Hellenic Republic

INTERNATIONAL HELLENIC UNIVERSITY

University Center for International Programmes of Studies

Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας

Πρόγραμμα Μεταπτυχιακών Σπουδών:

Διοίκηση Επιχειρήσεων και Οργανισμών για Στελέχη

Ποσοτικές Μέθοδοι για Στελέχη Επιχειρήσεων
Quantitative Methods for Managers
Γραμμική Παλινδρόμηση_2
Linear Regression Analysis_2

by
Dr. Efstathios Dimitriadis
Mathematic
Ph.D in Applied Statistics
M.Sc in Statistics and Demography
M.Sc in Quality Assurance

Έλεγχος Σημαντικότητας του μοντέλου

Για να ελεγχθεί πόσο καλά η συνάρτηση της παλινδρόμησης προσαρμόζεται στα δεδομένα είναι απαραίτητο να κατασκευαστεί ο πίνακας ANOVA.

Model	Sum of Squares	d.f	Mean Square	F	Sig.
Regression	RSS	1	RMS=RSS/1	**RMS/EMS**	Sig.F
Residuals	ESS	n-2	EMS=ESS/n-2		
Total	**TSS**	**n-1**			

Αυτός ο πίνακας δείχνει αν το μοντέλο της παλινδρόμησης προβλέπει την εξαρτημένη μεταβλητή σημαντικά καλά.

Αν Sig.F < a τότε το παλινδρομικό μοντέλο προβλέπει στατιστικά σημαντικά την εξαρτημένη μεταβλητή (καλή προσαρμογή των δεδομένων).

Έλεγχος σημαντικότητας των συντελεστών

Ο συντελεστής παλινδρόμησης $\hat{b}_1$ πρέπει να ελεγχθεί για την σημαντικότητα.

Υποθέσεις $H_0: \hat{b}_1 = 0$ $H_1: \hat{b}_1 \neq 0$

Στατιστικός έλεγχος: $t = \dfrac{\hat{b}_1}{s_{\hat{b}_1}}$

όπου: $s_{\hat{b}_1} = \dfrac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$

Απορρίπτουμε την H_0 αν Sig.< a ή $|t| > t_{a/2,n-2}$

Η t- κατανομή χρησιμοποιείται, με n-2 βαθμούς ελευθερίας

Η απόρριψη της μηδενικής υπόθεσης σημαίνει ότι υπάρχει γραμμική και στατιστικά σημαντική σχέση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής

Διάστημα Εμπιστοσύνης για το $\widehat{b_1}$

Η μορφή του διαστήματος εμπιστοσύνης για το $\widehat{b_1}$ είναι:

$$\widehat{b_1} \pm t_{a/2} \cdot s_{b_1}$$

Ο σημειακός εκτιμητής είναι $\widehat{b_1}$ και το περιθώριο του σφάλματος $t_{a/2} \cdot s_{b_1}$ με **α** επίπεδο σημαντικότητας και **n-2 d.f**.

Αν η υποθετική τιμή 0 του $\widehat{b_1}$ δεν περιλαμβάνεται στο διάστημα εμπιστοσύνης, μπορούμε να απορρίψουμε την μηδενική υπόθεση και να συμπεράνουμε ότι υπάρχει μια στατιστικά σημαντική σχέση μεταξύ των μεταβλητών.

Example of simple linear regression

Ice cream sales and average monthly temperature are given in the table below:

Months	Average Temperature °C	Sales (in thousands €)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

Develop an estimated regression equation for these data

1. Variables' Definition: y= Sales, χ= Temperature

2. Test for Linearity

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
sales * Temperature	Between Groups	(Combined)	7546,000	9	838,444	13,101	0,073
		Linearity	7420,176	1	7420,176	115,940	0,009
		Deviation from Linearity	125,824	8	15,728	,246	0,940
	Within Groups		128,000	2	64,000		
	Total		7674,000	11			

Based on the ANOVA table, value Sig. Deviation from Linearity of 0,940>0,05 it can be concluded that there is a linear relationship between the variables "Temperature" and "Sales".

