b}$'s signals large uncertainty concerning how well $\hat{b}$ reflects the real b, as compared to a narrow distribution of the $\hat{b}$'s.

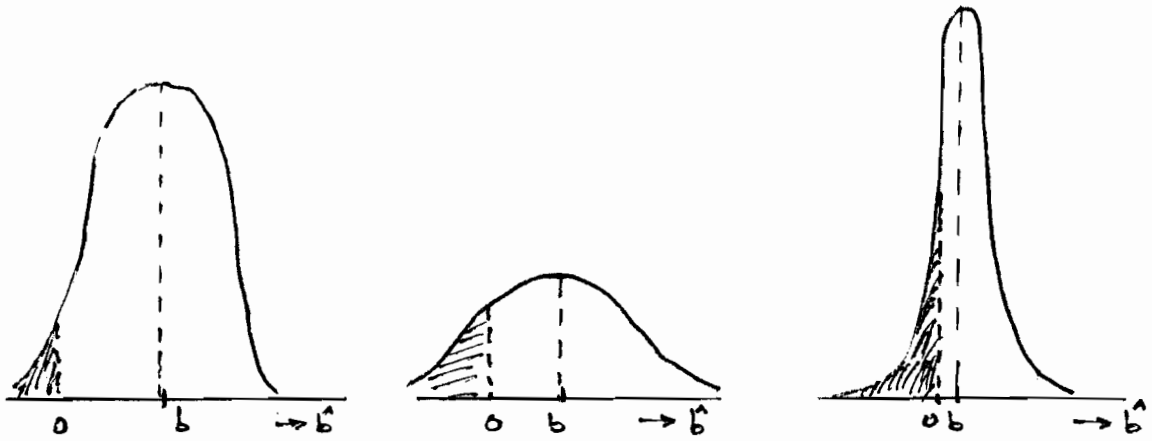
Whether the distribution is flat or narrow is reflected in the *standard deviation* of the $\hat{b}$'s: given a series of estimates $\hat{b}_1, \hat{b}_2, \hat{b}_3 \ldots \hat{b}_k$ based on k samples, we can compute the standard deviation of $\hat{b}$ as follows:

$$s_{\hat{b}} = \sqrt{\frac{\sum_{j=1}^{k}(\hat{b}_j - \bar{\hat{b}})^2}{k-1}}$$, where $\bar{\hat{b}}$ is the average of the estimates $\hat{b}_j$ (which is equal to b!).

Normally we have just one sample, and just one value of $\hat{b}$ (and $\hat{a}$). In that case we cannot compute the standard deviation of $\hat{b}$ (nor that of $\hat{a}$). But statistical theory gives us the

possibility to make an *estimate* of this standard deviation. It is called the "(estimated) standard error of the estimate", written as $se_{\hat{a}}$ and $se_{\hat{b}}$. Statistical programs, including Excel, give us readily the values of these standard errors of the estimates.

In case the standard error of an estimate is large compared to the estimate's value itself, are we uncertain about the real value of the parameter.



Little uncertainty: $se_{\hat{b}}$ is small as compared to b

Large uncertainty: $se_{\hat{b}}$ is large as compared to b

Large uncertainty: $se_{\hat{b}}$ is small, but so is b

The same holds for $a, \hat{a}$, and $se_{\hat{a}}$. But for b in particular this is important: if we cannot trust the estimate of b, we do not know whether the real b is positive, negative, or perhaps even zero (no relationship between X and Y)!

$\frac{\hat{b}}{se_{\hat{b}}}$ is a measure for the uncertainty in the estimate of b.

This ratio is called the *t-value of* $\hat{b}$.

Rule of thumb: if the t-value of $\hat{b}$ exceeds 2 in absolute value, then there is at least a 97.5 per cent probability that the real (but unknown) value of b is unequal zero, and that it has the same sign as $\hat{b}$. The larger the t-value in absolute value, the larger that probability. The value of b *can very well be* larger or smaller than $\hat{b}$, but it is likely close to $\hat{b}$.

More formally: if the t-value of $\hat{b}$ exceeds 2 in absolute value, we say that b is "significant" or "significantly different from zero" at the 5 per cent level.

The same holds for the estimate of the slope a.

