УНИВЕРСИТЕТ ИТМО

Methods & Models for Multivariate Data Analysis

Lecture 2. Multivariate random variables: correlation and prediction

Ass. Prof. Anna Kalyuzhnaya, PhD

УНИВЕРСИТЕТ ИТМО

Lecture 2 (1). Basic probabilistic definitions for multivariate random variable

ITMO more than a UNIVERSITY

Distribution of Multivariate Random Value

УНИВЕРСИТЕТ ИТМО

Continuous multivariate random variable (MRV) definition

MRV — Univariate RV

$$\boldsymbol{Z} = (z_1, \ldots, z_n)$$

Univariate random outcomes

$$\boldsymbol{\Xi} = (\xi_1, \ldots, \xi_n)$$

General definition of multivariate probability function:

$$F_{\Xi}(z_1, \ldots, z_n) = P(\xi_1 < z_1, \ldots, \xi_n < z_n) =$$

$$= P(\xi_1 < z_1) \cdot P(\xi_2 < z_2 | \xi_1) \cdot \ldots \cdot P(\xi_n < z_n | \xi_1, \ldots, \xi_{n-1})$$

$F_i(z_{m+1}, \ldots, z_n | z_1, \ldots, z_m)$ - **Conditional probability functions**

$F(z_1), \ldots, F(z_n)$ - **Unconditional probability functions**

Properties of MRV probability functions

УНИВЕРСИТЕТ ИТМО

$$0 \le F_{\Xi}(z_1, \dots, z_n) \le 1;$$
$$F_{\Xi}(-\infty, \dots, -\infty) = 0;$$
$$F_{\Xi}(-\infty, \dots, z_k, \dots -\infty) = 0;$$
$$F_{\Xi}(+\infty, \dots, +\infty) = 1;$$
$$F_{\Xi}(+\infty, \dots, z_k, \dots, +\infty) = F_{\xi_k}(z_k)$$

All the properties are caused by the general definition of MRV probability function

Generalization for discrete MRV distribution (in terms of distribution laws)

$$P_{\Xi}(z_1, \dots, z_n) = P(\xi_1 = z_1, \dots, \xi_n = z_n) =$$
$$= P(\xi_1 = z_1) \cdot P(\xi_2 = z_2 | \xi_1) \cdot \dots \cdot P(\xi_n = z_n | \xi_1, \dots, \xi_{n-1})$$

Unconditional moments of MRV

MRV mathematical expectation:

$$m_{\xi_i} = M[\xi_i] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} z_i f_{\Xi}(z_1, \dots, z_n) dz_1 \dots dz_n = \int_{-\infty}^{\infty} z_i f_i(z_i) dz_i$$

MRV variance:

$$D_{\xi_i} = M[(\xi_i - M[\xi_i])^2] = \int_{-\infty}^{\infty} (z_i - M[\xi_i])^2 f_i(z_i) dz_i$$

Covariance of MRV

УНИВЕРСИТЕТ ИТМО

Covariance of MRV is a measure of linear dependency of two random variables.

Probabilistic definition of covariance:

$$\mathrm{K}_{\xi_i\xi_j} = M[(\xi_i - M[\xi_i])(\xi_j - M[\xi_j])] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\xi_i - M[\xi_i])(\xi_j - M[\xi_j])\, f_{\xi_i\xi_j}(z_i, z_j)dz_i\, dz_j$$

Properties:

$$-\infty \le \mathrm{K}_{\xi_i\xi_j} \le +\infty$$

The sign of the covariance shows the tendency in the linear statistical relationship between the variables.

$$\mathrm{K}_{\xi_i\xi_j} = \mathrm{K}_{\xi_j\xi_i}$$

Covariance measure is symmetric

$$\mathrm{K}_{\xi_i\xi_i} = D_{\xi_i} \ge 0,\ \mathrm{K}_{\xi_i\xi_j} \le \sqrt{D_{\xi_i}D_{\xi_j}}$$

Covariance of URV is equal to it's variance and always non-negative

$$\mathrm{K}_{a\xi_i b\xi_j} = ab\mathrm{K}_{\xi_i\xi_j}, \mathrm{K}_{(a+\xi_i)(b+\xi_j)} = \mathrm{K}_{\xi_i\xi_j}$$

Linearity of covariance allow for useful consequences in a case of linear transformations of RV's

Conditional moments of bivariate RV

УНИВЕРСИТЕТ ИТМО

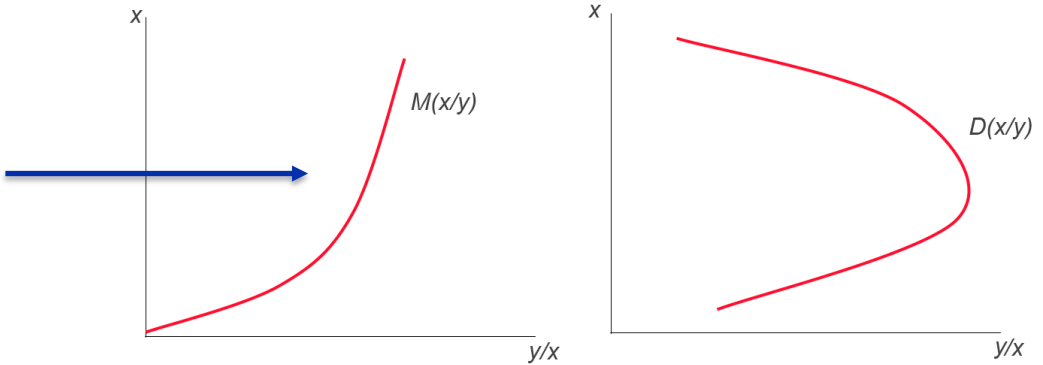
Conditional mathematical expectation – regression

$$m_{x|y}(\xi_1|\xi_2 = y) = \int_{-\infty}^{\infty} x f(x|y)dx$$

Conditional variance – (in Russian – скедастическая кривая)

$$D_{x|y}(\xi_1|\xi_2 = y) = \int_{-\infty}^{\infty} (x - m_{x|y})^2 f(x \mid y)dx$$

It can be seen that conditional ME serves as non-parametric regression



УНИВЕРСИТЕТ ИТМО

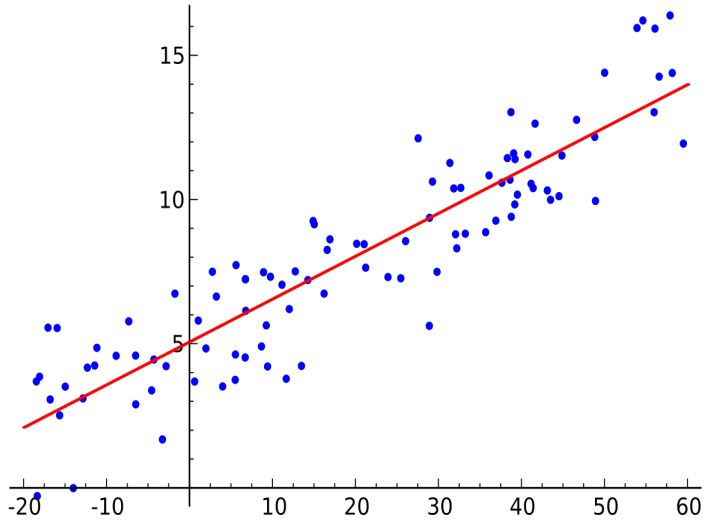
Lecture 2 (2). Multivariate random variable analysis. Statistical relations (basics)

