CFAspace

Provided by APF

Academy of Professional Finance 专业金融学院

CFA Level II

Correlation and Regression

Part I

CFA Lecturer: Jiahao Gu

Content

Correlation Analysis (Los a, b)	Sample covariance Sample correlation coefficient Scatter plot Limitations
Hypothesis Test (Los c)	Population correlation coefficient t-statistics
Simple Linear Regression (Los d, e)	Dependent / independent variables Assumptions Regression coefficient
Linear Regression Calculation (Los f)	Standard error of estimate Coefficient of determination (R^2) Confidence interval for a regression coefficient

Correlation Analysis

Sample covariance

Covariance is a statistical measure of the degree to which the two random variables move together.

A positive covariance indicates that the variables tend to move together; a negative covariance indicates that the variables tend to move in opposite directions.

$$\text{cov}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

n = sample size

X_i = ith observation on variable X

X = mean of the variable X observations

Y_i = ith observation on variable Y

Y = mean of the variable Y observations

The covariance may range from negative to positive infinity, and it is presented in terms of squared units. That's why we calculate the correlation coefficient, which converts the covariance into a standardized measure that is easier to interpret.

No.	Age(X)	Salary(Y)	X_i-Xbar	Y_i-Ybar	(X_i-Xbar)(Y_i-Ybar)
1	42	356	6.9	112.5	776.3
2	23	70	-12.1	-173.5	2099.4
3	37	418	1.9	174.5	331.6
4	24	12	-11.1	-231.5	2569.7
5	29	22	-6.1	-221.5	1351.2
6	51	455	15.9	211.5	3362.9
7	38	339	2.9	95.5	277.0
8	27	252	-8.1	8.5	-68.9
9	41	278	5.9	34.5	203.6
10	39	233	3.9	-10.5	-41.0
Average	35.1	243.5		COV(Age,Salary)	1206.8

Correlation Analysis

Sample correlation coefficient

The correlation coefficient, r, is a measure of the strength of the linear relationship (correlation) between two variables. The correlation coefficient has no unit of measurement.

$$r_{XY} = \frac{\text{cov}_{xy}}{(s_X)(s_Y)}$$

s_X = standard deviation of variable X
s_Y = standard deviation of variable Y

No.	Age(X)	Salary(Y)	X_i-Xbar	Y_i-Ybar	$(X_i\text{-Xbar})(Y_i\text{-Ybar})$
1	42	356	6.9	112.5	776.3
2	23	70	-12.1	-173.5	2099.4
3	37	418	1.9	174.5	331.6
4	24	12	-11.1	-231.5	2569.7
5	29	22	-6.1	-221.5	1351.2
6	51	455	15.9	211.5	3362.9
7	38	339	2.9	95.5	277.0
8	27	252	-8.1	8.5	-68.9
9	41	278	5.9	34.5	203.6
10	39	233	3.9	-10.5	-41.0
Average	35.1	243.5		COV(Age,Salary)	1206.8

s_X= 9.04 s_Y=160.24, so r_{XY} = 1206.8 / (9.04*160.24) = 0.833

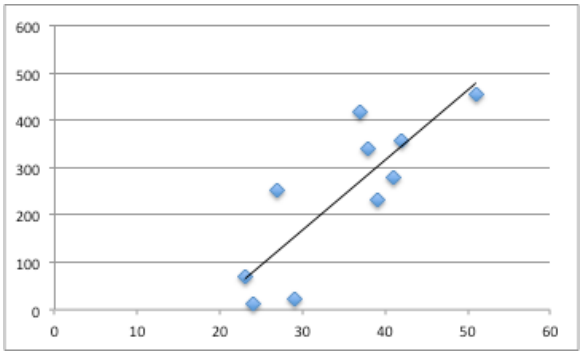
The correlation coefficient is bounded by positive and negative one (i.e., $-1 \le r \le +1$).