3. Test for Normality

Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ice cream sales	,181	12	,200*	,948	12	0,604

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Temperature	,163	12	,200*	,914	12	0,242

Based on Shapiro Wilk's test, the significance value for the two variables are 0,604 (sales) and 0,242 (temperature). Based on this test, the sales and temperature significance values are >0,05 and we can conclude that the two variables are normally distributed.

Note!!!

Kolmogorov- Smirnov test for large samples is used.

4. Development of the estimated regression equation

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7420,176	1	7420,176	292,335	,000[b]
	Residual	253,824	10	25,382		
	Total	7674,000	11			

This ANOVA table indicates that the regression model predicts the dependent variable significantly well because the Sig.F =0,000<0,05.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,983[a]	,967	,964	5,03809	2,179

a. Predictors: (Constant), Temperature

b. Dependent Variable: ice cream sales

In the table Model Summary we have:

1. Correlation coefficient R=0,983
2. Coefficient of Determination R^2=0,967 =$(0,983)^2$
3. Adjusted R^2=0,964 and
4. Durbin –Watson index of Autocorrelation =2,179

R^2 with a value of 0,967 indicates that 96,7% of the variation of the "sales" is explained by the "temperature". **D.W** =2,197 Indicates that there is no serious symptom of autocorrelation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	45,520	3,503		12,996	0,000
	Temperature	5,448	,319	,983	17,098	0,000

a. Dependent Variable: ice cream sales

The estimated regression model is:

Sales = 45,520+ 5,448*Temperature

The regression coefficient $\widehat{b}_1$ = 5,448 is statistically significant as the value 17,098 of t, is significant (sig.t=0,000<0,05).

That means that there is a linear and statistically significant relation between sales and temperature.

5. Έλεγχος για Heteroskedasticity

Heteroskedasticity: Είναι χρήσιμο να εξεταστεί εάν υπάρχει διαφορά στην διακύμανση του καταλοίπου της περιόδου παρατήρησης με άλλη περίοδο παρατήρησης. Σε ένα καλό μοντέλο παλινδρόμησης δεν πρέπει να υπάρχει πρόβλημα ετεροσκεδαστικότητας. Μια στατιστική μέθοδος που μπορεί να χρησιμοποιηθεί για να προσδιοριστεί εάν ένα μοντέλο είναι απαλλαγμένο από το πρόβλημα της ετεροσκεδιστικότητας είναι ο έλεγχος Glejser.

Για να εφαρμόσετε αυτό τον έλεγχο, θα πρέπει πρώτα να αποθηκευτούν ως νέα μεταβλητή (RES_1) τα μη τυποποιημένα κατάλοιπα. Μετά από αυτό, θα πρέπει να υπολογιστούν και να αποθηκευτούν ως νέα μεταβλητή (ABS_RES_1) οι απόλυτες τιμές των καταλοίπων. Τέλος, εκτελέστε ένα μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή τη μεταβλητή ABS_RES_1.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8,291	1,333		6,219	0,000
	Temperature	-,459	,121	-,767	-3,784	0,004

a. Dependent Variable: ABS_RES_1

If the value Sig.> 0,05 then there is no problem of Heteroskedasticity

If the value Sig.< 0,05 then there is problem of Heteroskedasticity

Based on output "coefficients" the obtained value of Sig. Temperature of 0,004 <0,05 means that there is problem of Heteroskedasticity

Πώς μπορείτε να βελτιώσετε την ακρίβεια ενός μοντέλου παλινδρόμησης;

Υπάρχουν λίγες ενέργειες που μπορεί να γίνουν όταν τα δεδομένα παραβιάζουν τις υποθέσεις παλινδρόμησης. Μια προφανής λύση είναι η χρήση αλγορίθμων που εντοπίζουν τη μη γραμμικότητα αρκετά καλά. Αν είστε όμως αποφασισμένοι να χρησιμοποιήσετε οπωσδήποτε παλινδρόμηση, ακολουθείστε μερικές συμβουλές που μπορείτε να εφαρμόσετε.