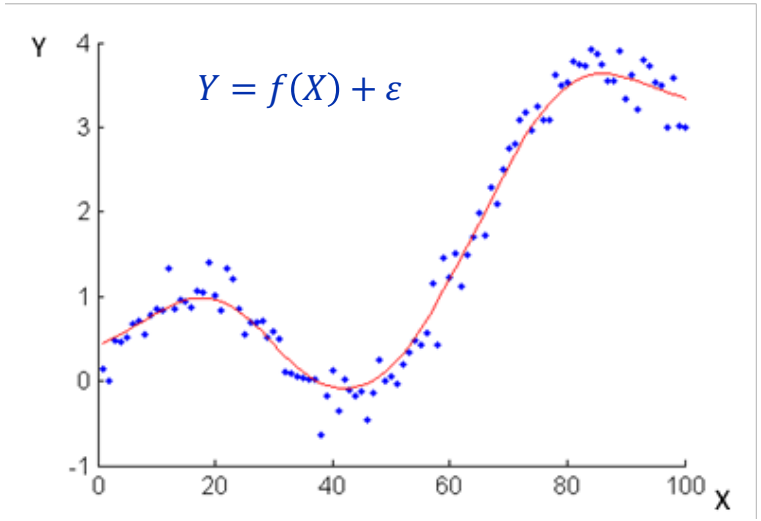
ITMO more than a UNIVERSITY

What is statistical relation?

УНИВЕРСИТЕТ ИТМО

What could we say about relation between X and Y?





How strong is the relation between X and Y?

↓

Statistical inference

Object of correlation analysis

$\hat{f}(X) = ??$

↓

Statistical prediction

Object of regression analysis

Statistical relationships

УНИВЕРСИТЕТ ИТМО

- **Correlation analysis**

On this lesson:

- ✓ **pair** coefficient of correlation
- ✓ confidence interval and significance level
- ✓ correlation matrix
- ✓ **multiple** coefficient of correlation
- ✓ determination coefficient
- ✓ **partial** coefficient of correlation

- **Regression analysis**

On this lesson:

- ✓ regression using **pair** coefficient of correlation
- ✓ multiple linear regression
- ✓ confidence interval for coefficients
- ✓ curvilinear regression
- ✓ regression paradoxes

ITMO more than a UNIVERSITY

УНИВЕРСИТЕТ ИТМО

Lecture 3 (3). Multivariate random variable analysis. Correlation analysis

ITMO more than a UNIVERSITY

Correlation analysis. Basics

УНИВЕРСИТЕТ ИТМО

Covariance of MRV

$$\mathrm{K}_{\xi_i\xi_j} = M[(\xi_i - M[\xi_i])(\xi_j - M[\xi_j])] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\xi_i - M[\xi_i])(\xi_j - M[\xi_j])\, f_{\xi_i\xi_j}(z_i, z_j)dz_i\, dz_j$$

Correlation of MRV is a measure of linear statistical dependency that is equal to normalized by product of standard deviations covariance.

$$R_{\xi_i\xi_j} = \frac{K_{\xi_i\xi_j}}{\sigma_{\xi_i}\sigma_{\xi_j}}$$

After normalization this measure changes in range [-1, 1] with the same interpretation as for covariance:

$$-1 \le R_{\xi_i\xi_j} \le 1$$

ITMOre than a UNIVERSITY

Correlation analysis. Basics

УНИВЕРСИТЕТ ИТМО

Correlation **is a statistical relationship that can show whether and how strong linear dependency between variables is.**

Pair coefficient of correlation or Pearson's product-moment coefficient ☺

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_x)(\mathrm{Y}-\mu_Y)]}{\sigma_X \sigma_Y}$$

Estimation of pair coefficient of correlation:

$$r_{xy} = \frac{\sum_{\mathrm{i}=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{\mathrm{i}=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{\mathrm{i}=1}^{n}(x_i - \bar{x})^2 \sum_{\mathrm{i}=1}^{n}(y_i - \bar{y})^2}}$$

ITMO more than a UNIVERSITY

Confidence interval and significance level

УНИВЕРСИТЕТ ИТМО

Correlation property:

If $U = cX + k_1, c \neq 0$, $V = dY + k_2, d \neq 0$ *then* $K_{XY} = K_{UV}$

Significance level:

$$S = t_{n-2,\alpha}(n - 2 + t_{n-2,\alpha}^2)^{-\frac{1}{2}}$$

Confidence interval for ρ:

$$P[r^* - t_{1-\alpha/2}\frac{\sigma_\mu}{\sqrt{n}} < \rho < r^* + t_{1-\alpha/2}\frac{\sigma_\mu}{\sqrt{n}}] \approx 1 - \alpha$$

