Linear regression model

π

Models in general

π

Introduction

Model - representation of some phenomenon

Non-math/stats models

Introduction

What is math/stats model?

1. **Often describes relationship between variables**
2. Types:

a) **Deterministic model** (no randomness)

b) **Probabilistic model** (with randomness)

Introduction

Deterministic models

1. Hypothesize exact relationships
2. Suitable when predicting error is negligible
3. Example: body mass index (BMI) is a measure of body fat based:

Metric formula: $\text{BMI} = \dfrac{\text{Weight in kilograms}}{(\text{Height in meters})^2}$

Non-metric formula: $\text{BMI} = \dfrac{\text{Weight in pounds x 703}}{(\text{Height in inches})^2}$

Introduction

Probabilistic models

1. Hypothesize two components:
   a) Deterministic
   b) Random error
2. Example: systolic blood pressure of newborns is 6 times the age in days + random error

$$\text{SBP} = 6 \text{ x age(d)} + \varepsilon$$

Random error may be due to factors other than age in days (for example birth weight)

Introduction

Types of probabilistic models

- Probabilistic models
  - Regression models
  - Correlation models
  - Other models

Regression models

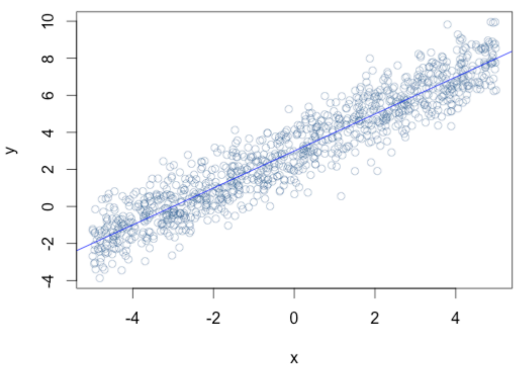
Introduction

Regression analysis is perhaps the most widely used method for the analysis of dependence – that is, for examining the **relationship between** a **set of independent variables** (X's) and **a single dependent variable** (Y).

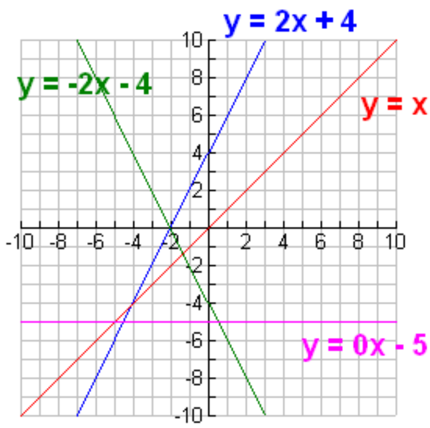
Regression (in general) is a linear combination of independent variables that corresponds as closely as possible to the dependent variable.

Regression models are used for purposes of description, inference and prediction.

Regression model vs mathematic function (model)



Linear regression model



Some linear mathematical functions

Introduction

1. **Description** – how can we describe the relationship between the dependent variable and the independent variables? How strong is the relationship captured by the model?
2. **Inference** – is the relationship described by the model statistically significant (i.e., is this level of association between the fitted values and the actual values likely to be the result of chance alone?) Which independent variables are most important?
3. **Prediction –** how well does the model generalize the observations outside the sample?

How do we construct regression model?

1. Research problem
2. Independent variable selection
3. Regression model formulation
4. Estimation
5. Verification
6. Prediction / Application

Model specification

Is based on theory:

1. Theory of field (economic, epidemiology, etc.)
2. Mathematical theory
3. Previous research
4. „Common sense”

Variable selection for the model

We have to choose independent variables for our model. In general two approaches are proposed:

1. **Substantive approach** – we choose variables according to some theory, experts, former regression models, etc.
2. **Substantive – formal approach** – first we use substantive approach to build a list of possible variables, then we use some formal approaches to select best of them

Formal approaches for variable selection

Some (it's not a complete list) of the formal approaches:

1. Coefficient of variation:

$$V_j = \frac{s_j}{\bar{x}_j}$$

We calculate this coefficient for each variable. Then some critical value V* is set (usually V* = 0.1). If the variable j has its coefficient grater than V* it can be in the model, otherwise it should not be considered in the model

Formal approaches for variable selection

2. Hellwig's method

Three steps:

a) Number of combinations: 2^{m-1}

b) Individual capacity of every independent variable in the combination:

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

c) Integral capacity of information for every combination*:*

$$H_k = \sum h_{kj}$$

Formal approaches for variable selection

2. Hellwig's method

The main goal of Hellwig's method is to choose independent variables that are highly correlated with dependent variable (they provide information) but which are not highly correlated with other independent variables (we do not repeat the same information)

a) In Hellwig's method the number of combination is provided by the formula 2^{m-1}, where m – is the number of independent variables

Formal approaches for variable selection

2. Hellwig's method

b) Individual capacity of each independent variable in the combination is given by the formula:

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

h_{kj} – individual capacity of information for *j*-th variable in *k*-th combination

r_{0j} – correlation coefficient between *j*-th variable (independent) and dependent variable

r_{ij} – correlation coefficient between *i*-th and *j*-th variable (both independent)

Formal approaches for variable selection

2. Hellwig's method

c) Integral capacity of information for every combination

The next step is to calculate Hk – integral capacity of information for each combination as the sum of individual capacities of information within each combination:

