Probaility Theory and Statistics

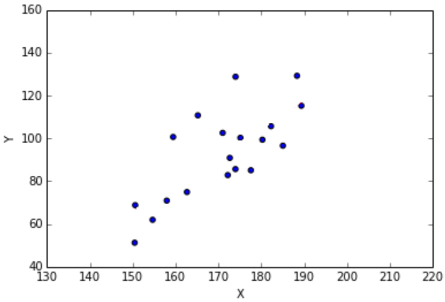
Lecture 7: Correlation analysis and regression

22 May 2025

Lecturer: Batashov Ruslan

Email: rusbatashov@gmail.com

Regression problem setup

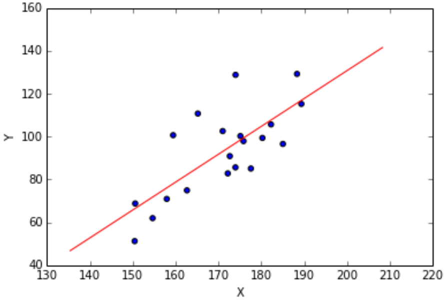


- There are several datapoints given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- There is some dependency between y and x i.e. $y = f(x)$.
- However, observed y_i differ from $f(x_i)$ due to some random error ε_i ("noise"):

$$\varepsilon_i = y_i - f(x_i)$$

Linear regression: look for linear dependency $f(x) = \alpha + \beta x$

i.e. try to understand the **trend**



The data points are located along the line $y = \alpha + \beta x$.

- We don't know true values α and β
- We need to **estimate** them based on data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ available i.e. find estimators $\widehat{\alpha}$, $\widehat{\beta}$
- And build regression line $\widehat{y} = \widehat{\alpha} + \widehat{\beta} x$

How to choose coefficients α, β? Intuition

Suppose we know the joint distribution of (X, Y) and want to find the coefficients α, β which make the model as precise as possible.

The precision criterion: $E(Y-(\alpha+\beta X))^2 \to min$ i.e. to make expected difference as small as possible.

Solution:

$$\beta = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X} = \rho \frac{\sigma_y}{\sigma_x}, \qquad \alpha = \mu_y - \beta \mu_x$$

Financial intuition

Y - return on an asset or portfolio,

X - market return

Then β becomes a parameter in Capital Asset Pricing Model (CAPM).

Linear regression model

Consider the model

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

where

- X is called an independent variable (also regressor or factor),
- Y is called a dependent variable (also response variable),
- ε is called a random error,
- the coefficients α, β are constant (α is called intercept, β is called slope).

Given values $x_1, x_2, \ldots, x_n$ of the independent variable, the values of the response variable $Y_1, Y_2, \ldots, Y_n$ follow the model

$$Y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n$$

where

- $x_1, \ldots, x_n$ are non-random values
- $Y_1, \ldots, Y_n$ are random variables
- $\varepsilon_1, \ldots, \varepsilon_n$ are iid random variables
- $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, \sigma^2)$, where σ^2 is the same for all $i = 1, \ldots, n$
- α, β, σ^2 are fixed unknown model parameters which needs to be estimated

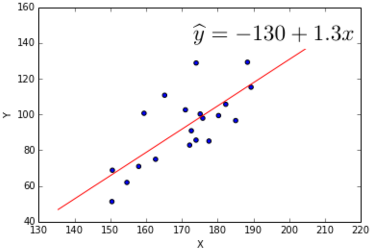
Estimation of α, β from data

Given values $(x_1, y_1), \ldots, (x_n, y_n)$ we can estimate intercept and slope and obtain $\widehat{\alpha}$ and $\widehat{\beta}$.

Then we can predict the value of y for a given value of x by the formula

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x.$$

The line $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$ is called the regression line.



How to obtain estimators $\widehat{\alpha}$ and $\widehat{\beta}$?

Ordinary Least Squares (OLS) method

The formulas for $\widehat{\alpha}$, $\widehat{\beta}$ can be obtained by the following method.