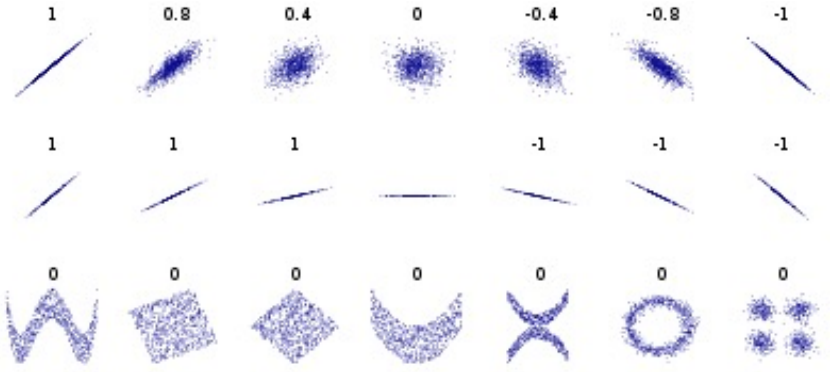
t -Student distribution

ITMO more than a UNIVERSITY

Correlation paradoxes

УНИВЕРСИТЕТ ИТМО

Correlation is only a measure of linear relationship

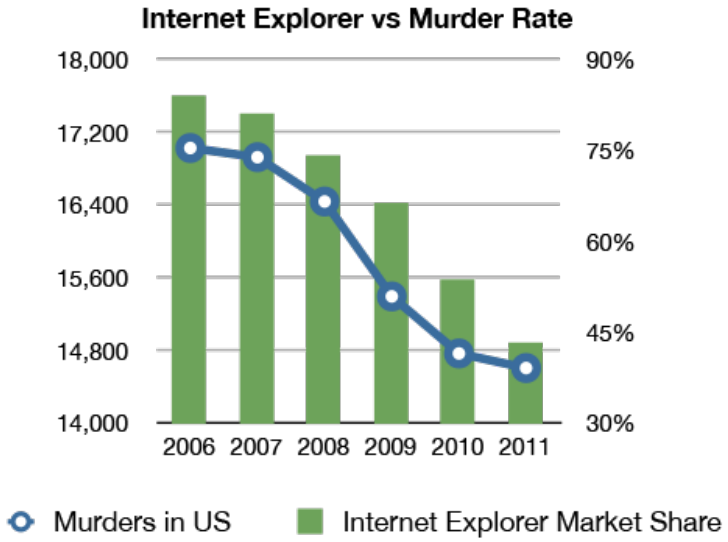


from Wikipedia

Paradox 1. If Corr(X,Y) = 0 it doesn't mean that X and Y are independent

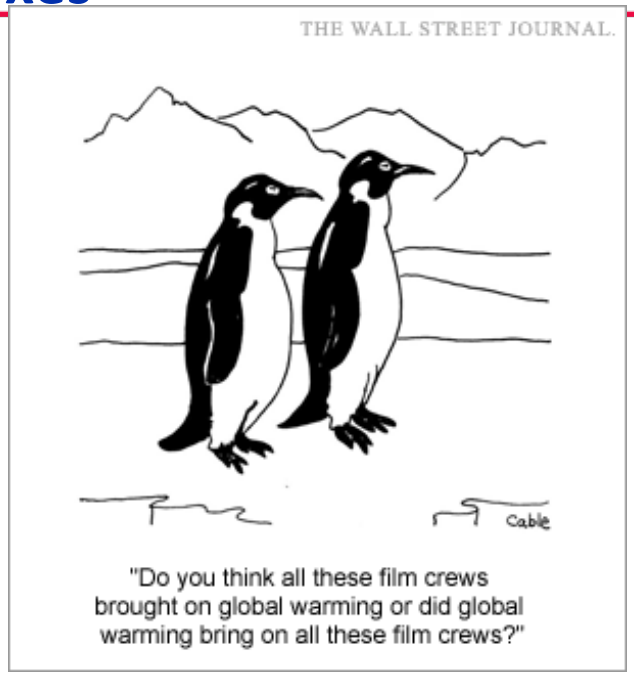
ITMOre than a UNIVERSITY

Correlation paradoxes



Paradox 2. We need to be careful in interpreting the value of correlation because correlation can be spurious

Correlation paradoxes



Paradox 3. Correlation value could not be used to say anything about cause and effect relationship.

Multivariate correlation analysis

УНИВЕРСИТЕТ ИТМО

Matrix with pair coefficients of correlation:

$$D = \begin{array}{|c|c|c|c|c|c|} \hline & y & x_1 & x_2 & \cdots & x_m \\ \hline y & 1 & r_{yx_1} & r_{yx_2} & \cdots & r_{yx_m} \\ \hline x_1 & r_{x_1y} & 1 & r_{x_1x_2} & \cdots & r_{x_1x_m} \\ \hline x_2 & r_{x_2y} & r_{x_2x_1} & 1 & \cdots & r_{x_2x_m} \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline x_m & r_{x_my} & r_{x_mx_1} & r_{x_mx_2} & \cdots & 1 \\ \hline \end{array}$$

Multiple coefficient of correlation (from 0 to 1) :

$$R_0 = \sqrt{1 - \frac{\det(Q)}{Q_{yy}}}$$

Q_{yy} - algebraic complement to yy element of D

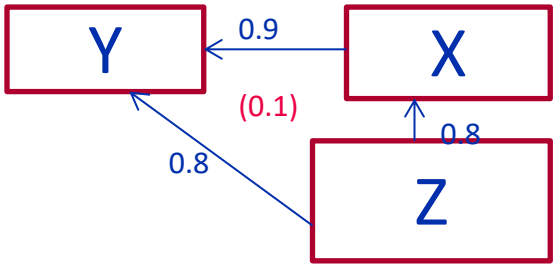
OR

$$R_0 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}}$$

Partial coefficient of correlation:

$$r_{jk \cdot 1,2,\ldots,n} = \frac{d_{jk}}{\sqrt{d_{kk} d_{jj}}}$$

$$d_{jk} = (-1)^{i+j} \mid D_{jk} \mid$$



From correlation to regression

УНИВЕРСИТЕТ ИТМО

Correlation and regression coefficients relation:

$$M(Y|X) = M(Y) + r^* \frac{\sigma_Y^*}{\sigma_X^*}\left[X - M(X)\right] =$$

$$= \left[M(Y) - r^* \frac{\sigma_Y^*}{\sigma_X^*} M(X)\right] + r^* \frac{\sigma_Y^*}{\sigma_X^*} X$$

r^* - Sample estimation of correlation coefficient

σ_Y^* - Sample estimation of standard deviation

ITMO more than a UNIVERSITY

УНИВЕРСИТЕТ ИТМО

Lecture 2 (4). Multivariate random variable analysis. Statistical prediction

ITMO more than a UNIVERSITY

Several types of regression

УНИВЕРСИТЕТ ИТМО

Types of regression:

- simple regression $y = ax + b + \varepsilon$
- multivariate simple regression $y_i = \sum_{k=1}^{n} \beta_k x_{ik} + \varepsilon_i$
- multivariate polynomial regression $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_n x^n + \varepsilon$
- quantile regression (simple, polynomial) $y_i = \sum_{k=1}^{n} \beta_k x_{pk} + \varepsilon_i,\ p = [0,1]$
- regularized regression $\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$
- logistic regression $f(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$

↓

Method for classification!

Linear regression. Basics

УНИВЕРСИТЕТ ИТМО

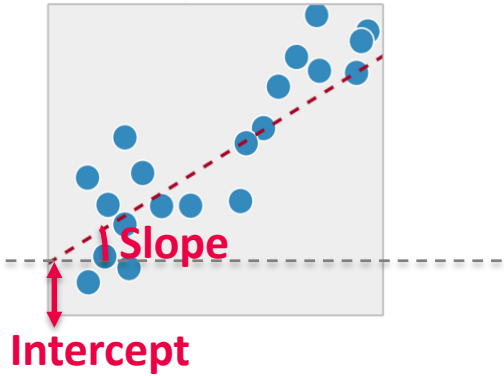
General definition: $E(Y|X) = f(X, \beta)$ — Conditional mean

Simple linear regression - assessment of linear relation between two numerical features

The essence: calculation of intercept and slope coefficients for line equation, which describes mean statistical dependency between variables optimally in terms of squared error

$$y_i = ax_i + b + \varepsilon$$

Slope — Intercept — Differences from an average trend

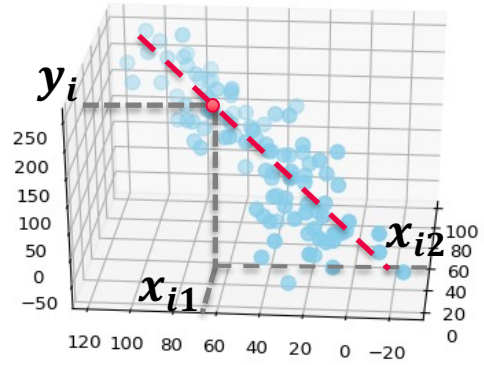


Multiple linear regression - assessment of linear dependence of target numerical feature on many other features

The essence: the same, as simple regression, but in a multidimensional equivalent

$$y_i = \sum_{k=1}^{n} \beta_k x_{ik} + \varepsilon_i$$

Multidimensional "slope"



Least Squares method (LS) (1/2)

УНИВЕРСИТЕТ ИТМО

In the regression analysis, usually, "best“ solution is understood in terms of model and target feature difference squares sum minimization :