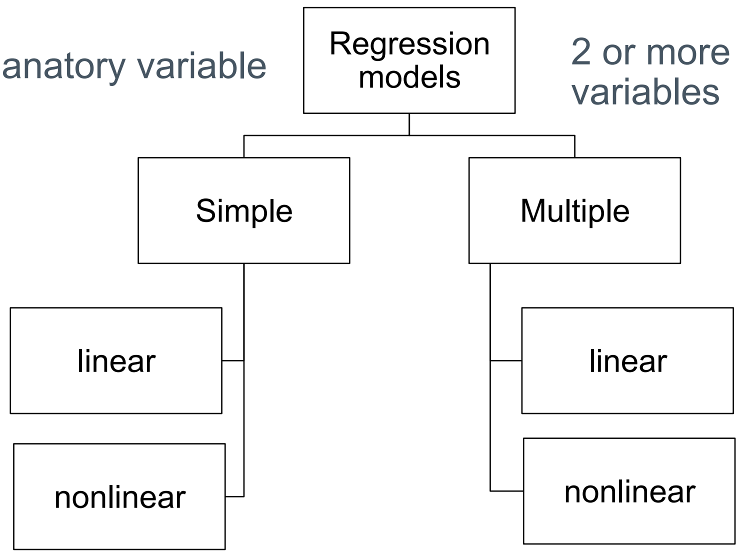
$$H_k = \sum h_{kj}$$

We choose the variables from the combination with the highest integral capacity of information

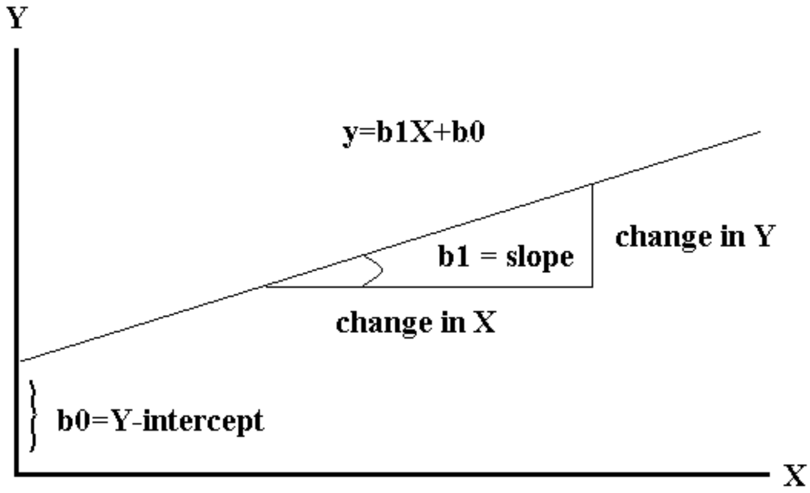
Types of regression models

1 explanatory variable

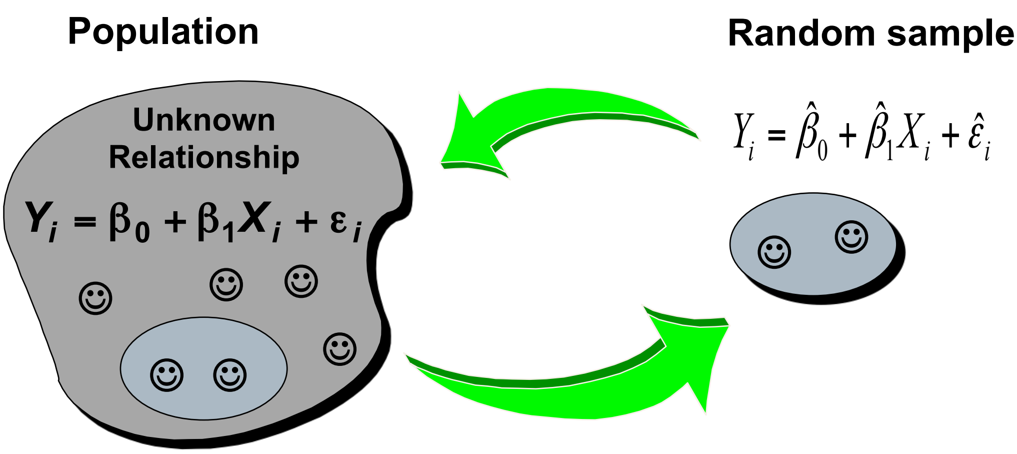
2 or more explanatory variables



Linear equations – simple



Population and sample regression models



Regression model in general

The basic formulation is:

$$Y_1 = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m + \varepsilon_1$$

where:

Y_1 – is dependent (response) variable

$X_1,\ldots,X_m$ – are independent (explanatory) variables

b_0 – is called intercept

$b_1,\ldots,b_m$ – are called coefficients

ε_1 – random error

Why we assume errors?

› To capture the fact that our expectations are not perfectly accurate, we introduce a random error to reflect the difference between the actual value of the dependent variable and our expectations

› It is almost impossible to take into account all variables that have significant influence on our dependent variable

› Variables in the model can have errors in measurement

› The mathematic formula we use may not really reflect the real data

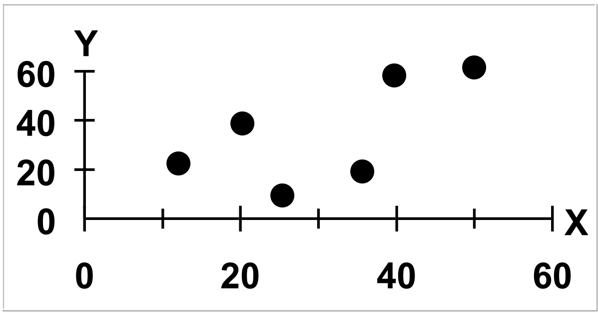
› Some phenomenons (i.e. economic) have some randomness in their nature