Some statistical programs do not compute t-values, just the standard errors $se_{\hat{a}}$ and $se_{\hat{b}}$ In that case you have to compute the t-values (also called t-statistics) $\hat{a} / se_{\hat{a}}$ and $\hat{b} / se_{\hat{b}}$ yourself. Other programs, such as Excel, report t-values for the estimates.

NB! The standard error of $\hat{b}$ and its t-value can only be used to determine the *uncertainty* in the b-estimate. It *cannot* be used to find out to what extent X influences Y. In other words, a *low* estimate for b (*little* impact of X on Y) can be *strongly* significant (high t-value), if only the standard error is low enough.

Example

Marriage satisfaction =	6.045 +	0.067*income
standard error	0.460	0.019
t-value	13.1	3.5

t-values are larger than 2 in this example, both for the estimate of a and that of b. In other words, the real parameters are significantly different from zero at the 5 per cent level.

NB The "critical value" for the t-ratio equal to 2 (or minus 2, in case the estimate is negative) is only correct for data sets that are not too small, that is, data sets that contain at least n = 50 data points. For smaller data sets you have to be more cautious, and check against a larger value than 2/smaller value than -2 in order to obtain significance at the 5-per cent level. For instance, for n = 10, the critical value is 2.25; for n = 30 it is 2.05.

R^2

An important indicator that is commonly used to characterize how well the model fits the data is the coefficient of determination, written as R^2. It tells us how much of the uncertainty in the original data set has been reduced by the estimated model.

Consider a certain person i. Let us assume that we know his or her value for the independent variable (X_i), but not the corresponding value for dependent variable (Y_i). What is our best guess for this person's value of Y? Before we estimated the model, our best guess would have been the average value $\overline{Y}$ of the data set. In that case, we would have made a prediction error

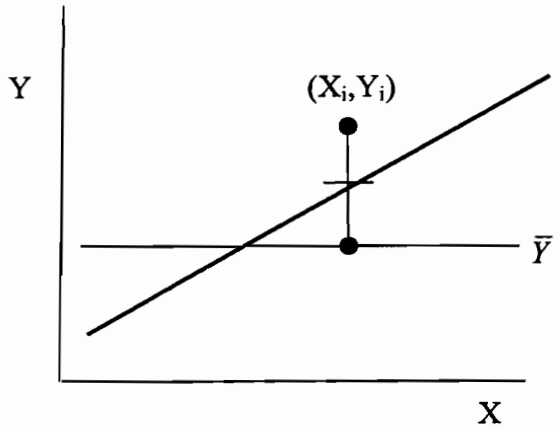
equal to $Y_i - \overline{Y}$. For <u>all</u> individuals we compute the sum of squared errors $\sum_i (Y_i - \overline{Y})^2$. This sum of squares reflects total ignorance of the relationship between X and Y. After we have estimated the model (i.e. model parameters a and b), we can improve this person's Y-prediction, when we also know his or her X-value. In that case our best guess for this person's Y-value is of course $\hat{Y}_i = \hat{a} + \hat{b}X_i$, where $\hat{a}$ and $\hat{b}$ are the estimated values for the parameters. In that case the prediction error becomes the difference between the observed value and the predicted value, or $Y_i - \hat{Y}_i$. In other words, the prediction error becomes $Y_i - \hat{a} - \hat{b}X_i = e_i$.

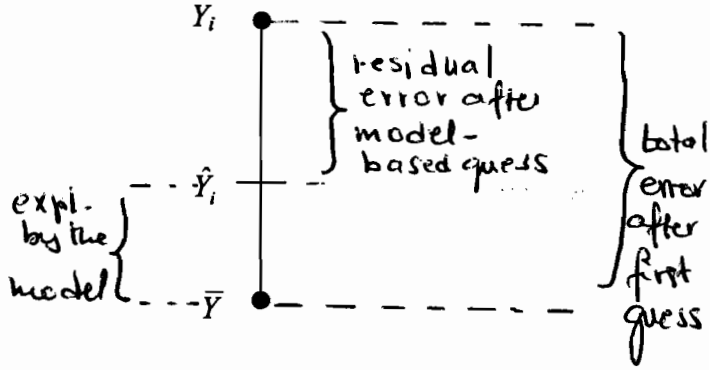
Thus, knowledge of the model has improved the prediction from $\overline{Y}$ to $\hat{Y}_i$.

Still the prediction error equals $Y_i - \hat{Y}_i$. $\sum_i (Y_i - \hat{Y}_i)^2$ for all individuals.