- Εάν τα δεδομένα σας πάσχουν από ***μη γραμμικότητα***, τροποποιήστε τις ανεξάρτητες μεταβλητές χρησιμοποιώντας sqrt, log, square κ.λπ.
- Εάν τα δεδομένα σας πάσχουν από ***ετεροσκεδικότητα***, τροποποιήστε την εξαρτημένη μεταβλητή χρησιμοποιώντας sqrt, log, square κ.λπ. Επίσης μπορείτε να χρησιμοποιήσετε την μέθοδο ελαχίστων τετραγώνων για να αντιμετωπίσετε αυτό το πρόβλημα.
- Εάν τα δεδομένα σας πάσχουν από ***πολυγραμμικότητα***, χρησιμοποιήστε μια μήτρα συσχέτισης για να ελέγξετε τις συσχετισμένες μεταβλητές. Ας υποθέσουμε ότι οι μεταβλητές Α και Β είναι πολύ συσχετισμένες. Τώρα, αντί να αφαιρέσετε μία από αυτές, χρησιμοποιήστε αυτήν την προσέγγιση: Βρείτε τη μέση συσχέτιση των Α και Β με τις υπόλοιπες μεταβλητές. Όποια μεταβλητή έχει τον υψηλότερο μέσο όρο σε σύγκριση με άλλες μεταβλητές, καταργήστε την.
- Μπορείτε να κάνετε επιλογή μεταβλητή με βάση τις τιμές p. Εάν μια μεταβλητή εμφανίζει τιμή p> 0,05, μπορούμε να καταργήσουμε αυτήν τη μεταβλητή από το μοντέλο, καθώς όταν p> 0,05, αποτύγχάνουμε πάντα να απορρίψουμε την μηδενική υπόθεση.

2. Πολλαπλή Παλινδρόμηση

(δύο ή περισσότερες ανεξάρτητες μεταβλητές)

Πολλαπλή Παλινδρόμηση

- Πολλαπλή παλινδρόμηση είναι η μελέτη του πως μια εξαρτημένη μεταβλητή σχετίζεται με δύο ή περισσότερες ανεξάρτητες μεταβλητές.

Μοντέλο Πολλαπλής Παλινδρόμησης

- Το μοντέλο πολλαπλής παλινδρόμησης είναι:

$$y = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 + \ldots\ldots\ldots + \beta_\kappa \chi_\kappa + \varepsilon$$

όπου:

β_0 και $\beta_1, \beta_2, \ldots\ldots, \beta_\kappa$ ονομάζονται ***παράμετροι*** του μοντέλου,

ε είναι η τυχαία μεταβλητή που ονομάζεται ***όρος σφάλματος***.

y είναι γραμμική συνάρτηση των $\chi_1, \chi_2, \ldots \chi_\kappa$ ***συν*** ε

Γίνεται η υπόθεση ότι η ε ***ακολουθεί*** $N(0, \sigma^2)$

Multiple Regression Equation

- Under the assumption that the mean or expected value of ε is ***zero*** the mean or expected value of ***y***, denoted E(y) is:

$$E(y) = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \ldots\ldots\ldots + \beta_\kappa\chi_\kappa$$

The equation above that describes how the mean value of ***y*** is related to $\chi_1, \chi_2, \ldots\chi_\kappa$ ***is called multiple regression equation.***

Estimated Multiple Regression Equation

- The values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_\kappa$ of the regression equation are not known and must be estimated from sample data. A simple random sample is used to compute samples statistics $\hat{b}_0, \hat{b}_1 \ldots\ldots, \hat{b}_k$ that are used as the point estimators of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_\kappa$.
- The estimated multiple regression equation is:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots\ldots + \hat{b}_k x_k$$

Where $\hat{b}_0, \hat{b}_1 \ldots\ldots, \hat{b}_k$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_\kappa$

$\hat{y}_i$ = *estimated value of y*

Least Squares Method

- The estimated multiple regression equation is:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots\ldots + \hat{b}_k x_k$$

Error or Residual: $d_i = y_i - \hat{y}_i$

The development of estimated regression equation is based in the ***last square method***. The presentation of the formula for the regression coefficients $\hat{b}_0, \hat{b}_1 \ldots\ldots, \hat{b}_k$ involves the use of matrix algebra and is beyond the scope of this course.