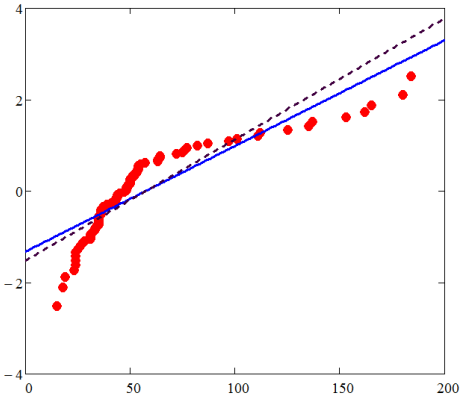
$$\sum_{i=1}^{N}\left(y_i-\sum_{k=1}^{n}\beta_k x_{ik}\right)^2 \rightarrow \min$$

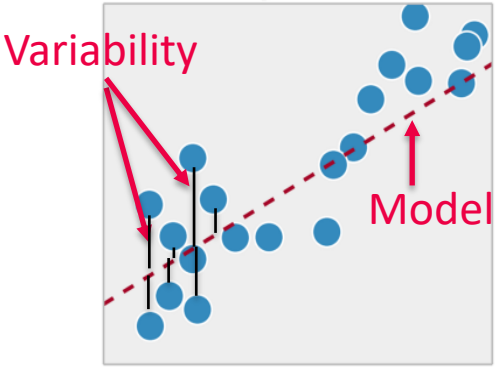
Problem of search of function minimum (reminder):

$$S(a,b)=(a+bx_1-y_1)^2+\cdots+(a+bx_n-y_n)^2$$

$$\frac{\partial S}{\partial a}=\sum_i 2(a+bx_i-y_i) \qquad \frac{\partial S}{\partial b}=\sum_i 2(a+bx_i-y_i)x_i.$$

But what if the case is not fully linear?





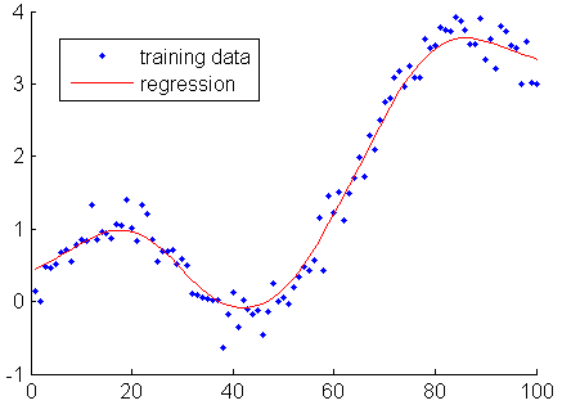
Types of non-linear regression

УНИВЕРСИТЕТ ИТМО

Polynomial regression - assessment of linear dependence between features which can be involute

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_n x^n + \varepsilon$$

BUT, in terms of numerical methods - all this still is the linear system of the equations therefore we solve it as before (LS).



Nonlinear regression - assessment of concrete non-linear dependence between features

Example: $y = \alpha\beta^x$

To solve that, it is necessary to linearize: $\ln y = \ln \alpha + \ln \beta x$

Confidence interval for coefficients

УНИВЕРСИТЕТ ИТМО

General expression (matrix form):

$$\hat{\theta}_k - t_{N-2,1-\frac{\alpha}{2}} \hat{\sigma}\left(\left[X^T X\right]^{-1}\right)_{kk} < \theta < \hat{\theta}_k + t_{N-2,1-\frac{\alpha}{2}} \hat{\sigma}\left(\left[X^T X\right]^{-1}\right)_{kk}$$

here $\hat{\sigma} = \sqrt{\frac{(Y - XB)}{n-2}}$ - deviation of error or noise

$t_{N-2,1-\frac{\alpha}{2}}$ - quantile of Students distribution

For simple linear regression it is useful to express $\hat{\sigma}\left(\left[X^T X\right]^{-1}\right)_{kk}$ **as**

$\hat{\sigma}_{b0} = \hat{\sigma}$ - for intercept coefficient

$\hat{\sigma}_{bk} = \frac{\hat{\sigma}}{\sigma_{X_k}}$ - for slope coefficient

ITMO more than a UNIVERSITY

Bayesian view on regression fitting (1/3)

The regression of the value y to x reflects the conditional average value or otherwise the estimate of the conditional mathematical expectation *y* with the fixed value *x*. I.e. by identified coefficients we try to get closer to true values $E[p(y|x)]$:

$$p(y_n|f^*(x_n,\theta^*),\sigma) = N(y_n|f^*(x_n,\theta^*),\sigma)$$

We need to find distribution parameters θ and σ. At first, we look for θ, fixing σ:

$$\theta^* = argmax \prod_{n=1}^{N} N(y_n|x_n,\theta,\sigma)$$

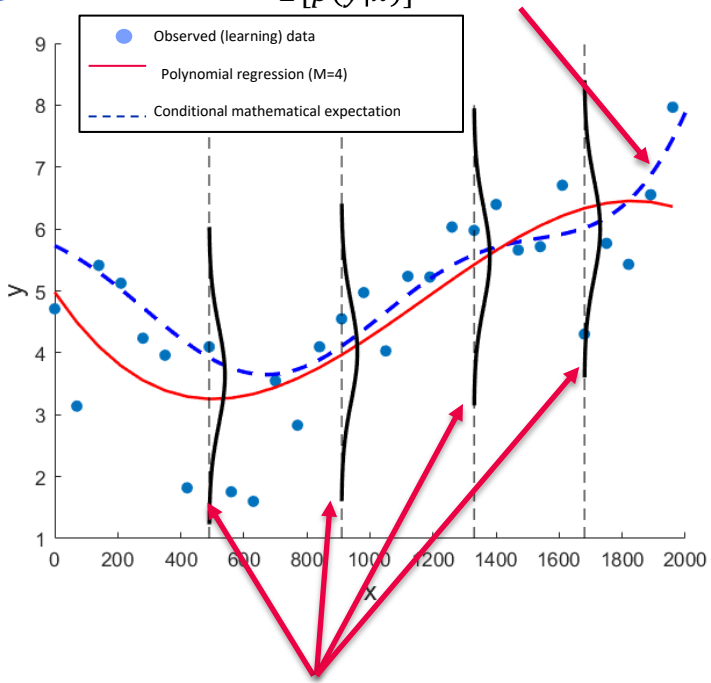
After logarithm and transform, we get:

$$E[(f(x,\theta) - y)^2] \to 0$$

Trying θ^* , we find σ^* :

$$E[(f^*(x,\theta^*) - y)^2] = \sigma^{2^*}$$

Estimation of conditional mathematical expectation $E[p(y|x)]$



Conditional distribution of probabilities $p(y|x)$

Bayesian view on regression fitting (2/3)

However, the usage of conditional distributions does not yet make the setting Bayesian. The applied meaning of Bayes theorem is the ability to use an a priori distribution of a hidden variable.