But the model has "explained" $\hat{Y}_i - \overline{Y}$. $\sum_i (\hat{Y}_i - \overline{Y})^2$ for all individuals.

The total error equals $\sum_i (Y_i - \overline{Y})^2$.





In other words, the relative improvement in predictive power equals the ratio $\frac{\sum_i (\hat{Y}_i - \overline{Y})^2}{\sum_i (Y_i - \overline{Y})^2}$.

This is the definition of the coefficient of determination R^2, with $0 \leq R^2 \leq 1$.

$R^2 = 1$: perfect fit: all values Y_i lie on the straight line $a+bX_i$, and all residuals e_i are zero.

$R^2 = 0$; no relationship between X and Y. Knowledge of a person's X-value does not help in predicting his or her Y-value. The estimated value of b is zero.

The larger R^2, the better the model explains the (variation in) the data.

NB1. To find a model with largest possible R^2 is not a goal in itself in regression analysis. It is useless to find two variables X and Y that have a high R^2-value, unless we can explain (have a theory) why they are related this way. Remember: the R^2 between X and Y for the model Y=a+bX is the same as that between Y and X in the model X=c+dY!

NB2. One can prove that R^2 has the same value as the square of the sample's correlation coefficient r_{XY}=corr(X,Y): $R^2 = (r_{XY})^2$.

Example Marriage satisfaction and income:

Marriage satisfaction =	6.045 +	0.067*income	$R^2 = 0.11$
t-value	13.1	3.5	

The model explains only 11% of the variation in marriage satisfaction across individuals, in spite of the strongly significant relationship between marriage satisfaction and income.

Multivariate regression

Consider the following situation: we want to analyse food consumption for a family, and the impact of family size and household income.