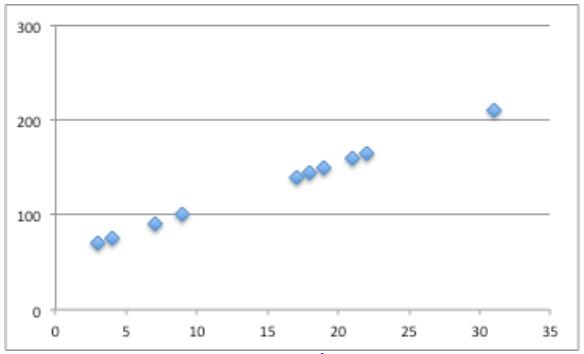
r=1	perfect positive correlation
0<r<1	positive linear relationship
r=0	no linear relationship
-1<r<0	negative linear relationship
r=-1	perfect negative correlation

Correlation Analysis

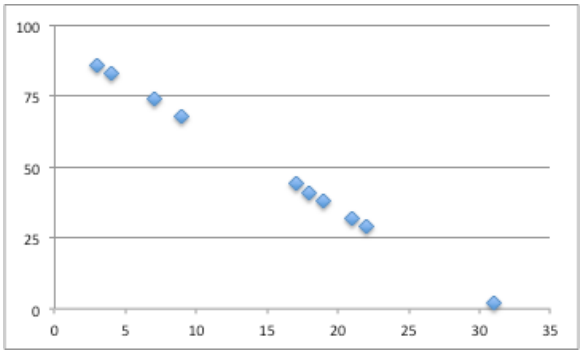
Scatter plot

A scatter plot is a collection of points on a graph where each point represents the values of two variables (i.e., an X/Y pair). Note that for r = 1 and r = -1 the data points lie exactly on a line, but the slope of that line is not necessarily +1 or -1.

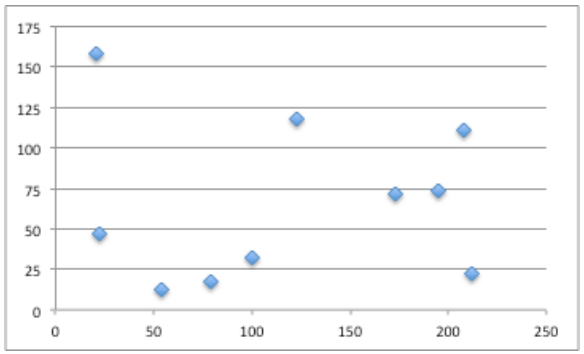




r = +1



r = -1



r = 0

Correlation Analysis

Limitations

-- Outliers

Outliers represent a few extreme values for sample observations. Relative to the rest of the sample data, the value of an outlier may be extraordinarily large or small. Outliers can result in apparent statistical evidence that a significant relationship exists when, in fact, there is none, or that there is no relationship when, in fact, there is a relationship.

-- Spurious Correlation

Spurious correlation refers to the appearance of a causal linear relationship when, in fact, there is no relation. Certain data items may be highly correlated purely by chance. For example, the relationship between weather and stock market. Obviously there is no economic explanation for this relationship, so this would be considered a spurious correlation.

-- Nonlinear Relationships

Correlation measures the linear relationship between two variables. However, two variables could have a nonlinear relationship such as $Y = X^2$. Therefore, another limitation of correlation analysis is that it does not capture strong nonlinear relationships between variables.