Theoretical derivation:

$$p(x,c) = p(c)p(x|c) = p(x)p(c|x) \Rightarrow p(c|x) = \frac{p(c)p(x|c)}{\sum_k p(c_k)p(x|c_k)}$$

A priori distribution of the latent variable

Posteriori distribution of the latent variable

Means normalization of conditional distribution

For our task θ is hidden variable, for which you can introduce an a priori distribution $p(\theta)$:

$$p(\theta|x,y,\sigma) \propto p(y|x,\theta,\sigma)p(\theta)$$

Then the predictive model can be described in the probabilistic form (taking into account uncertainty):

$$p(y|x,\boldsymbol{x},\boldsymbol{y}) = \int p(y|x,\theta)p(\theta|x,y)\mathrm{d}\theta$$

Since all conditional distributions are Gaussian, the probability model of prediction is also Gaussian:

$$p(y|x,\boldsymbol{x},\boldsymbol{y}) = N(y|\mu(x),s^2(x))$$

Bayesian view on regression fitting (3/3)

УНИВЕРСИТЕТ ИТМО

The Bayesian approach makes it possible to construct a probabilistic prediction in which uncertainty is taken into account by learning sample and a priori distribution of the parameters of the polynomial model.

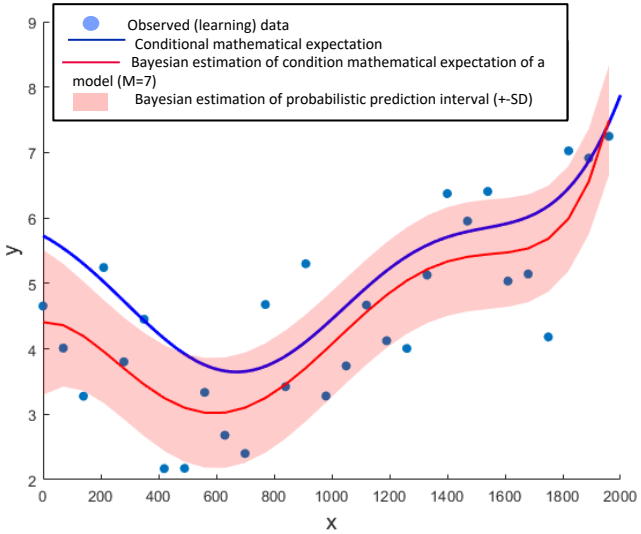
Assessment of conditional mathematical expectation and variance in matrix form (for brevity):

$$\mu^* = \sigma^{-2^*} \boldsymbol{X}^T \boldsymbol{S} \boldsymbol{X} y \qquad s^{2^*} = \sigma^{2^*} + \boldsymbol{X}^T \boldsymbol{S} \boldsymbol{X}$$

The matrix $\boldsymbol{S}^{-1}$ contains references to two variances: $1/\alpha$ (variance of a priori distribution of regression parameters) and σ^2 (regression model error variance):

$$\boldsymbol{S}^{-1} = \alpha \boldsymbol{I} + \sigma^{-2^*} \boldsymbol{X} \boldsymbol{X}^T$$

Bayesian probabilistic interval(+- standard deviation):

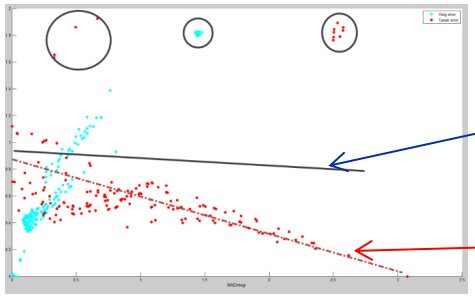


Robustness of regression

УНИВЕРСИТЕТ ИТМО

Both for x- and y axis: for predictor variables and for resulting variable (outlying cases)

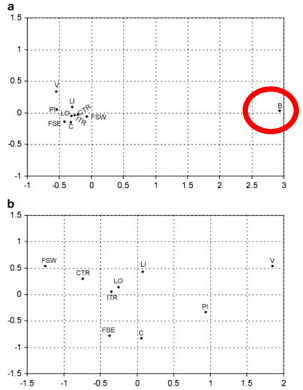
For example



Regression line taking into account “suspicious” values

Correct regression line

For example



Simple regression is not robust for non-homogenous data

To improve robustness of regression model or to give the bigger weight to distribution tails we can use quantile regression.

The same as ordinary regression but uses quantiles instead of sample values

$$y_i = \sum_{p=1}^{n} \beta_p x_p + \varepsilon_i$$

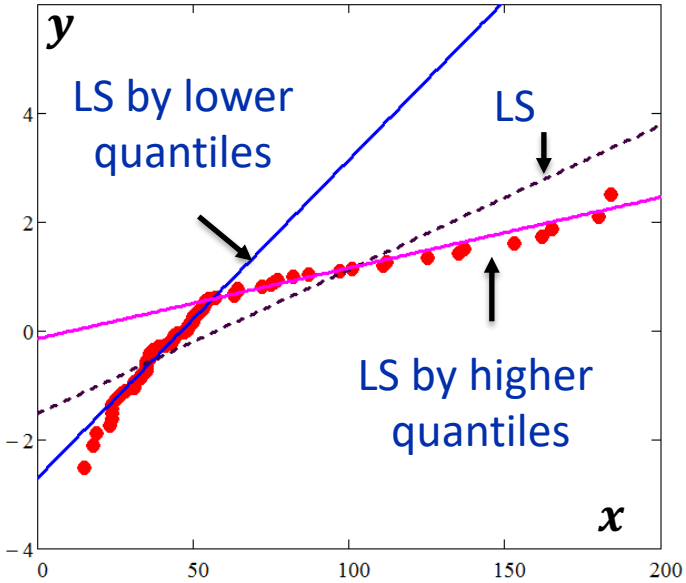
Quantiles

Least Squares method (LS) (2/2)

УНИВЕРСИТЕТ ИТМО

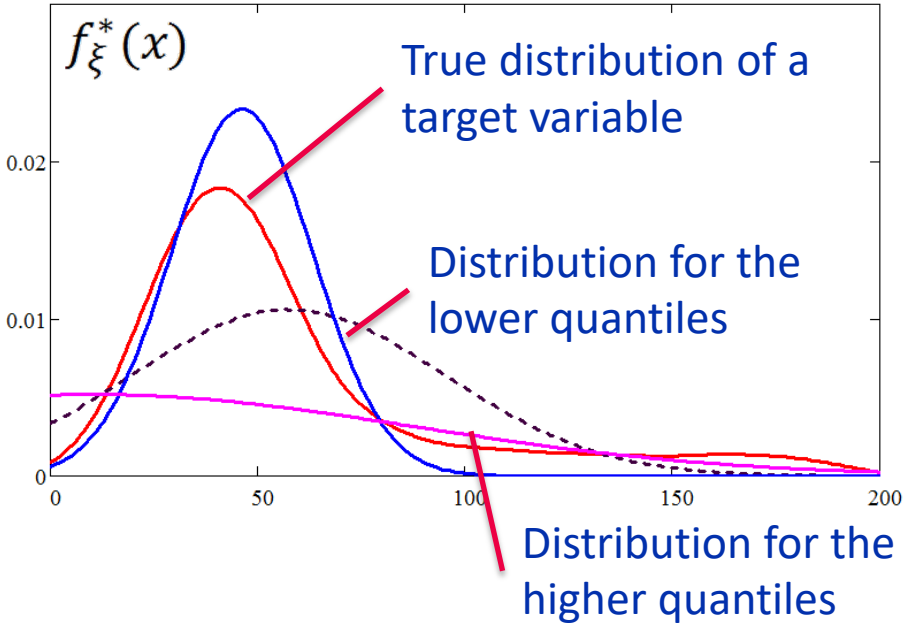
How we can use possibilities of quantile regression?