Y_i = food consumption pr. family

X_i = income

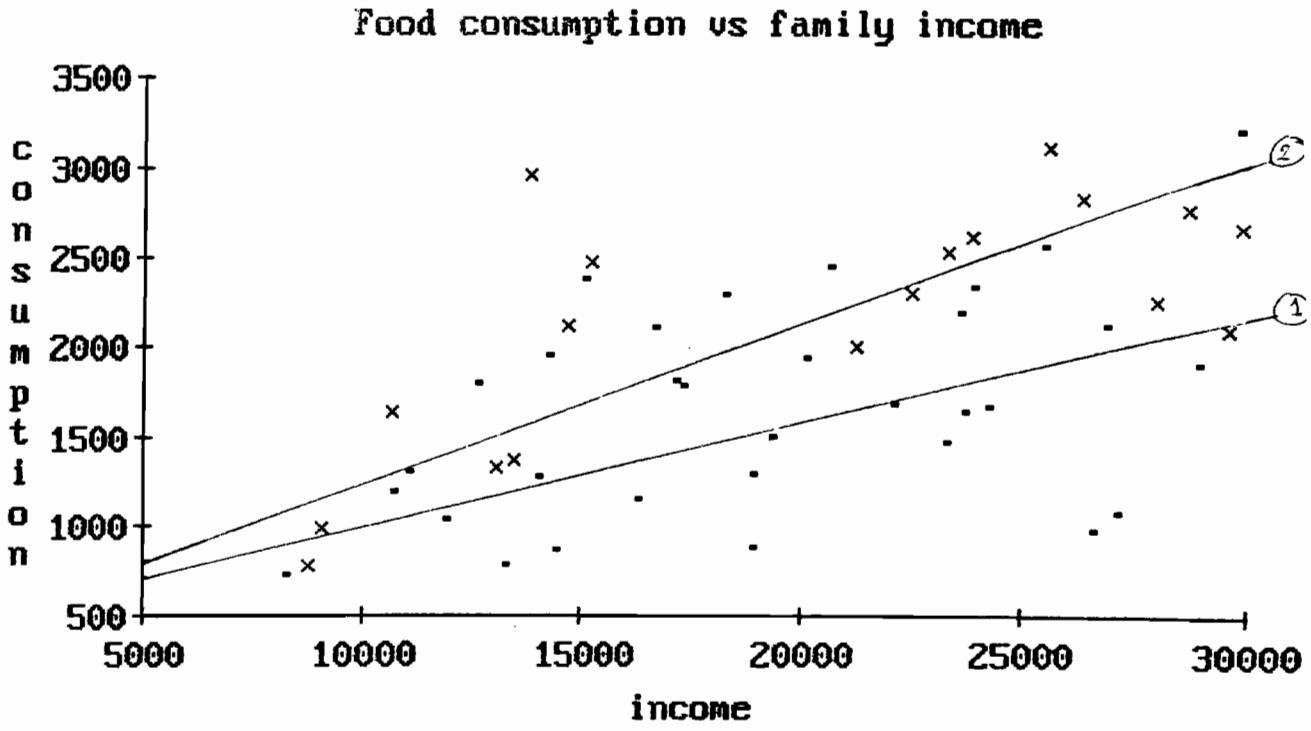
Z_i = family size

The assumption is that both high income and large family size lead to a high consumption level. We have data for 50 families on all three variables. Consumption and income are given in US dollars. We split up the data set in two smaller sets: one for family size up to three, and one for families of four or more persons. Regression nr. 1 shows the relationship between food consumption and income for small families ($Z_i \leq 3$). Regression 2 gives the same relationship, but now for large families ($Z_i > 3$). t-values are given in parentheses.

1. Y_i = 731.3 +	0.0486X_i + e_i	$R^2 = 0.24$
(2.29)	(3.10)	n = 33

2. Y_i = 759.6 +	0.0728X_i + e_i	$R^2 = 0.56$
(2.24)	(4.39)	n = 17

See graph on next page.



$x : Z_i > 3$ (2) $\quad\quad$ $\bullet : Z_i \leq 3$ (1)

Line 1 lies under line 2: large families have higher consumption than small families. This way we can analyse how both income and family size affect food consumption. We could eventually estimate a separate model for $Z_i = 1$, $Z_i = 2$, $Z_i = 3$, ... etc. But this way we cannot answer the question: by how much does consumption increase when there is one additional family member? The reason is that the estimated regression lines are not parallel. We could of course estimate the model $Y_i = c + dZ_i$, but the problem here is the fact that the estimates $\hat{c}$ and $\hat{d}$ are heavily influenced by family income.

The solution to this problem is to include the associations between consumption and income on the one hand, and consumption and family size on the other hand simultaneously, by estimating the multivariate model

$$Y = a + bX + cZ.$$

We follow the same principles as with bivariate regression:

- estimate coefficients a, b, and c by means of the Method of Least Squares;
- compute standard errors for the estimates:
- compute R^2.

Example 1

$$Y_i = 331.3 + 0.0561X_i + 129.6Z_i + e_i \qquad R^2 = 0.56$$
$$\quad (1.30) \quad (4.95) \quad (3.59) \qquad n = 50$$

One extra family member leads to an increase in consumption equivalent to 129.6 dollars, other things being equal (in particular the family's income!). The estimated constant is no longer significant at the 5 per cent level. There is a fair chance that the estimate of 331.3 only by chance was larger than zero.

Example 2

Satisfaction with the current marriage

We have measurements among 800 married women for the following variables:

- marriage satisfaction (dependent variable; 1 = very dissatisfied, ..., 10 = very satisfied)
- income (1000 NOK)
- husband unemployed? (1 = yes, 0 = no)
- has the couple one or more children aged less than 2 years? (1 = yes, 0 = no)
- has the woman education at medium level? (1 = yes, 0 = no)
- has the woman higher education? (1 = yes, 0 = no)
- marriage duration (years)
- age of the woman (years)
- attitude with respect to emancipation (1 = traditional, ..., 4 = radical)

Excel gave the following estimation results, see next page:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.218
R Square	0.048
Adjusted R Square	0.038
Standard Error	5.719
Observations	800

NB. Adjusted R Square:

$$1-\overline{R}^2 = \frac{n-1}{n-k-1}(1-R^2)$$

Estimation results:

	Coefficients	*Standard Errors*	*t Stat*	*Lower 95%*	*Upper 95%*
Intercept	14.180	1.882	7.533	10.485	17.874
Variable 1 (income)	0.023	0.019	1.223	-0.014	0.061
Variable 2 (husb.unempl?)	-1.125	1.611	-0.699	-4.288	2.037
Variable 3 (child under 2?)	-0.705	0.553	-1.275	-1.79	0.38
Variable 4 (medium ed.?)	-1.266	0.554	-2.285	-2.353	-0.178
Variable 5 (higher ed.?)	-0.993	0.715	-1.388	-2.397	0.411
Variable 6 (marr. duration)	0.004	0.076	0.059	-0.144	0.153
Variable 7 (age)	-0.048	0.072	-0.675	-0.189	0.092
Variable 8 (emancipation)	-1.148	0.223	-5.141	-1.586	-0.709