Least Squares Criterion: $\sum d_i^2 = \min$

Test of significance of the regression model

In order to test how well the regression equation fits the data (i.e., predicts the dependent variable) the table ANOVA is needed to be constructed.

Model	Sum of Squares	d.f	Mean Square	F	Sig.
Regression	RSS	k	RMS=RSS/k	**RMS/EMS**	Sig.F
Residuals	ESS	n-k-1	EMS=ESS/n-k-1		
Total	**TSS**	**n-1**			

k = the number of independent variables

This table indicates if the regression model predicts the dependent variable significantly well. If Sig.F < a indicates that, overall the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

Test of significance of the Regression Coefficients

The regression coefficient s $\hat{b}_i$ should be tested for significance.

Hypotheses $H_0: \hat{b}_i = 0$ $H_1: \hat{b}_i \neq 0$ Test Statistics: $t = \dfrac{\hat{b}_i}{s_{\hat{b}_i}}$

Where: $s_{\hat{b}_i} = \dfrac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$

Reject H_0 if Sig.< a or $|t| > t_{a/2,n-k-1}$

t- distribution is used, with n-k-1 degrees of freedom

The rejection of null hypothesis means that there is a linear and statistically significant relation between dependent and independent variable.

Confidence Interval for β_i

The form of a confidence interval for β_i is as follows:

$$b_i \pm t_{a/2} \cdot s_{b_i}$$

The point estimator is $\widehat{b_i}$ and the margin of error is $t_{a/2} \cdot s_{b_i}$ with ***a*** level of significance and n-k-1 d.f.

If 0, the hypothesized value of β_i is not included in the confidence interval, we can reject the null hypothesis and conclude that a significant relationship exists between the variables.

Multiple Coefficient of Determination

The multiple coefficient of determination can be interpreted as the percentage of the variance of y that can be explained by the estimated regression equation.

R- square: $R^2 = \frac{RSS}{TSS}$

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

Determining When to Add or Delete Variables

An ***F test*** can be used to determine whether it is advantageous to add one or more independent variables to a multiple regression model.
This test is based on a determination of the amount of reduction in the error sum of squares resulting from adding one or more independent variables to the model.
The steps that we should follow to determine when to add or delete variables are:

- Run a regression model with one or more independent variables.
- Calculate the error sum of squares denoted by $ESS(\chi_1, \chi_2,\ldots..,\chi_q)$
- Run a new regression model with the added new variables
- Calculate the error sum of squares denoted by $ESS(\chi_1, \chi_2,\ldots..,\chi_q, \chi_{q+1},\ldots..,\chi_p)$
- Calculate the reduction in ESS resulting from adding the new variables.
- Perform an F test to determine whether this reduction is significant.

$$F = \frac{\dfrac{ESS(x_1, x_2, \ldots, x_q) - ESS(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{ESS(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}}$$

Where, q= the number of variables in the initial regression model

p= the number of variables after adding the new variables

n= the number of observations

If $F > F_a$ we reject the null hypothesis and conclude that the set of additional independent variables is statistically significant.

F statistic is used with p-q numerator degrees of freedom and n-p-1 denominator degrees of freedom.

Example

The Butler Trucking Company is an independent trucking company in south California. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to estimate the total daily travel time for the drivers. Initially the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries. In attempting to identify another independent variable, the managers felt that the number of deliveries could also contribute to the total travel time. A simple random sample of 10 driving assignments provided the data shown in the table.

Driving Assignment	χ_1=Miles Traveled	χ_2 = Deliveries	y=Travel time (hours)
1	100	4	9,3
2	50	3	4,8
3	100	4	8,9
4	100	2	6,5
5	50	2	4,2
6	80	2	6,2
7	75	3	7,4
8	65	4	6,0
9	90	3	7,6
10	90	2	6,1

The ***estimated regression equation*** in the case in which the only independent variable is the “miles traveled” is:

$\hat{y} = 1{,}27 + 0{,}0678x_1$ and the ESS(χ_1)=8,029

Adding the variable χ_2 “number of deliveries”

the ***estimated regression equation is:***

$\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}932x_2$ and the ESS(χ_1,χ_2)=2,299

ESS(χ_1)-ESS(χ_1,χ_2)=8,029-2,299=5,730

An ***F*** test will performed to determine whether this reduction is significant.

$$F = \frac{\frac{ESS(x_1) - ESS(x_1, x_2)}{2-1}}{\frac{ESS(x_1, x_2)}{10-2-1}} = \frac{\frac{5,730}{1}}{\frac{2,299}{7}} = 17,45$$

At 5% level of significance the value of ***F statistic*** with 1 d.f for the numerator and 7 for denominator is 5,59.