Linear regression model – assumptions

1. Linearity of the phenomenon measured
2. Independent variables are independent, so none of them is a linear combinations made from any of them
3. Independent variables are not random
4. Constant variance of the error terms
5. Independence of the error terms
6. Normality of the error term distribution

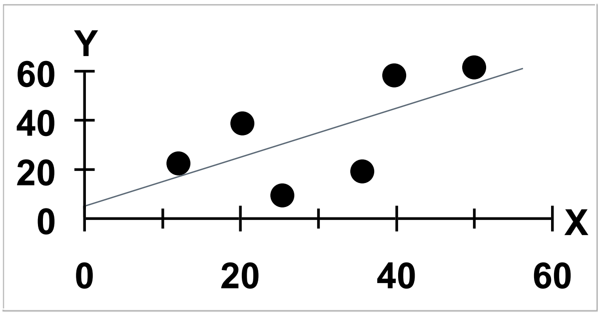
Linear regression model Scatter plot

1. Plot of all (X_i, Y_i) pairs
2. Suggests how well model will fit



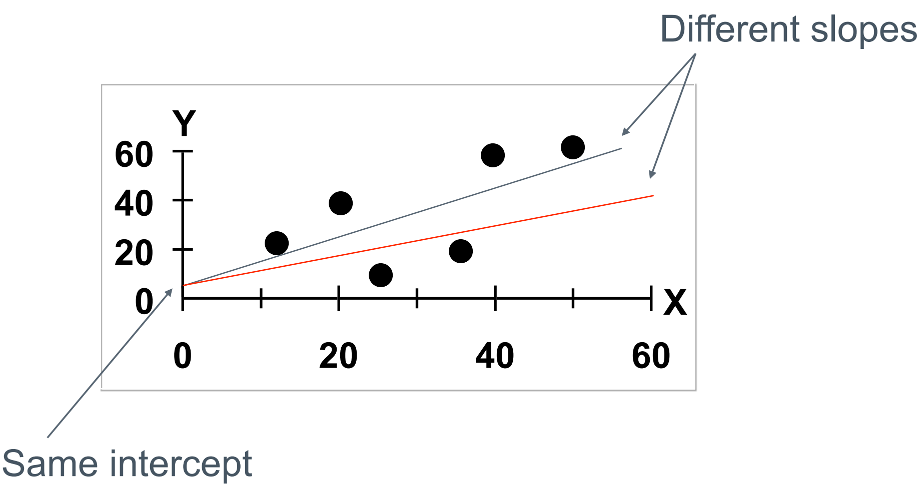
Challange

How to draw a line through the points? How to determine which line „fits best”?



Challange

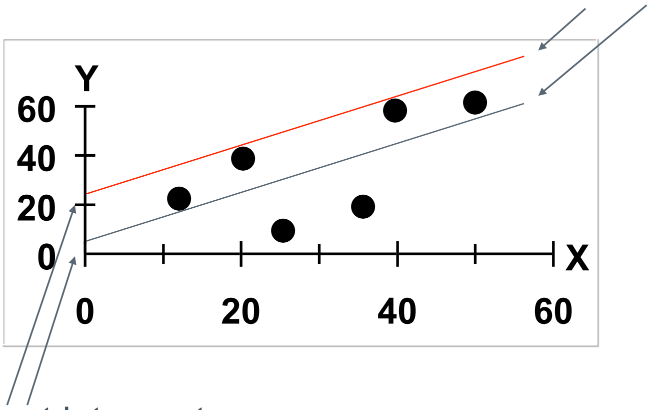
How to draw a line through the points? How to determine which line „fits best”?



Challange

How to draw a line through the points? How to determine which line „fits best”?

Same slopes

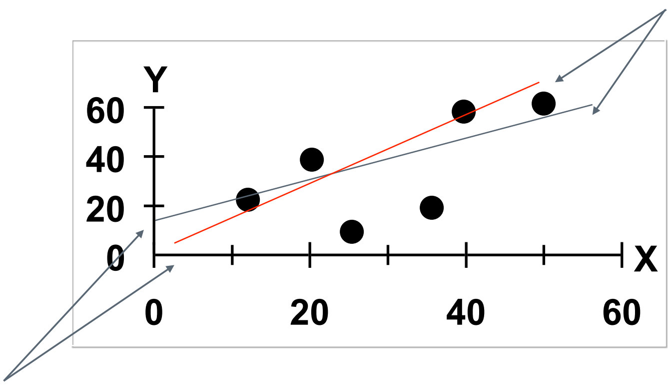


Different intercepts

Challange

How to draw a line through the points? How to determine which line „fits best”?

Different slopes



Different intercepts

Ordinary Least Squares (OLS)

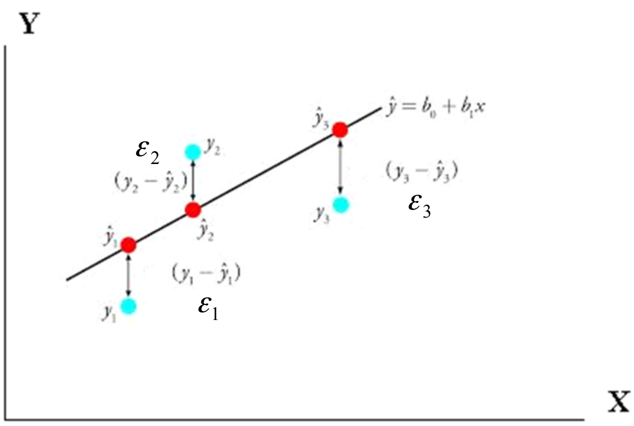
„Best fit” means differences between actual Y values and predicted Y values, that are minimum. But positive differences off-set negative ones – **square errors**