Suppose we have a dataset $(x_1, y_1), \ldots, (x_n, y_n)$. For *given* coefficients α, β define the residuals of the linear model by

$$y_i - (\alpha + \beta x_i)$$

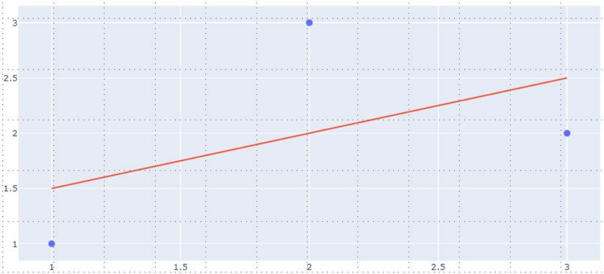
The OLS estimates of the regression coefficient are the coefficients α, β such that

$$(y_1 - \alpha - \beta x_1)^2 + (y_2 - \alpha - \beta x_2)^2 + \ldots + (y_n - \alpha - \beta x_n)^2 \text{ is minimal.}$$

Example

Linear regression by 3 datapoints:

$$(x_1, y_1) = (1, 1), (x_2, y_2) = (2, 3), (x_3, y_3) = (3, 2)$$



$$\Downarrow$$

regression line

The OLS formulas for $\widehat{\alpha}$ and $\widehat{\beta}$

We have the quadratic function in α and β:

$$f(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

The OLS estimators $\widehat{\alpha}, \widehat{\beta}$ are found by minimizing the function $f(\alpha, \beta)$, which is done by finding $\widehat{\alpha}, \widehat{\beta}$ such that

$$\begin{cases} f'_\alpha(\widehat{\alpha}, \widehat{\beta}) = 0, \\ f'_\beta(\widehat{\alpha}, \widehat{\beta}) = 0. \end{cases}$$

We obtain formulas

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \qquad \widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}.$$

Sample correlation

Recall the definition of the correlation coefficient from probability theory:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)},$$

where $\mathrm{Cov}(X,Y)$ is the covariance:

$$\mathrm{Cov}(X,Y) = E\big((X-EX)(Y-EY)\big),$$

and $\sigma(X)$, $\sigma(Y)$ are the standard deviations:

$$\sigma(X) = \sqrt{E(X-EX)^2}, \quad \sigma(Y) = \sqrt{E(Y-EY)^2}.$$

What will be analogues of $\rho(X,Y)$ and $\mathrm{Cov}(X,Y)$ in Statistics?

Sample correlation coefficient

Suppose we have a dataset $(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$

Define the following sums:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = (n-1)s_x^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = (n-1)s_y^2,$$

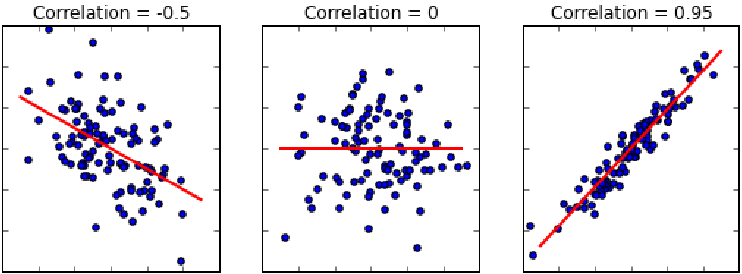
$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

To estimate $\operatorname{cov}(X, Y),\ \rho(X, Y)$, we will use the following analogues:

$\operatorname{cov}(X, Y)$	$\dfrac{1}{n-1} S_{xy}$
$\rho(X, Y)$	$r_{xy} = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

r_{xy} is called sample correlation coefficient or Pearson correlation coefficient

Examples



Basic properties of the sample correlation coefficient

r_{xy} has properties similar to that of $\rho(X, Y)$.

1. r_{xy} is always between -1 and 1.

2. r_{xy} does not change under a linear transformation of samples: if $v_i = a + bx_i,\ \ u_i = c + dy_i$, and $b, d > 0$, then $r_{uv} = r_{xy}$.