Thus, $F=17,45>5,59=F_a$ and we can conclude that the reduction in the error sum of squares is significant. That means that it is advantageous to add the χ_2 variable.

Correlation Analysis

Correlation Analysis

Correlation Analysis is statistical method that is used to discover if there is a relationship between two variables, and how strong that relationship may be.

Correlations are useful because if you can find out what relationship variables have, you can make ***predictions about future behavior***. Businesses use these statistics for budgets and business plans.

Coefficient of Linear Correlation:

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \quad \text{or} \quad r = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{n\sigma_x \sigma_y}$$

- Values of the correlation coefficient are always between –1 and +1.
- A value of +1 indicates that the two variables are perfectly related in a positive linear sense.
- A value of -1 indicates that the two variables are perfectly related in a negative linear sense.
- Values of the correlation coefficient close to zero indicate that the two variables are not linearly related.
- Values of the correlation coefficient >+0,7 or <-0,7 indicate a strong linear relation between the two variables.

Χαρακτηρισμός Συσχέτισης

r > +0.70	Πολύ δυνατή θετική σχέση
+0.40 < r < +0.69	Δυνατή θετική σχέση
+0.30 < r < +0.39	Μέτρια δυνατή θετική σχέση
+0.20 < r < +0.29	Αδύνατη θετική σχέση
+0.01 < r < +0.19	Καμία ή αμελητέα σχέση
0	Καμία σχέση
-0,19< r< -0.01	Καμία ή αμελητέα σχέση
-0,29< r< -0.20	Αδύνατη αρνητική σχέση
-0,39< r< -0.30	Μέτρια δυνατή αρνητική σχέση
-0.69< r< -0.40	Δυνατή αρνητική σχέση
r <-0.70	Πολύ δυνατή αρνητική σχέση

Test of significance of the Correlation Coefficient

Performing the Hypothesis Test

Null Hypothesis: H_0: $\rho_{xy} = 0$

Alternate Hypothesis: H_1: $\rho_{xy} \neq 0$

Test Statistics: $t = \dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ with n-2 d.f