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

OLS minimizes the sum of the squared differences (errors) (SSE)

Ordinary Least Squares (OLS)

OLS minimizes: $\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2$



Ordinary Least Squares (OLS)

The objective function in the matrix form:

$$\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)=\min$$

To solve this problem, one differentiates of this expression with respect to the elements $\hat{\mathbf{b}}$, sets it equal to zero, and solves the result (known as the first-order conditions) for $\hat{\mathbf{b}}$. The first-order conditions are given by:

$$2\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)=0$$

Expanding the expression yields:

$$\mathbf{2X'y}-\mathbf{2X'X}\hat{\mathbf{b}}=0$$

Dividing through 2 and solving for $\hat{\mathbf{b}}$ yields:

$$\hat{\mathbf{b}}=(\mathbf{X'X})^{-1}\mathbf{X'y}$$

only if: $\det(\mathbf{X'X})\neq 0$

Ordinary Least Squares (OLS)

The variance of the residuals are estimated as follows:

$$S_e^2=\frac{\sum_{i=1}^{n}e_i^2}{n-m-1}=\frac{1}{n-m-1}\left(\mathbf{y'y}-\mathbf{b'X'y}\right)$$

where: *n* – number of observations, *m* – number of independent variables

Variance-covariance matrix is calculated as follows:

$$\mathbf{D}^2(\hat{\mathbf{b}})=S_e^2\left(\mathbf{X'X}\right)^{-1}$$

On the main diagonal we have standard squared errors for the estimates.

Ordinary Least Squares (OLS) – simple example

Let's assume we have the dependent variable Y (let's assume it is the production of simple chairs in thousands) and two independent variables (resources) X1 (wood usage in cubic cm) and X2 (man-hour used):

Y	X1	X2
2,3	1	2
3,4	1,5	3
4,5	3	4
5	2,5	5
6,2	3	6
8	3,5	7
9,1	4	8
10,1	4,5	9

Ordinary Least Squares (OLS) – simple example

The elements of: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Matrix **X**:

1	1	2
1	1,5	3
1	3	4
1	2,5	5
1	3	6
1	3,5	7
1	4	8
1	4,5	9

Matrix **Y**:

2,3
3,4
4,5
5
6,2
8
9,1
10,1

Matrix **X**$'$:

1	1	1	1	1	1	1	1
1	1,5	3	2,5	3	3,5	4	4,5
2	3	4	5	6	7	8	9

The vector of 1's in the first column of **X** corresponds to a dummy variable that is multiplied by the intercept term

Ordinary Least Squares (OLS) – simple example

Let's calculate elements of: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Matrix $\mathbf{X}'\mathbf{X}$:

8	23	44
23	76	146
44	146	284

Matrix $(\mathbf{X}'\mathbf{X})^{-1}$:

0,9710	-0,3913	0,0507
-0,3913	1,2174	-0,5652
0,0507	-0,5652	0,2862

Matrix $\mathbf{X}'\mathbf{y}$:

48,6
161,85
314,7

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

b0	-0,1783
b1	0,1435
b2	1,0620

So our estimated model looks like this:

$$\hat{y} = -0{,}1783 + 0{,}1435X_1 + 1{,}0620X_2$$

Ordinary Least Squares (OLS) – simple example

Model interpretation:

$$\hat{y} = -0{,}1783 + 0{,}1435X_1 + 1{,}0620X_2$$

- If we increase amount of X_1 used by a cubic cm, the production will increase by 0,1435 thousands of chairs (ceteris paribus), and vice versa
- If we increase amount of X_2 used by one man-hour, the production will increase by 1,0620 thousands of chairs (ceteris paribus), and vice versa

Ordinary Least Squares (OLS) – simple example

Let's calculate the residuals:

$$\hat{y} = -0{,}1783 + 0{,}1435X_1 + 1{,}0620X_2$$

y	X_1	X_2	$\hat{y}$	e_i	e_i^2
2,3	1	2	2,08913	0,2109	0,0445
3,4	1,5	3	3,222826	0,1772	0,0314
4,5	3	4	4,5	0,0000	0,0000
5	2,5	5	5,490217	-0,4902	0,2403
6,2	3	6	6,623913	-0,4239	0,1797
8	3,5	7	7,757609	0,2424	0,0588
9,1	4	8	8,891304	0,2087	0,0436
10,1	4,5	9	10,025	0,0750	0,0056
			Sum	**0,0000**	**0,6038**

Ordinary Least Squares (OLS) – simple example

The variance of the residuals:

$$S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} = \frac{0{,}6038}{8-2-1} = 0{,}1208$$

The variance-covariance matrix: $\mathbf{D}^2(\hat{\mathbf{b}}) = S_e^2(\mathbf{X}'\mathbf{X})^{-1}$

0,1173	-0,047	0,0061
-0,047	0,147	-0,068
0,0061	-0,068	0,0346

The errors for all estimates:

$$S(\hat{b}_0) = \sqrt{0{,}1173} = 0{,}3424$$
$$S(\hat{b}_1) = \sqrt{0{,}1470} = 0{,}3834$$
$$S(\hat{b}_2) = \sqrt{0{,}0346} = 0{,}1859$$

Ordinary Least Squares (OLS) – simple example

The interpretation of the errors for all estimates:

› when we estimate parameter b_0, if we could take many times a sample from the same population, we make mistake by ±0,3424 (b_0 = -0,1783 ±0,3424)

› when we estimate parameter b_1, if we could take many times a sample from the same population, we make mistake by ±0,3834 (b_1 = 0,1435±0,3834)

› when we estimate parameter b_2, if we could take many times a sample from the same population, we make mistake by ±0,1859 (b_2 = 1,0620±0,1859)

How good is the fit?