The relation between x and y is not linear



Replacement of selected values with quantiles: $x_i \rightarrow x_p$

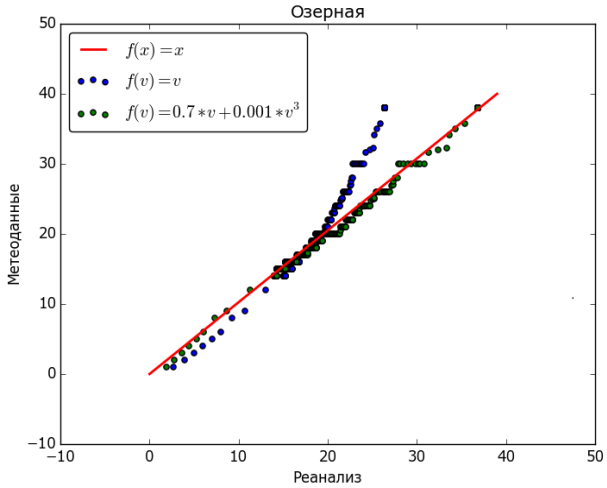
The connection between regression and distributions

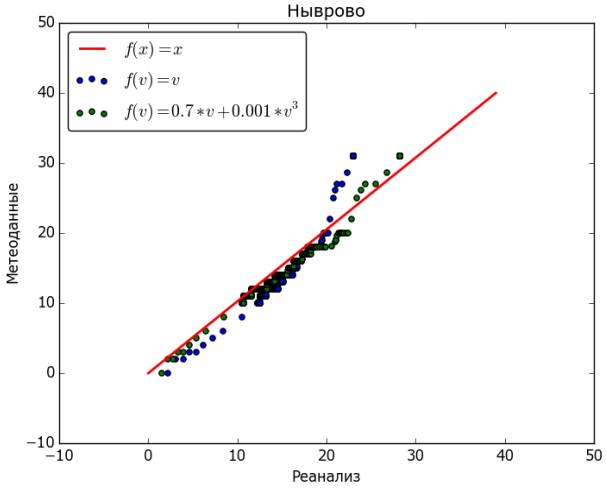


Example of quantile regression

УНИВЕРСИТЕТ ИТМО

Example: quantile polynomial regression $y = a_1 x + a_3 x^3 + \varepsilon$



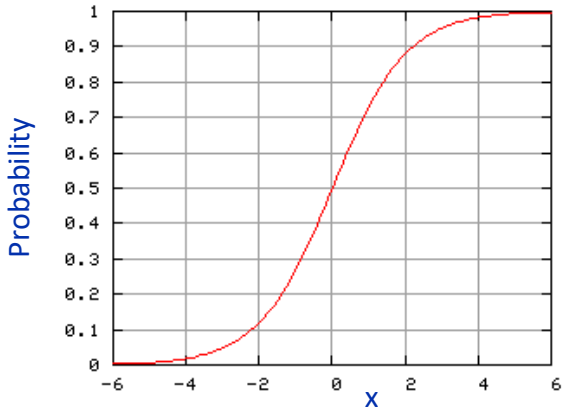


ITMO more than a UNIVERSITY

Logistic regression

Logistic regression is statistic model, which is used for probability prediction of appearance of an event by data adjusment to logistic curve.

Logistic curve



Logistic function (for one variable):

$$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}},$$

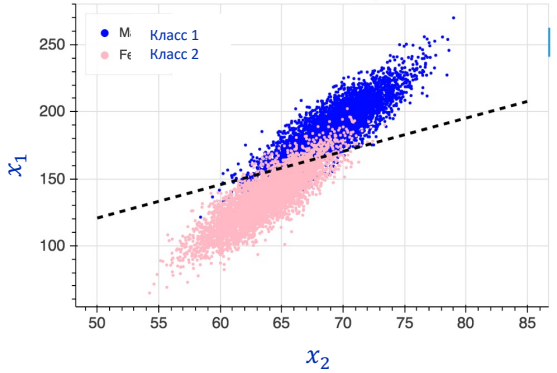
где x – predictor β - Regression coefficients

Logistic function (for several variables):

$$F(x_1, \dots, x_n) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}},$$ где

где x – predictors , β - Regression coefficients

Decision boundary



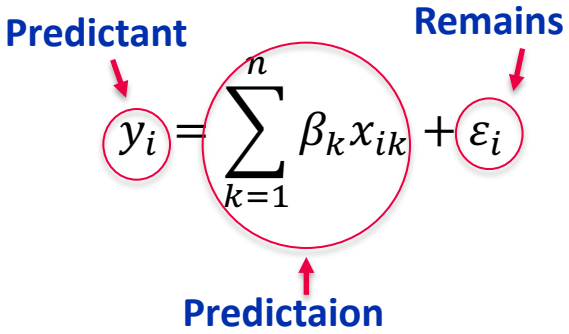
The method of logistic regression keeps all disadvantages of multidimensional linear regression therefore practical realization has to provide standardization of data, elimination of outliers, regularization of scales, feature selection.

How to estimate quality of regression and classification?

Concept of regression model quality (1/2)

УНИВЕРСИТЕТ ИТМО

Intuitively, "quality" of regression model is described by its ability to reproduction of regularities in data



How to estimate it formally?

1) Correlation coefficient between a predictant and prediction:

$$|r| \to 1$$

2) Correlation coefficient between a predictant and the remains:

$$|r| \to 0$$

3) Testing the hypothesis of normality of remains distribution (with mat. expectation equals 0):

$$\varepsilon \sim N(0,\sigma)$$

4) Determination coefficient assessment:

$$R^2 \equiv 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Variance of remains

Variance of predicant

5) With different metrics (s. 46)

Concept of regression model quality (2/2)

УНИВЕРСИТЕТ ИТМО

Relative measures of quality - when there are several models from which it is possible to choose

Akaike's information criterion, AIC

Penalty on excessive complexity of model

$$AIC = 2k + 2 \cdot \ln(L)$$

k -number of predictants

$\ln(L)$ - log likehood function

When the amount of samples is the same: $AIC = 2k + n[ln(RSS)]$,

$RSS = \sum_{n} (e_i)^2$ - remains squares sum

Bayesian information criterion

$$BIC = k \cdot \ln(n) + 2 \cdot \ln(L)$$

n - number of observations

The strengthened penalty

ITMO more than a UNIVERSITY

Metrics for estimation of an error of regression model

УНИВЕРСИТЕТ ИТМО

1) Average integrated error (BIAS):

$$BIAS = \frac{\sum_{i=1}^{n}(y_i - f_i)}{n}$$

2) Mean Absolute Integral Error (MAE):

$$MAE = \frac{\sum_{i=1}^{n}|(y_i - f_i)|}{n}$$

3) Mean square error (MSE):

$$MSE = \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}$$

4) Standart Deviation of Integral Error (STD):

$$STD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}}$$

ITMO more than a UNIVERSITY

Confidence interval for mean model result

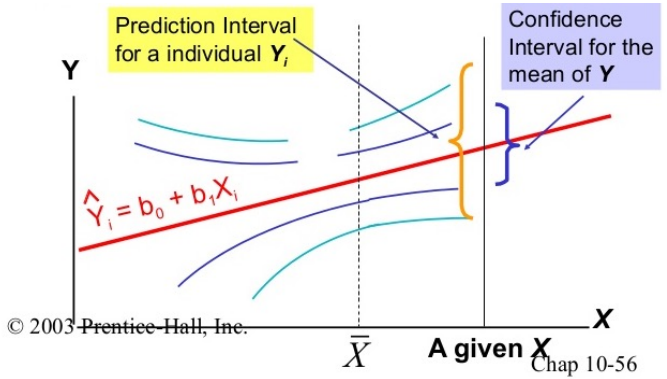
УНИВЕРСИТЕТ ИТМО

General expression (matrix form):

$$(XB)_i - t_{N-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{X_i(X^TX)^{-1}(X_i)^T} < Y < (XB)_i + t_{N-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{X_i(X^TX)^{-1}(X_i)^T}$$