3. $r_{xy} = -1$ when there exists $\beta < 0$ such that $y_i = \alpha + \beta x_i$ for all i.

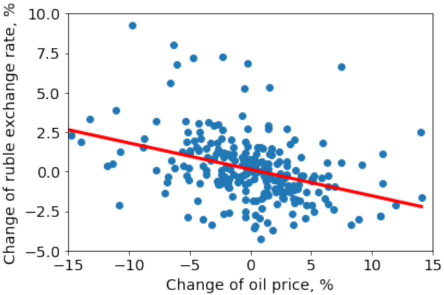
4. $r_{xy} = +1$ when there exists $\beta > 0$ such that $y_i = \alpha + \beta x_i$ for all i.

Based on notations above we can rewrite

$$\boxed{\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} = r_{xy}\frac{s_y}{s_x}}$$

Example

Regression of weekly changes of Ruble exchange rate and oil prices:



$r_{xy} = -0.39$, regression line: $y = 0.14 - 0.17x$

Coefficient of determination R^2

For a linear regression

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i, \quad e_i = y_i - \widehat{y}_i$$

define the Sum of Squares Total, the Sum of Squares Regression, and Sum of Squares Error:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2, \quad SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2,$$
$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

SST shows the total variability in the response variable.

SSR shows the variability explained by the regression equation.

SSE shows the variability which cannot be explained by the regression.

It can be shown that for regression line

$$SST = SSR + SSE.$$

The coefficient of determination R^2 is by definition

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

R^2 shows how well the model explains the data:

"$R^2 \cdot 100\%$ of the variability in y can be explained by the regression equation."

- $R^2 = 0$ means the linear model does not explain the variability
- $R^2 = 1$ means the linear model completely explains the variability (y is a linear function of x)

Relation between the coefficients of correlation and determination

For the linear regression model with one independent variable

$$\boxed{R^2 = r_{xy}^2}$$

As a result, $r_{xy} = \sqrt{R^2}$ if $\beta > 0$ and $r_{xy} = -\sqrt{R^2}$ if $\beta < 0$.

Interpretation of r_{xy}

$\lvert r\rvert$	Strength of the relationship
$\geqslant 0.8$	Very strong
$0.6 - 0.8$	Strong
$0.4 - 0.6$	Moderate
$0.2 - 0.4$	Weak
< 0.2	Very weak

Confidence intervals and hypothesis testing

Sampling distributions of $\widehat{\alpha}$ and $\widehat{\beta}$

The following properties are known:

- $\widehat{\alpha}$ and $\widehat{\beta}$ have normal distributions,
- $\widehat{\alpha}$, $\widehat{\beta}$ are unbiased: $E(\widehat{\alpha}) = \alpha$, $E(\widehat{\beta}) = \beta$,
- the variances of $\widehat{\alpha}$, $\widehat{\beta}$ are

$$\mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}} \qquad \mathrm{Var}(\widehat{\alpha}) = \frac{\sigma^2}{S_{xx}} \cdot \overline{x^2} \qquad \mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = \frac{-\overline{x} \cdot \sigma^2}{S_{xx}}$$

where $\overline{x^2} = \frac{1}{n} \sum_i x_i^2$.

Since our model $Y_i = \alpha + \beta x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$ is based on normal distribution we have

$$\widehat{\beta} \sim N(E(\widehat{\beta}), \mathrm{Var}(\widehat{\beta})),$$
$$\widehat{\alpha} \sim N(E(\widehat{\alpha}), \mathrm{Var}(\widehat{\alpha}))$$

$$\frac{\widehat{\beta} - \beta}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

σ is unknown $\Rightarrow$ how to estimate it?