R^2 is the standard measure how good is the fit:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

But one drawbacks of R^2 is that whenever an independent variable is added to the model it always increases, no matter how small the contribution of that variable is.

A better solution for such problems is adjusted R^2:

$$R^2_{adj} = \bar{R}^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / (n - m - 1)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$$

Is the model significant?

To make statistical inferences about the goodness of fit of the model or the value of model parameters, we will proceed by assuming that error terms are normally distributed:

$$\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

To test the significance of the overall model we have to test the hypothesis:

$$H_0 : b_1 = b_2 = \ldots b_m = 0$$

$$H_1 : |b_1| \neq |b_2| \neq \ldots |b_m| \neq 0$$

To do this we use ratio:

$$F = \frac{\sum_i (y_i - \bar{y})^2 / m}{\sum_i (y_i - \hat{y}_i)^2 / (n - m - 1)}$$

That is distributed as *F*-statistic with (m, n-m-1) degrees of freedom

Is the model significant?

A significant *F*-statistic does not necessarily mean that all regression model parameters are different from zero. We have to test the hypothesis:

$$H_0 : b_i = 0$$

$$H_1 : b_i \neq 0$$

we use the term:

$$t = \frac{\hat{b}_i}{S\left(\hat{b}_i\right)}$$

which has a *t*-distribution with (n-m-1) degrees of freedom

Detecting problems with the model

Multicollinearity

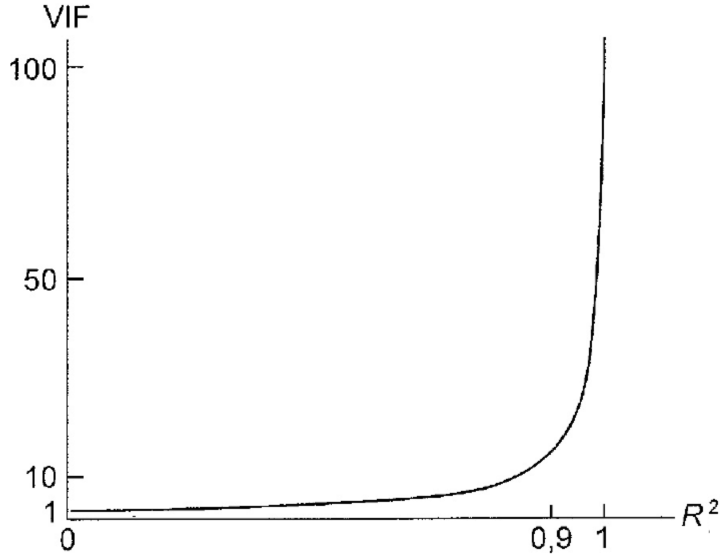
One of the value aspects of the regression is that it is able to deal with some amount of correlation among independent variables. However to much multicollinearity in the data can be a problem.

One measure of multicollinearity is the *variance inflation factor* (VIF):

$$VIF\left(\hat{b}_i\right) = \frac{1}{\left(1 - R_i^2\right)}$$

where: is the R^2 for the model where Xi is used as dependent variable, and other $X_1,\ldots,X_k$ are used as independent variables

Detecting problems with the model



Relation between VIF and R^2 for the variable

Detecting problems with the model

Multicollinearity

If we have no multicollinearity in case of one independent variable $VIF(\hat{b}_i) = 1$

$VIF(\hat{b}_i) > 10$ suggests we have the problem of multicollinearity in our model

Detecting problems with the model

Heteroscedasticity

As we assume all error terms εi have the same variance σ2. This assumption is called **homoscedasticity**. When this assumption is violated (i.e. not all variances are the same), we deal with **heteroscedasticity**.

One way do detect heteroscedasticity is the Goldfeld–Quandt test.

1. Divide the sample (size *n*) in two subsamples A (n_1 elements) and B (n_2 elements) ($n_1 + n_2 = n$). It is possible to omit some observations in the middle of the data – so $n_1 + n_2 < n$.

   We choose the subsamples arbitrary.

Detecting problems with the model

Heteroscedasticity

2. For each subsample we calculate:

$$s_1^2 = \frac{1}{n_1 - m - 1} \sum_{i \in A} e_i^2 \qquad s_2^2 = \frac{1}{n_2 - m - 1} \sum_{i \in B} e_i^2$$

3. We test the hypothesis:

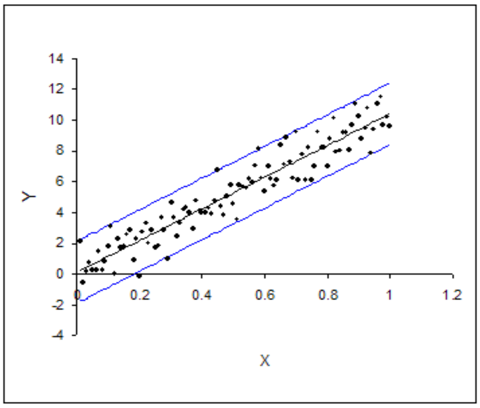
$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

following term is used:

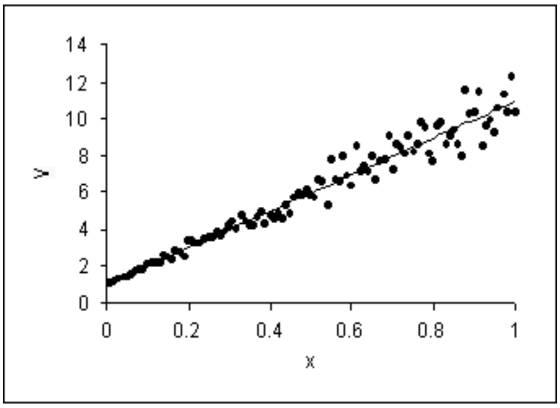
$$F_e = \frac{s_1^2}{s_2^2}$$

which has *F*-distribution with (n_1-m-1, n_2-m-1) degrees of freedom

Detecting problems with the model



Homoscedasticity in a simple, bivariate model



Heteroscedasticity in a simple, bivariate model

Detecting problems with the model

Autocorrelation

To check autocorrelation we have to check the hypothesis:

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

we use Durbin – Watson statistic (DW) in this case:

$$DW = \frac{\sum_t (e_t - e_{t-1})^2}{\sum_t e_t^2}$$

The null hypothesis tells us there is no autocorrelation, the alternative hypothesis tells us there is a positive autocorrelation

Detecting problems with the model

Normality of the residuals

To check the normality of the residuals we usually use the Shapiro-Wilk normality test:

- Rearrange the data in ascending order $x_i \leq \ldots \leq x_n$
- Calculate SS as follows: $SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$
- If n is even, let m=n/2 if n is odd let m=(n-1)/2
- Calculate b as follows, taking a_i weights from Shapiro-Wilk tables for coefficients:

$$b = \sum_{i=1}^{m} a_i (x_{n+1-i} - x_i)$$

- Calculate the statistics: $W = b^2 / SS$
- Find p-value in Shapiro-Wilk tables

Regression analysis – packages and functions of R software

Estimation of parameters and confidence intervals for them	**stats** package – **`lm`**, **`confint`** functions **car** package – **data.ellipse, confidence.ellipse**
Analysis of variance	**stats** package – **anova** function
Basic summaries	**stats** package – **summary.lm, extractAIC**
Model check	**stats** package – **influence.measures, cooks.distance, dfbeta, dfbetas, dffits, hatvalues, rstandard, rstudent**
Multicollinearity	**car** package – **vif** function, **DAAG** package – **vif** function **perturb** package – **colldiag** function
Testing linearity of the model	**lmtest** package – **harvtest** function
Normality tests	**stats** package – **shapiro.test** function, **nortest** package – **ad.test, cvm.test, lillie.test, sf, test** functions **lawstat** package – **rjb.test** function
Heteroscedasticity tests	**lmtest** package – **gqtest, bptest, hmctest** functions
Autocorrelation tests	**lmtest** package – **dwtest, bgtest** function **car** package – **durbinWatsonTest** function
Prediction	**stats** package – **predict.lm** function

Regression analysis – packages and functions of R software

ESTIMATION	
`lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)`	
`formula`	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted.
`data`	n optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called.
`model, x, y`	logicals. If TRUE the corresponding components of the fit (the model frame, the model matrix, the response, the QR decomposition) are returned
`levels`	confidence levels

Regression analysis – packages and functions of R software

ANALYSIS OF VARIANCE
`anova(object)`

BASIC MODEL SUMMARY	
`summary(object)` `extractAIC(fit, k=2)`	
`fit`	fitted model, usually the result of a fitter like lm
`k=2`	numeric specifying the 'weight' of the equivalent degrees of freedom AIC for k=2 and `BIC k=log(n)` n – numer of observations

Regression analysis – packages and functions of R software

MODEL CHECK	

`influence.measures(model); cooks.distance(model);`
`dfbeta(model); dfbetas(model); dffits(model);`
`hatvalues(model); rstandard(model); rstudent(model)`	

`influence.measures`	this suite of functions can be used to compute some of the regression (leave-one-out deletion) diagnostics for linear and generalized linear models discussed in Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), etc.
`cooks.distance`	
`dfbeta`	
`dfbetas`	
`dffits`	
`hatvalues`	
`rstandard`	
`rstudent`	
`model`	an R object, typically returned by lm or glm

Regression analysis – packages and functions of R software

Multicollinearity	
`vif(mod); vif(obj, digits=5)` `colldiag(mod, scale=TRUE, add.intercept=TRUE)`	
`mod, obj`	lm-like objects
`digits`	number of digits
`scale`	if FALSE, the data are left unscaled, TRUE is the default
`add.intercept`	if TRUE intercept is added

Testing linearity	
`harvtest(formula, order.by=NULL)`	
`formula`	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted.
`order.by`	Either a vector z or a formula with a single explanatory variable like ~ z. The observations in the model are ordered by the size of z. If set to NULL the observations are assumed to be ordered (e.g., a time series).