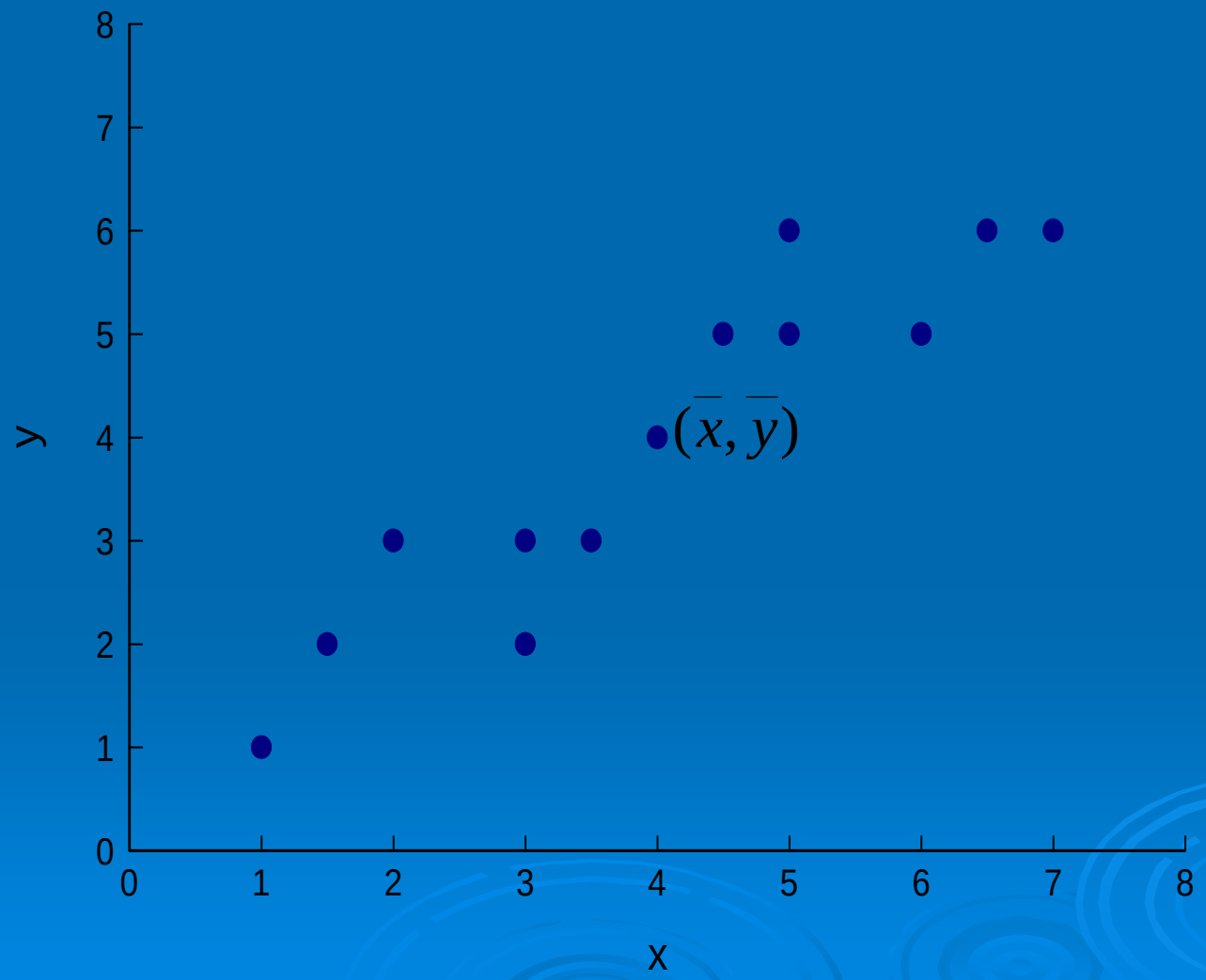
Reject H_0 if Sig.< a or $|t| > t_{a/2,n-2}$