An estimator for σ^2

An unbiased estimator for σ^2 is

$$s^2 = \frac{SSE}{n-2} = \frac{(1-r_{xy}^2)S_{yy}}{n-2}$$

Replacing σ by s, we define standard errors of $\widehat{\alpha}$ and $\widehat{\beta}$:

$$se(\widehat{\beta}) = \frac{s}{\sqrt{S_{xx}}} = \frac{\sqrt{1-r_{xy}^2}}{\sqrt{n-2}} \cdot \frac{s_y}{s_x}, \qquad se(\widehat{\alpha}) = se(\widehat{\beta}) \cdot \sqrt{\overline{x^2}}$$

Why do we divide by $n-2$, not by $n-1$?

t-statistics for the regression coefficients

Theorem

If ε_i are normal random variables, then the t-statistics

$$t_{slope} = \frac{\widehat{\beta} - \beta}{se(\widehat{\beta})}, \qquad t_{intercept} = \frac{\widehat{\alpha} - \alpha}{se(\widehat{\alpha})}$$

have the $t(n-2)$ distribution.

Confidence intervals and hypothesis testing

Confidence intervals for α, β

Using that $t_{slope} \sim t(n-2)$, $t_{intercept} \sim t(n-2)$ we get

$$\alpha = \widehat{\alpha} \pm t_{a/2}(n-2) \cdot se(\widehat{\alpha}),$$

$$\beta = \widehat{\beta} \pm t_{a/2}(n-2) \cdot se(\widehat{\beta}).$$

One-sided intervals are obtained similarly, replacing $t_{a/2}(n-2)$ with $t_a(n-2)$.

Hypothesis testing

To test the null hypothesis

$$H_0\colon \beta = \beta_0$$

use the t-statistic

$$t_{st} = \frac{\widehat{\beta} - \beta_0}{se(\widehat{\beta})}.$$

If H_0 is true, then it has $t(n-2)$ distribution $\Longrightarrow$ compute rejection regions or p-values as usual for two-sided or one-sided alternatives.

In the same way, tests for α can be performed.

Test of regression significance

To test the hypothesis

$$H_0: \beta = 0 \text{ (regression is not significant)}$$
$$H_1: \beta \neq 0 \text{ (regression is significant)}$$

we use

$$t_{st} = \frac{\widehat{\beta} - 0}{se(\widehat{\beta})} = \frac{r\frac{s_y}{s_x}}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}} \cdot \frac{s_y}{s_x}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Test rules

- For $H_1: \beta \neq 0$: reject H_0 is $|t_{st}| > t_{a/2}(n-2)$
- For $H_1: \beta > 0$: reject H_0 is $t_{st} > t_a(n-2)$
- For $H_1: \beta < 0$: reject H_0 is $t_{st} < -t_a(n-2)$

The null hypothesis of the absence of correlation

We consider a test for the null hypothesis of the absence of correlation

$$H_0 : \rho(X, Y) = 0$$

and the two-sided or one-sided alternatives

$$H_1 : \rho(X, Y) \neq 0 \quad \text{or} \quad H_1 : \rho(X, Y) > 0 \quad \text{or} \quad H_1 : \rho(X, Y) < 0$$

based on a pairs of values $(x_1, y_1), \ldots, (x_n, y_n)$.

Since

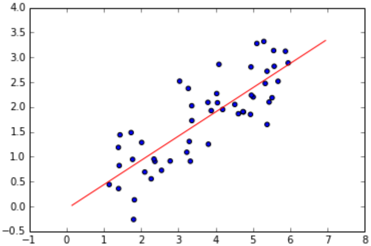
$$\beta = \frac{\mathrm{cov}(Y, X)}{\mathrm{Var}(X)} = \rho(X, Y) \frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{\mathrm{Var}(X)}}$$

we have $\beta = 0 \Leftrightarrow \rho(X, Y) = 0$. Therefore, for testing H_0 we use

$$t_{st} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

Computer output of fitting a regression model

Example



Regression output:

Variable	Coefficient	Std. Error	t-statistic	P-value
C	-0.05	0.194	-0.263	0.794
X	0.48	0.049	10.02	0

The p-values correspond to the two-sided tests that the coefficients are zero.

Reading

Newbold, Carlson, Thorne: § 11.1 – 11.4

Mann: § 13.1, 13.2