Regression analysis – packages and functions of R software

Normality tests	
`shapiro.test(x); ad.test(x); cvm.test(x); lillie.test(x); sf.test(x); rjb.test(x, option=c(”RJB”, ”JB”), crit.values=c(”chisq.approximation”, ”empirical”), N=0)`	
`x`	residuals
`crit.values`	a character string specifying how the critical values should be obtained, i.e. approximated by the chisq-distribution (default) or empirically
`option`	the choice of the test must be "RJB" (default, robust Jaque-Bera test) or "JB" (Jarque-Bera test)
`N`	number of Monte Carlo simulations for the empirical critical values

Regression analysis – packages and functions of R software

HETEROSCEDASTICITY	

`gqtest(formula, point=0.5, fraction=0, order.by=NULL)`
`bptest(formula, varformula=NULL)`
`hmctest(formula, point=0.5, order.by=NULL)`	

`formula`	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted
`point`	numerical. If point is smaller than 1 it is interpreted as percentages of data, i.e. n*point is taken to be the (potential) breakpoint in the variances, if n is the number of observations in the model. If point is greater than 1 it is interpreted to be the index of the breakpoint
`fraction`	numerical. The number of central observations to be omitted. If fraction is smaller than 1, it is chosen to be fraction*n if n is the number of observations in the model
`varformula`	a formula describing only the potential explanatory variables for the variance (no dependent variable needed). By default the same explanatory variables are taken as in the main regression model

Regression analysis – packages and functions of R software

AUTOCORRELATION	

`dwtest(formula, order.by=NULL, alternative=c("grater", "two.sided", "less"))`
`durbinWatsonTest(model, alternative=c("two.sided", "positive", "negative"))`
`bgtest(formula, order=1, order.by=NULL, type=c("Chisq", "F"))`	

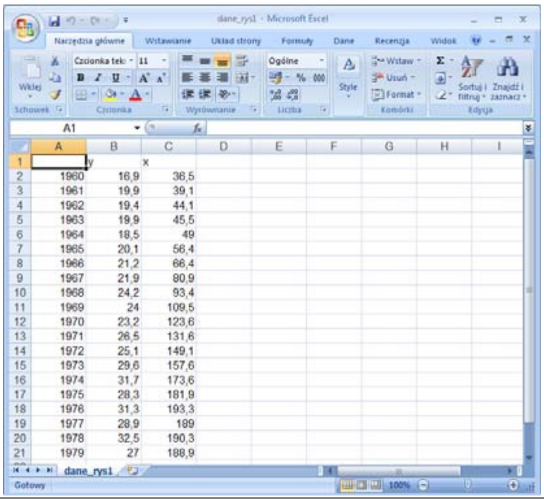
`alternative`	a character string specifying the alternative hypothesis: **grater, positive** – autocorrelation coefficient grater than 0; **less, negative** – autocorrelation coefficient less than 0; **two.sided** – autocorrelation coefficient is not zero
`order`	integer. maximal order of serial correlation to be tested
`model`	a linear model, or a vector of residuals from a linear model
`type`	the type of test statistic to be returned. Either "Chisq" for the Chi-squared test statistic or "F" for the F test statistic

Regression analysis – packages and functions of R software

PREDICTION	
`predict(object, newdata, interval="prediction", level=0.95)`	
object	object of class inheriting from "lm"
`newdata`	an optional data frame in which to look for variables with which to predict. If omitted, the fitted values are used
`interval`	type of interval calculation
`level`	tolerance/confidence level

Regression analysis – Example in R software

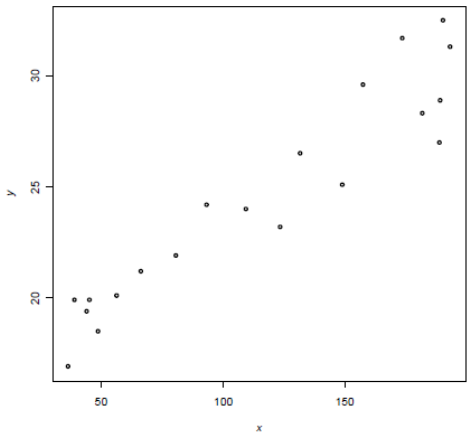
The data: wheat harvests in Poland within the years 1960-1979 (Y) depedning on the use of mineral fertilizers in kg of pure NPK (nitrogen-phosphorus-potassium) (X)



	y	x
1960	16,9	36,5
1961	19,9	39,1
1962	19,4	44,1
1963	19,9	45,5
1964	18,5	49
1965	20,1	56,4
1966	21,2	66,4
1967	21,9	80,9
1968	24,2	93,4
1969	24	109,5
1970	23,2	123,6
1971	26,5	131,6
1972	25,1	149,1
1973	29,6	157,6
1974	31,7	173,6
1975	28,3	181,9
1976	31,3	193,3
1977	28,9	189
1978	32,5	190,3
1979	27	188,9

Regression analysis – Example in R software

Scatter plot for the data**:**



$$y = b_0 + b_1 X + \varepsilon$$

The dependence between wheat harvests (Y) and the use of mineral fertilizers in kg of pure NPK seems to be linear

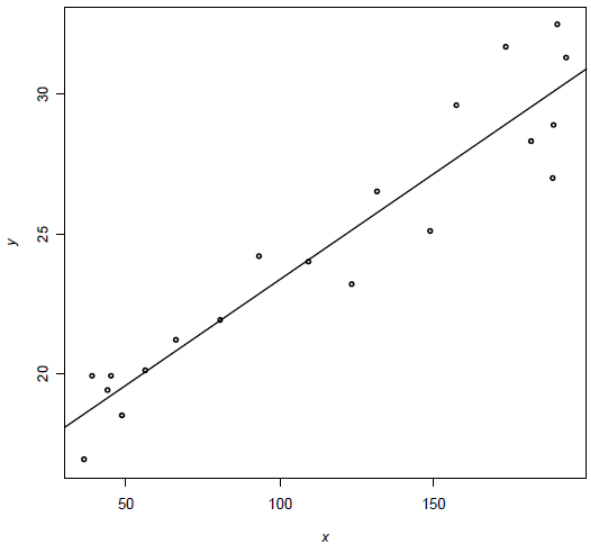
Regression analysis – Example in R software

```
[1] Estimation results
```

```
Call:
lm(formula = y ~ x, data = d, x = TRUE, y = TRUE)
Residuals:
    Min       1Q   Median       3Q      Max
-3,1063 -1,2294  0,1506  0,9316  2,7531
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15,791400   0,789077   20,01 9,53e-14 ***
x            0,075780   0,006124   12,37 3,08e-10 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 1,592 on 18 degrees of freedom
Multiple R-Squared: 0.8948,     Adjusted R-squared: 0.889
F-statistic: 153.1 on 1 and 18 DF,  p-value: 3,077e-10
```