Covariance

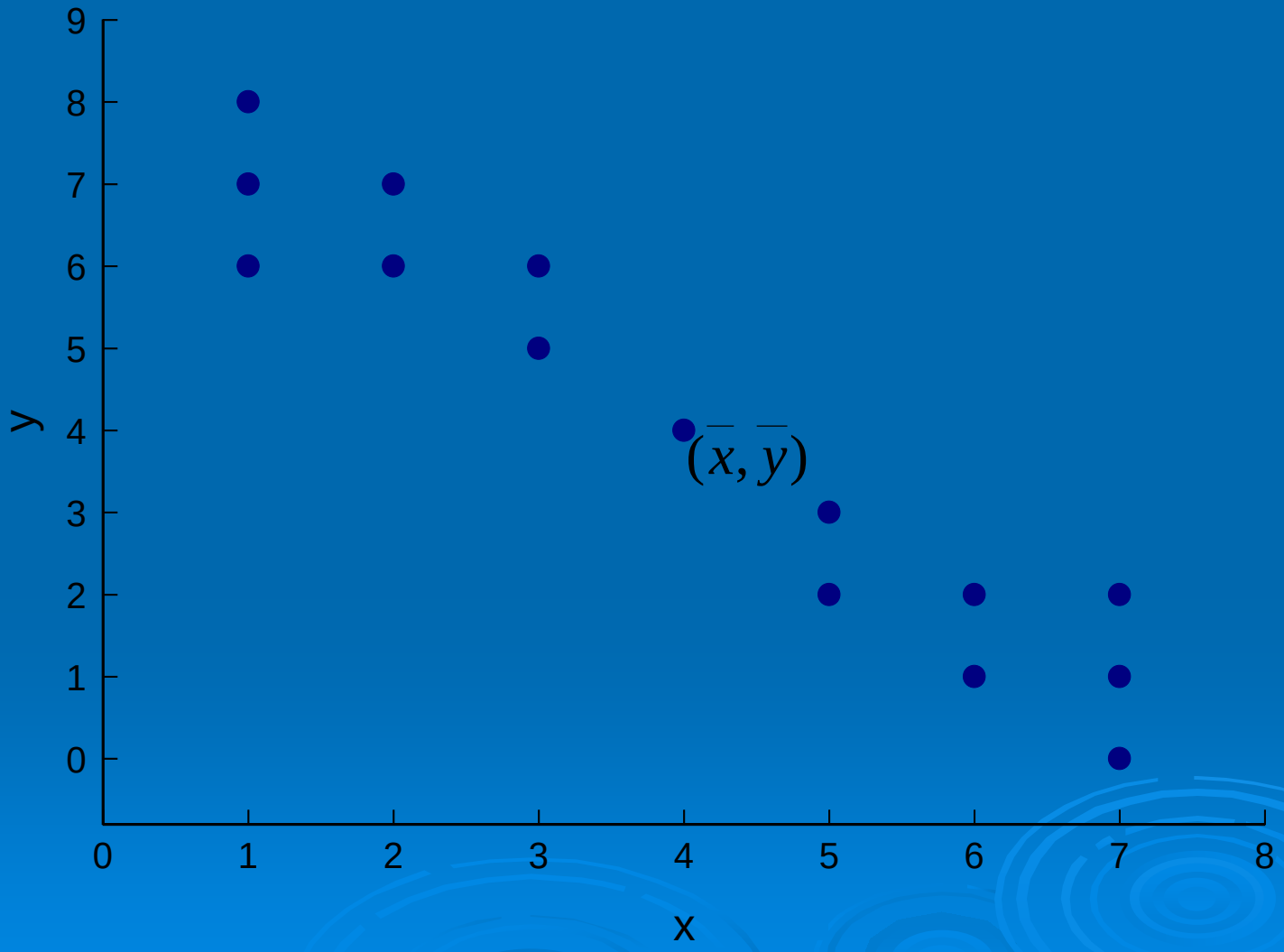
$$\mathbf{COV(X,Y)} = \frac{1}{n}\sum X_i Y_i - \overline{X}\,\overline{Y}$$

1. **If COV(X,Y) >0 positive covariance**
2. **If COV(X,Y)<0 negative covariance**
3. **If COV(X,Y)=0 no covariance**