Note that the estimate for the constant ("Intercept") is highly significant (t-value = 7.5), as is the estimate for the emancipation variable (t-value = -5.1). Also the effect of medium education is significant at the 5 per cent level (t-value = -2.3), but none of the others are.
The last two columns, labeled as "Lower 95%" and "Upper 95%", give the upper and lower bound of the 95 per cent confidence interval for each parameter (constant, independent variable). A 95 per cent confidence interval for a certain parameter is an interval of which we can assume that it covers the real (but unknown) value of that parameter with 95 per cent probability. For example, the parameter for emancipation has confidence interval equal to [-1.586, -0.709]. This means that we can assume that the real coefficient for emancipation lies with 95 per cent probability between -1.586 and -0.709. We estimated that coefficient as being -1.148 (column "Coefficients"), but because that estimate is a random variable, we cannot be certain about its value. The real value of that coefficient may be different. The confidence interval tells us how much different.
Both the lower bound and the upper bound of the 95 per cent confidence interval for the emancipation coefficient are negative. Hence the value of zero lies outside the interval. This is

reassuring: although we cannot be sure about the effect of emancipation on marriage satisfaction, we know (at least with 97.5 per cent probability) that it is negative: the higher the score on the emancipation variable, that is. The more radical the woman's attitude, the less satisfied she is with her marriage, all other things being equal.

An estimate which is significant at the 5 % level, i.e. which has a t-value larger than 2 in absolute value, will have a corresponding confidence interval for which the lower bound has the same sign as the upper bound. In other words, zero is not contained in the confidence interval. In the example above this is the case for the constant, for medium education, and for emancipation. Indeed, the t-values for these three estimates are larger than 2 in absolute value. Confidence intervals for the other estimates include zero. None of those are significant at the 5 per cent level.

Thus: when the upper and the lower bound of the confidence interval have the same sign, the corresponding estimate is significant.

Dummy variables

Variables that can only attain two values: 0 and 1. Are used to describe a "yes/no-situation" for example "husband unemployed?" in the previous example. The coefficient that is estimated for such a dummy variable describes the effect of being in (or entering) the "yes-situation", compared to being in (entering) the "no-situation. For example: when the husband is (or becomes) unemployed, marriage satisfaction decreases by 1.13 points (but not significant).

Dummy variables are also used in regression analysis to model the effect of an independent ordinal or nominal variable. An example is the ordinal variable "educational level" (low – medium – high educational level, 3 categories). In this case we cannot just define 3 categories 1, 2, and 3 for this variable. The reason is that it is unclear whether the distance between low and medium educational level is the same as that between medium and high. The distance between categories in such an ordinal variable is undefined – we only know the order of the categories. Even less information when the independent variable is a nominal variable without a logical ordering, such as hair colour (fair, dark, brown, red) or continent (Europe, Asia, America, Africa, Oceania). In case we have an ordinal or a nominal variable with k categories (k = 3 for educational level here), we use k-1 dummy variables (here 2). Each dummy variable except one corresponds to one level of the variable. The level that is not represented

by the dummy variable is called the reference level. Each dummy variable defines whether or not the individual involved has the corresponding characteristic or not.

For example medium educational level 1 or 0

high educational level 1 or 0

A person with high educational level scores as follows on the two dummy variables:

dummy medium educational level 0

dummy high educational level 1

and vice versa for a person with medium educational level. A person who scores a zero on both dummy variables belongs to the reference group: by implication this person has low educational level (because he has neither medium nor high).

Since the categories are mutually exclusive (that is, a person can belong to one and only one category of the ordinal or nominal variable), it is sufficient with (k-1) dummy variables.

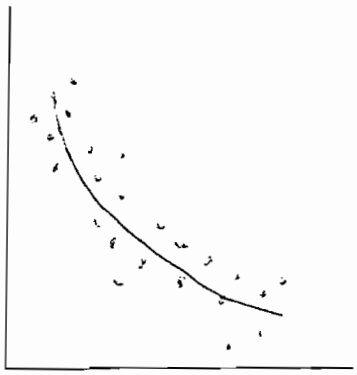
The consequence of such a situation with a number of dummy variables that describe the effect of one nominal or ordinal independent variable is that each dummy's coefficient only tells us how much the dependent variable is affected compared to the reference level.