For simple linear regression:

$$\hat{\sigma}\sqrt{\frac{X_i(X_i)^T}{X^TX}} = \hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}\right.}$$



from www.slideshare.net

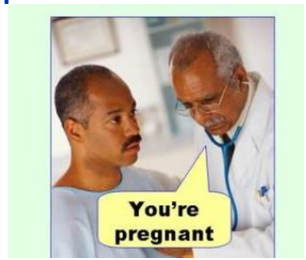
ITMOre than a UNIVERSITY

Type 1 and 2 errors as classification errors

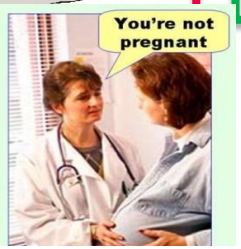
УНИВЕРСИТЕТ ИТМО

	Hypothesis - 0	Hypothesis - 1
In fact - 0 (Negative)	True negative 	False positive 
In fact - 1 (Positive)	False negative 	True positive 

Type 1 error:



Type 2 error:



ITMO more than a UNIVERSITY

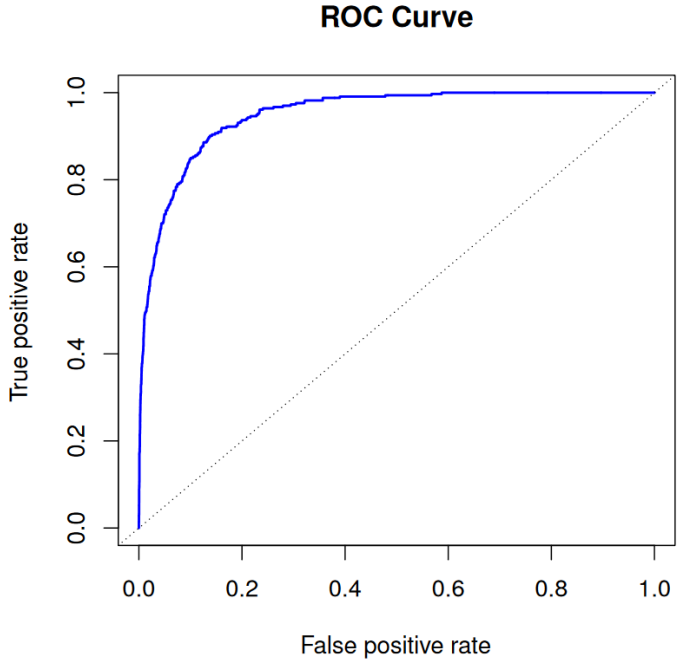
Assessment of the classifier model quality

AUC-ROC (or ROC AUC) — ***Area Under Curve or Receiver Operating Characteristic*** under "curve". This curve is a line from (0,0) to (1,1) in coordinates of True Positive Rate (TPR) and False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

In ideal case, when classifier doesn't make errors, (FPR = 0, TPR = 1)



Gini coefficient and Kolmogorov-Smirnov metric

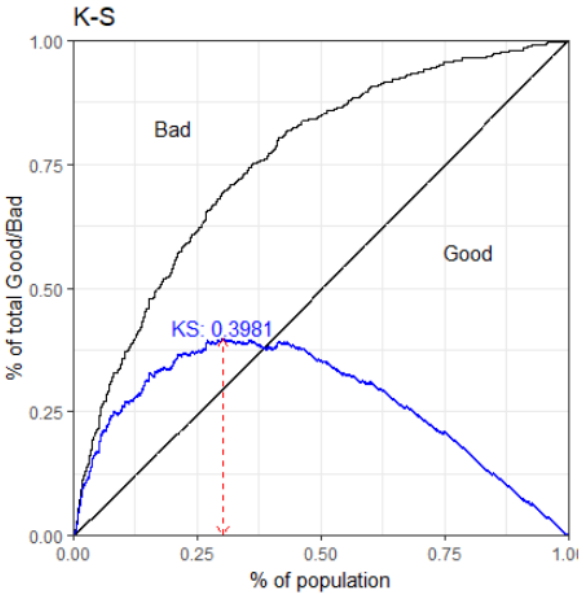
The Gini coefficient shows a statistical indicator of the degree of data stratification by investigated feature.
It can be calculated as

$$GINI = 1 - 2AUC$$

The Kolmogorov-Smirnov metric allows to compare probability distributions for real and predicted classes, and the criterion itself is calculated as the maximum distance between two curves:

$$KS = \max_i |cpB_i - cpG_i|$$

Where cpG – cumulative probability for «1» class, cpB – cumulative probability for «0» class.



Example of KS curve for various cutting-off thresholds for credit scoring model (the ROC curve is shown in black)

How to deal with multicollinearity problem?

Selection of predictors for regression

УНИВЕРСИТЕТ ИТМО

Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others.

Existence of considerable correlation between a target variable and predictors - a basis for creation of regression models.

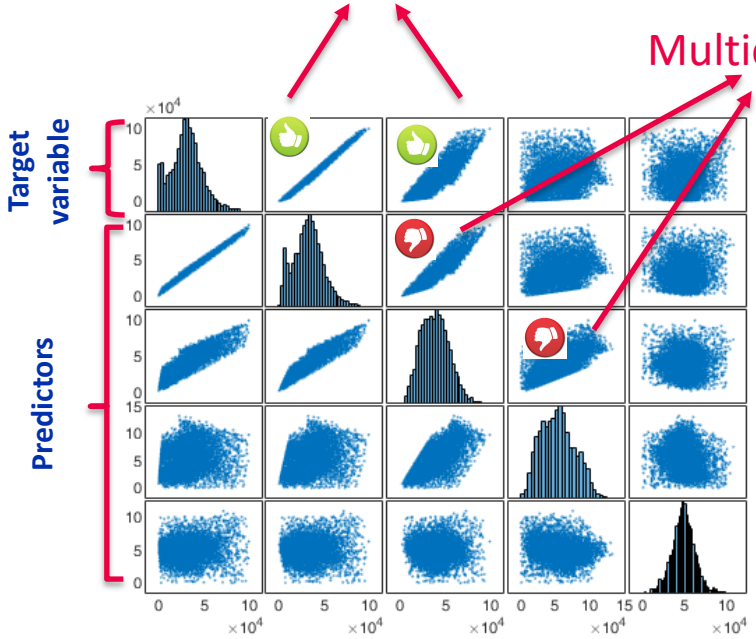
Correlation matrix:

	y	x_1	x_2	…	x_m
y	1	r_{yx_1}	r_{yx_2}	…	r_{yx_m}
x_1	r_{x_1y}	1	$r_{x_1x_2}$	…	$r_{x_1x_m}$
x_2	r_{x_2y}	$r_{x_2x_1}$	1	…	$r_{x_2x_m}$
…	…	…	…	…	…
x_m	r_{x_my}	$r_{x_mx_1}$	$r_{x_mx_2}$	…	1

Target variable

Predictors

For high-quality regression: $|r| \to 0$



Scatterplot matrix on the example

Multicollinearity

УНИВЕРСИТЕТ ИТМО

Multiple coefficient of correlation (from 0 to 1) :

$$R_0 = \sqrt{1 - \frac{\det(Q)}{Q_{yy}}}$$

Q_{yy} - algebraic complement to yy element of D

VIF – Variance Inflation Factor

$$VIF = \frac{1}{1 - R^2{}_j}$$

$R^2{}_j$ - Determination coefficient for j-factor

The solution 1. Selection of predictors on the basis of the analysis:

- Pair coefficients of correlation between predictors.
- Multiple coefficients of correlation.
- Determination coefficient.