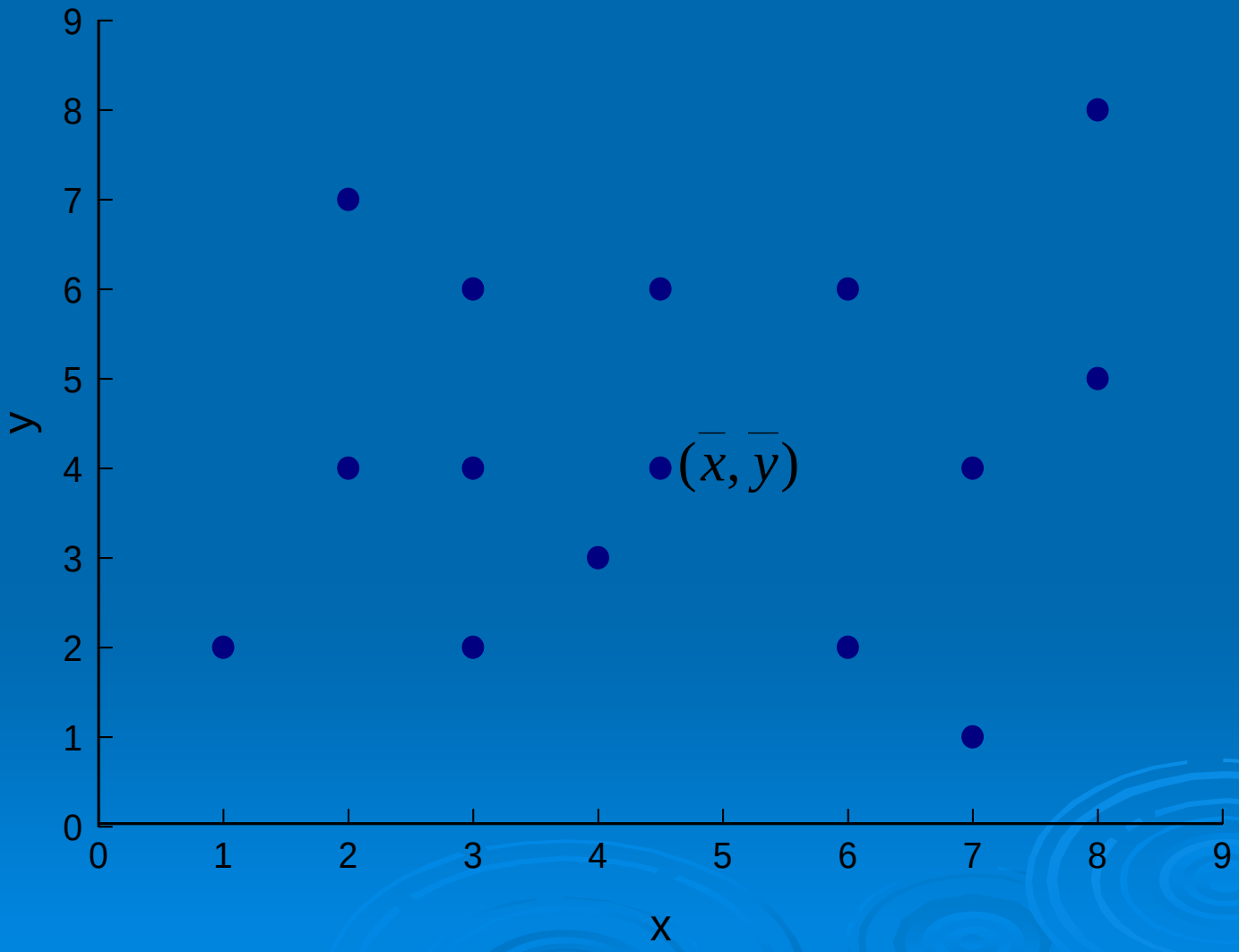
Positive Covariance



Negative covariance



Zero covariance



Partial Correlation

A ***partial correlation*** determines the linear relationship between two variables when accounting for one or more other variables. Typically, researchers and practitioners apply partial correlation analyses when (a) a variable is known to bias a relationship (b) or a certain variable is already known to have an impact, and you want to analyze the relationship of two variables beyond this other known variable.

Notice that the ***partial correlation*** is somewhat smaller than the simple correlation. This suggests that part of the simple correlation is due to each of the variables being related to control variable.

Formula for Partial Correlation

Formula for partial correlation coefficient for χ and y, controlling for z.

$$r_{xy.z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

We must first calculate the zero-order (bivariate correlation) coefficients between all possible pairs of variables (y and χ, y and z, χ and z) before solving this formula.

Example of Partial Correlation

Family	Husband's Housework y	Number of children x	Husband's Years of Education z
A	1	1	12
B	2	1	14
C	3	1	16
D	5	1	16
E	3	2	18
F	1	2	16
G	5	3	12
H	0	3	12
I	6	4	10
J	3	4	12
K	7	5	10
L	4	5	16

Calculate the partial correlation between husbands' housework (Y) and number of children (X), controlling for husbands' years of education (Z)

The table below contains all the bivariate correlations among the variables.

Zero- order Correlations

	y	x	z
y	1	0,499	-0,296
x	0,499	1	-0,466
z	-0,296	-0,466	1

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}} = \frac{0{,}499 - (-0{,}296)(-0{,}466)}{\sqrt{1 - (-0{,}296)^2}\sqrt{1 - (-0{,}466)^2}}$$

$$r_{yx.z} = 0{,}427$$

The partial correlation is somewhat smaller than the simple correlation between y and χ.

!!!! An other way to compute Partial Correlation

The Partial Correlation of variables A and B after removing the effect of variable C is estimated as follows:

- Regress variable A on C
- Regress variable B on C
- For each case, compute the residuals for each of the regression equations
- Compute the usual Pearson correlation between the two sets of residuals.

The residuals represent variables A and B with the effect of variable C removed.