Regression analysis – Example in R software



The estimated model and the real data

Regression analysis – Example in R software

$$\underset{(0{,}789077)}{\hat{y} = 15{,}7914} + \underset{(0{,}006124)}{0{,}07578\,x}$$

$\hat{b}_1$ = 0,07578 – by increasing (decreasing) the usage of fertilizers, will increase (decrease) the wheat production by 0,007578 q from each hectare (q = 100 kg)

$\hat{b}_0$ =15,7914 (the intercept) – anticipated wheat production without any fertilizers

$s(\hat{b}_1)$=0,006124 – when estimating b_1, if we could take sample from the same population, we make mistake ±0,006124

$s(\hat{b}_0)$=0,789077 – when estimating b_0, if we could take sample from the same population, we make mistake ±0,006124

Regression analysis – Example in R software

The **residual standard error** = 1,592 empirical values of dependent variable (wheat production in Poland) differ from theoretical values on the average 1,592 q from hectare

The **multiple R-squared** = 0,8948 – 89,48% of the dependent variable's variability was explained by the model

Adjusted R-squared = 0,889 – 88,9% of the dependent variable's variability was explained by the model

Regression analysis – Example in R software

Confidence intervals for the parameters

```
                  2,5 %      97,5 %
(Intercept) 14,13360953 17,44918997
x            0,06291333  0,08864732
```

The interval [14,134; 17,449] covers an unknown value of b_0 parameter with 95% probability

The interval [0,0629; 0,0886] covers an unknown value of b_1 parameter with 95% probability

Regression analysis – Example in R software

Analysis of variance

```
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 388,01  388,01   153,1 3,077e-10 ***
Residuals 18  45,62    2,53
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Regression analysis – Example in R software

Normality test:

```
          Shapiro-Wilk normality test
data:  model$residuals
W = 0,9798, p-value = 0,9317
```

As the α=0,05 ≤ p-value = 0,9317 there we can not decline that error is normally distributed

Regression analysis – Example in R software

Is the model significant?

T test:

```
t value  Pr(>|t|)
 20,01   9,53e-14
 12,37   3,08e-10
```

As for b_0 the α=0,05 > 9,53e-14 we have to reject the null hypothesis, it means b_0 is significantly different from zero

As for b_1 the α=0,05 > 3,08e-10 we have to reject the null hypothesis, it means b_1 is significantly different from zero.

Regression analysis – Example in R software

Is the model significant?

F Test

```
Test F
F-statistic: 153.1 on 1 and 18 DF,  p-value: 3,077e-10
```

Since α=0,05 > 3,008e-10 we have to reject the null hypothesis. b_1 is significantly different from zero. X has a significant influence on the values of y

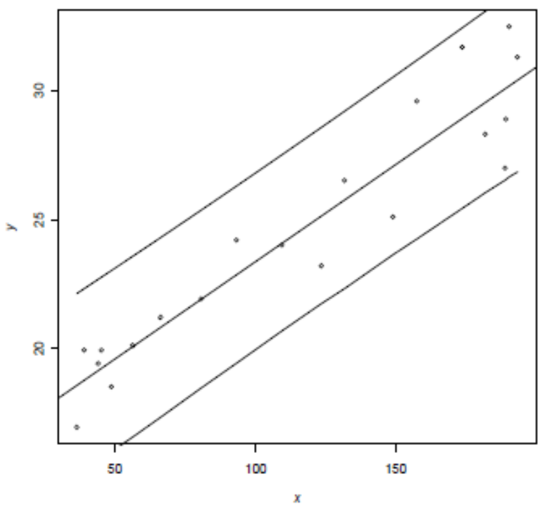
Regression analysis – Example in R software

Predicted values

```
          fit      lwr      upr
1960 18,55738 14,98451 22,13026
1961 18,75441 15,19085 22,31797
1962 19,13331 15,58684 22,67979
1963 19,23940 15,69752 22,78129
1964 19,50464 15,97385 23,03542
1965 20,06541 16,55630 23,57452
1966 20,82321 17,33948 24,30695
1967 21,92203 18,46690 25,37716
1968 22,86928 19,43086 26,30770
1969 24,08934 20,66143 27,51726
1970 25,15785 21,72887 28,58682
1971 25,76409 22,33024 29,19794
1972 27,09025 23,63506 30,54543
1973 27,73438 24,26361 31,20515
1974 28,94686 25,43767 32,45605
1975 29,57584 26,04216 33,10952
1976 30,43974 26,86748 34,01199
1977 30,11388 26,55684 33,67093
1978 30,21239 26,65084 33,77395
1979 30,10630 26,54960 33,66300
```

fit – fitted values
lwr – lower bound of the confidence interval
upr – upper bound of the confidence interval

Regression analysis – Example in R software



The model with confidence intervals

Regression analysis – Example in R software

Predicted values – new data