Hypothesis Test

Population correlation coefficient

To test whether the correlation between the population of two variables is equal to zero, the appropriate null and alternative hypotheses can be structured as a two-tailed test as follows:

$H_0: \rho = 0$, $H_a: \rho \neq 0$

We use a t-test to determine whether the null hypothesis should be rejected. The test statistic is computed using the sample correlation, r, with n - 2 degrees of freedom (df):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

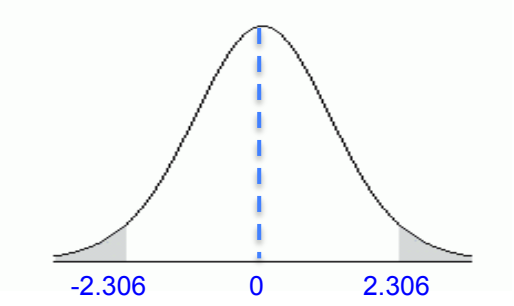
Reject H_0 if $t > t_{critical}$ or $t < -t_{critical}$

Using the previous example, r = 0.833, n = 10,

so $t = 0.833 \times \sqrt{(10 - 2)} / \sqrt{(1 - 0.833^2)} = 4.258$

The two-tailed critical t-values at a 5% level of significance with df = 8 are ±2.306.

Because $t > t_{critical}$, we reject H_0. We conclude that the correlation between age and salary is significantly different than zero at a 5% significance level.

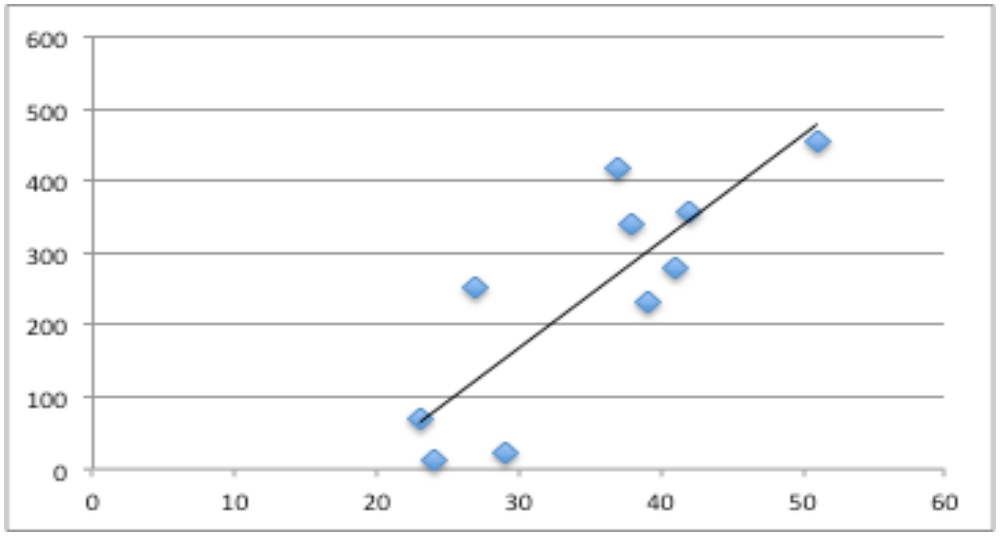


Simple Linear Regression

Dependent / independent variables

The dependent variable is the variable whose variation is explained by the independent variable. The dependent variable is also referred to as the explained variable, the endogenous variable, or the predicted variable.

The independent variable is also referred to as the explanatory variable, the exogenous variable, or the predicting variable.



Dependent variable (salary) is on the vertical axis, while independent variable (age) is on the horizon axis.

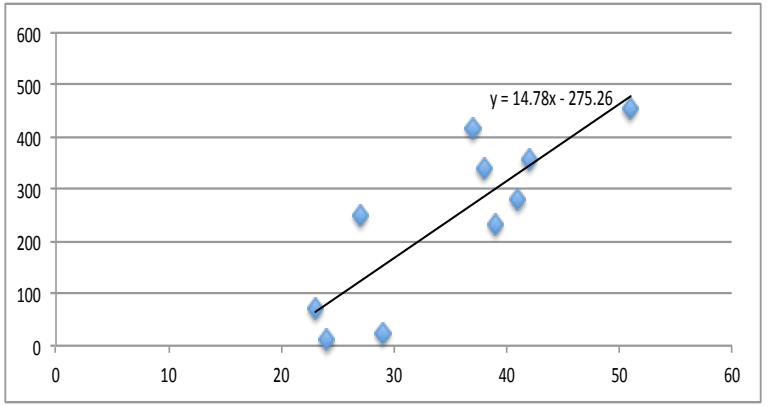
Simple Linear Regression

Interpret Regression Coefficient

Linear Regression Model:

$$Y_i = b_0 + b_1 X_i + e_i$$

b_0: regression intercept term
b_1: regression slope coefficient
e_i : residual for the ith observation