Example: A woman with medium educational level has 1.26 points lower marriage satisfaction than a woman with the reference educational level, that is, the low educational level. The estimate -1.26 as such has no meaning, only in relation to the reference level.

NB Dummy variables are used in ordinary regression solely as independent variables. A model in which the dependent variable is a dummy (for example employed/unemployed) is much more complicated than the relatively simple models presented here. Logistic regression (logit models) – later.

Transformations

Instead of a linear relationship between X and Y, the scattergram might indicate that a non-linear relationship would give a better fit to the data.



By transforming the dependent or the independent variable, we can nonetheless find a linear relationship between transformed variables, and use ordinary Least Squares.

Transformations that are often used are:

- logarithmic transformation for the independent variable

  $Y = a + b.\log(Y)$

  Linear relationship between log(X) and Y

- logarithmic transformation for the dependent variable

  $\log(Y) = a + bX$, or $Y = \exp(a + bX)$

  often used in when the dependent variable is strictly positive, such as in rate regression

- exponential transformation of the independent variable : $Y = a + b.\exp(X)$
- quadratic transformation of the independent variable: $Y = a + b(X^2)$
- inverse transformation of the independent variable: $Y = a + b/X$.

Maximum likelihood estimation – an intuitive explanation

In certain cases, the Method of Least Squares for estimating the parameters on the basis of the data has its drawbacks. A method which can be used more generally is the method of Maximum Likelihood – ML-method. Maximum Likelihood estimation can be used as soon as we are able to find an expression for the probability distribution of the dependent variable Y, given the model (and this may be difficult for some models).

Assume that we know the relationship (linear or non-linear) between Y and X, with a known parameter b. For example: $Y = 0.17X$. In general: $Y = Y(X;b)$. When we draw a sample, Y_i becomes a random variable with a certain probability distribution.

For example $Y_i = 0.17X_i + e_i$ e_i is the residual.

Assume that we can write down the expression for the probability that we observe the value Y_i for person i, given X_i and b. Write this probability as $p(Y_i;X_i,b)$. This probability can be interpreted as "the probability for the data point (X_i,Y_i), given the relationship Y(X;b)". This probability holds for person nr. i only. For all n persons in the sample, the probability for all data points equals

(*) $p(Y_1;X_1,b).p(Y_2;X_2,b).p(Y_3;X_3,b)..... (Y_n;X_n,b).$

Now back to reality. We do not know b. All we have, are the data (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_n), plus an assumption about the relationship Y(X;b) between Y and X. When b varies,

given the data, so does the probability (*). The method of Maximum Likelihood estimation involves that we choose that particular value of b which maximizes the probability (*). In other words, choose b such that $p(Y_1;X_1,b).p(Y_2;X_2,b).p(Y_3;X_3,b)..... (Y_n;X_n,b)$ is at its maximum.

Main idea behind ML : Given the data points (X_i,Y_i) (i=1,2,3,…,n) and given the model Y=Y(X;b), the Maximum Likelihood estimate of b is that b-value that results in maximum probability for observing the actual data set.
In other words, when we consider different values as possible b-estimates, and all of these can in principle be the value of the unknown parameter b, then we choose that b-value that results in highest probability for the actually observed sample.

Individual data versus aggregate data
In the examples so far we used the individual as the unit of analysis. The analysis was performed at the individual level – our data were individual data.

A different type of analysis is that for aggregate data.

Example 1: consider the Crude Death Rate for Norway in the years 1950, 1951, …, 2000. Is there a relationship with health expenditures in the years 1950-2000?.

Y_t: Crude Death Rate in year t

X_t: health expenditure in year t.

Assume that the relationship is $Y_t = a + bX_t + e_t$. An obvious question now is: is b less than zero?
This type of question is answered by doing a time-series analysis for the years 1950-2000. "Year" is the unit of analysis.

Example 2: Y_f is the death rate in county f in the year 2000 (f=1,2,…19), and X_f are the health expenditures in county f in that year. Can we still assume that $Y_f = a + bX_f + e_f$, where a and b are the same as those in the time series analysis? This latter type of analysis, with aggregate data for one point in time but for different regions is called cross-sectional analysis. The unit of analysis is "county".