The solution 2. Reduce linear dependency between predictors:

- Orthogonalization of predictor matrix.
- Regularization of model solution.

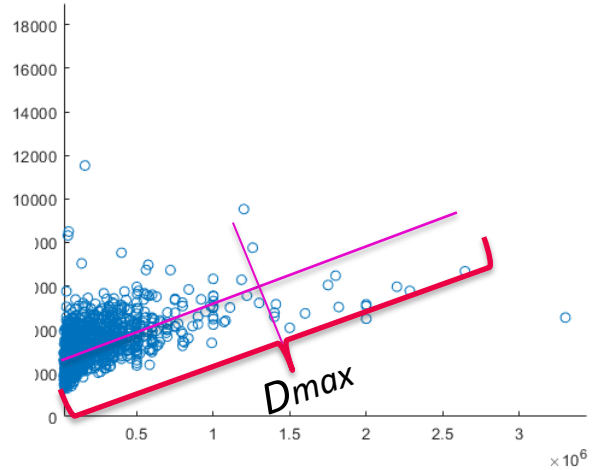
ITMO more than a UNIVERSITY

Principle component analysis

УНИВЕРСИТЕТ ИТМО

The principle component analysis is applied to reduce data dimensionality with preserving possible large amount of information (in terms of variance)

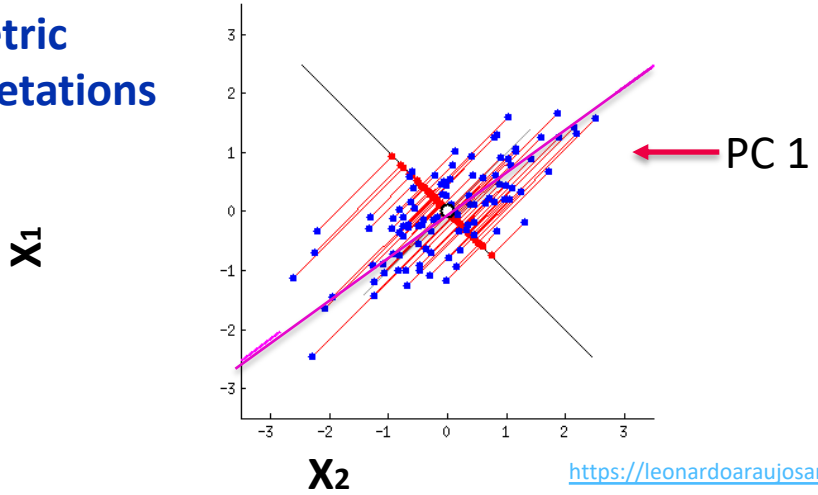
Essence: searching of new system of coordinates (axes) in which data wouldn't not be correlated (orthogonal) and would have the maximum variance.



Example: PC axes (pink) in transactional data

Illustration of the choice of a new axis with the smallest sum of distances squares to it:

Geometric interpretations



https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/

Methods of predictive models regularization

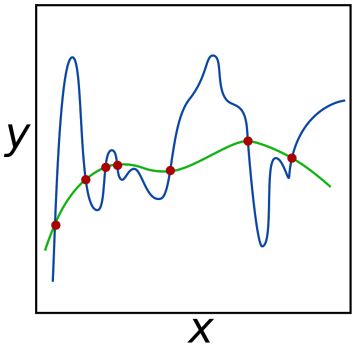
Regularization in statistics, machine learning, the theory of inverse problem - method of addition of some extra information to a condition, generally on purpose to prevent overfitting.

The overfitting appears when resulting polynomials have too large coefficients. The idea of regularization is to minimize quality functional using penalty functions for large weights in the model (a penalty for complexity).

Methods of regularization:

- **LASSO regression:** $L_1 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i |a_i|$
- **Ridge regression:** $L_2 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i a_i^2$

Penalty members of regression



Comparison of regularized (**green**) and simple (**blue**) models with identical values in the known points (**red**)

Coefficient of regularization λ is a hyper-parameter, which should be tuned:

* High λ – too simple models
*Low λ – risk of over-training
*Tune λ – by using cross-validation (decribed later on)

Regularization in classification task

УНИВЕРСИТЕТ ИТМО

When we use classification as logistic regression it is quite simple to transfer regularization idea to classification task.

Regularized logistic regression (example: credit scoring task)

$$P\{y\,|x\} = f(\theta^T x)^y \left(1 - f(\theta^T x)\right)^{1-y}, y \in \{0,1\}, f(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\theta = \text{argmin}\left(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \theta_j x_{ij})^2 + \lambda|\theta|\right)$$

Where P – probability of default, a - logistic function (sigmoid), x – profile parameters, y – default label, θ - regularized regression parameters, n – amount of loaners and m – variables amount, i – profile index, j - variable index

Example of quality increase of credit classifier after regularization application:

Model	KS criteria	GINI	AUC
Logistic regression	0.35	0.47	0.74
Logistic regression with regularization	**0.37**	**0.48**	**0.75**

Thanks!

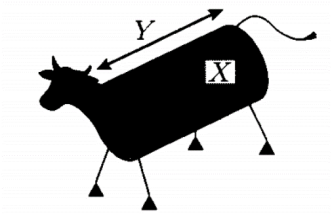
www.ifmo.ru

ITMOre than a UNIVERSITY

Task 1

УНИВЕРСИТЕТ ИТМО

Task: Guess the best and the most robust model



Лагутин. Наглядная математическая статистика

1. Linear regression $Z = a_0 + a_1 X + a_2 Y$
2. Power regression $Z = a_0 X^{a_1} Y^{a_2}$
3. Equation with physical meaning $Z = a_0 X^2 Y$

Z – cow weight, X – circuit of cow body, Y – cow length

Conditions:

1. Training sample – 10 cows with the largest weight
2. Cross-validation
3. Checking results on sample contained weights of 10 smallest cows

ITMO more than a UNIVERSITY

Task 2

УНИВЕРСИТЕТ ИТМО

Task: Guess a kind of mistake of English military-officers?

Example:

During the II World War English military officers investigate dependency between bomb-dropping accuracy (*Z*) and a set of factors.

Set of factors:

H – flight altitude of bomber

V – wind speed

X – number of fighter jets of enemy

Result: greater number of enemy's fighter jets – better the accuracy of English bombers

ITMO more than a UNIVERSITY