Regression Equation:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

SSE is the sum of the squared vertical distances between the estimated and actual Y-values. The regression line is the line that minimizes the SSE.

$$\hat{b}_1 = \frac{\text{COV}_{XY}}{\sigma_x^2}$$

the slope coefficient equals covariance divided by variance, it also indicates the change in the dependent variable for a 1-unit change in the independent variable

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

the intercept is an estimate of the dependent variable when the independent variable is zero

Using the previous example, $\hat{b}_1$ = 1206.8/(9.04)^2 = 14.78, $\hat{b}_0$ = −275.26
So salary = −275.26 + 14.78 × age.

Simple Linear Regression

Assumptions

Most of the major assumptions pertain to the regression model's residual term(e_i):

-- A linear relationship exists between the dependent and the independent variable

-- The independent variable is uncorrelated with the residuals

-- The expected value of the residual term is zero

-- The variance of the residual term is constant for all observations

-- The residual term is independently distributed; that is, the residual for one observation is not correlated with that of another observation

-- The residual term is normally distributed

Linear Regression Calculation

Standard error of estimate and R^2

-- Standard error of estimate (SEE)

The SEE is the standard deviation of the error terms (e_i) in the regression. SEE is also referred to as the standard error of the residual, or standard error of the regression.

The SEE gauges the "fit" of the regression line. SEE will be low if the relationship is very strong and high if the relationship is weak.

-- Coefficient of Determination (R^2)

The coefficient of determination (R^2) is defined as the percentage of the total variation in the dependent variable explained by the independent variable.

R^2 = r^2 only when there is one independent variable.

No.	Age(X)	Salary(Y)	X_i-Xbar	Y_i-Ybar	Y-hat	e_i
1	42	356	6.9	112.5	345.5	10.5
2	23	70	-12.1	-173.5	64.7	5.3
3	37	418	1.9	174.5	271.6	146.4
4	24	12	-11.1	-231.5	79.4	-67.4
5	29	22	-6.1	-221.5	153.3	-131.3
6	51	455	15.9	211.5	478.5	-23.5
7	38	339	2.9	95.5	286.4	52.6
8	27	252	-8.1	8.5	123.8	128.2
9	41	278	5.9	34.5	330.7	-52.7
10	39	233	3.9	-10.5	301.1	-68.1
Average	35.1	243.5	R^2	0.695	SEE	93.92

Linear Regression Calculation

Regression Coefficient Confidence Interval

The appropriate null and alternative hypotheses can be structured as a two-tailed test as follows:

H_0: $b_1 = 0$, H_a: $b_1 \neq 0$

If the confidence interval at the desired level of significance does not include zero, the null is rejected, and the coefficient is said to be statistically different from zero.

The confidence interval of b_1 is $\hat{b}_1 - t_c \times s_{\hat{b}1} < b_1 < \hat{b}_1 + t_c \times s_{\hat{b}1}$

s_{b1} is a function of the SEE. As SEE rises, s_{b1} also increases, and the confidence interval widens, which means less confidence of the regression coefficient.

The estimated slope coefficient, b_1 from the previous example is 14.79 with a standard error of 3.464. Calculate the 95% confidence interval of b_1

The confidence interval of b_1 is
14.79 - 2.306 x 3.464 ~ 14.79 + 2.306 x 3.464, or 6.80 ~ 22.78
Because this confidence interval does not include zero, we can conclude that the slope coefficient is significantly different from zero.

CFAspace

Thank You!