ALY-6015 – Intermediate Analytics

Module 1 – Recap Correlation and Regression

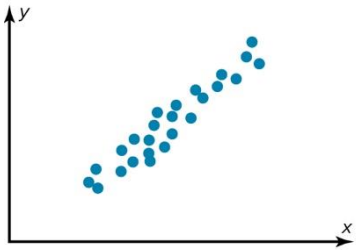
Jean-Sebastien Provost
j.provost@northeastern.edu

January 7th, 2025

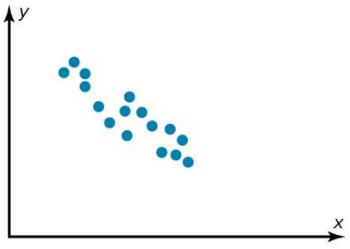
Northeastern University – Vancouver Campus

Linear Relationships Require…a Line

- A **positive linear relationship:** Plot roughly follow the path of an increasing line.
- A **negative linear relationship:** Plot roughly follow the path of a decreasing line.

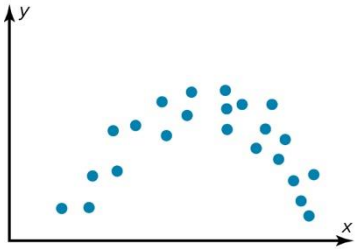


(a) Positive linear relationship

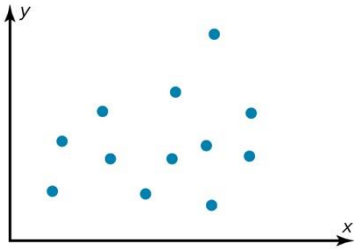


(b) Negative linear relationship

- Some scatterplots show no linear relationship. Sometimes they show a **curvilinear relationship**, where the points follow the path of a non-linear curve, but sometimes they show no relationship.



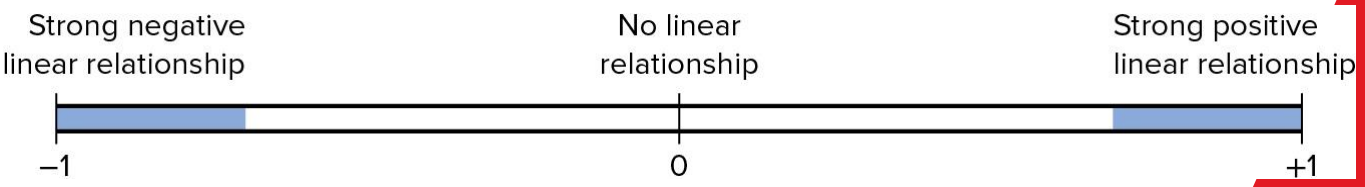
(c) Curvilinear relationship



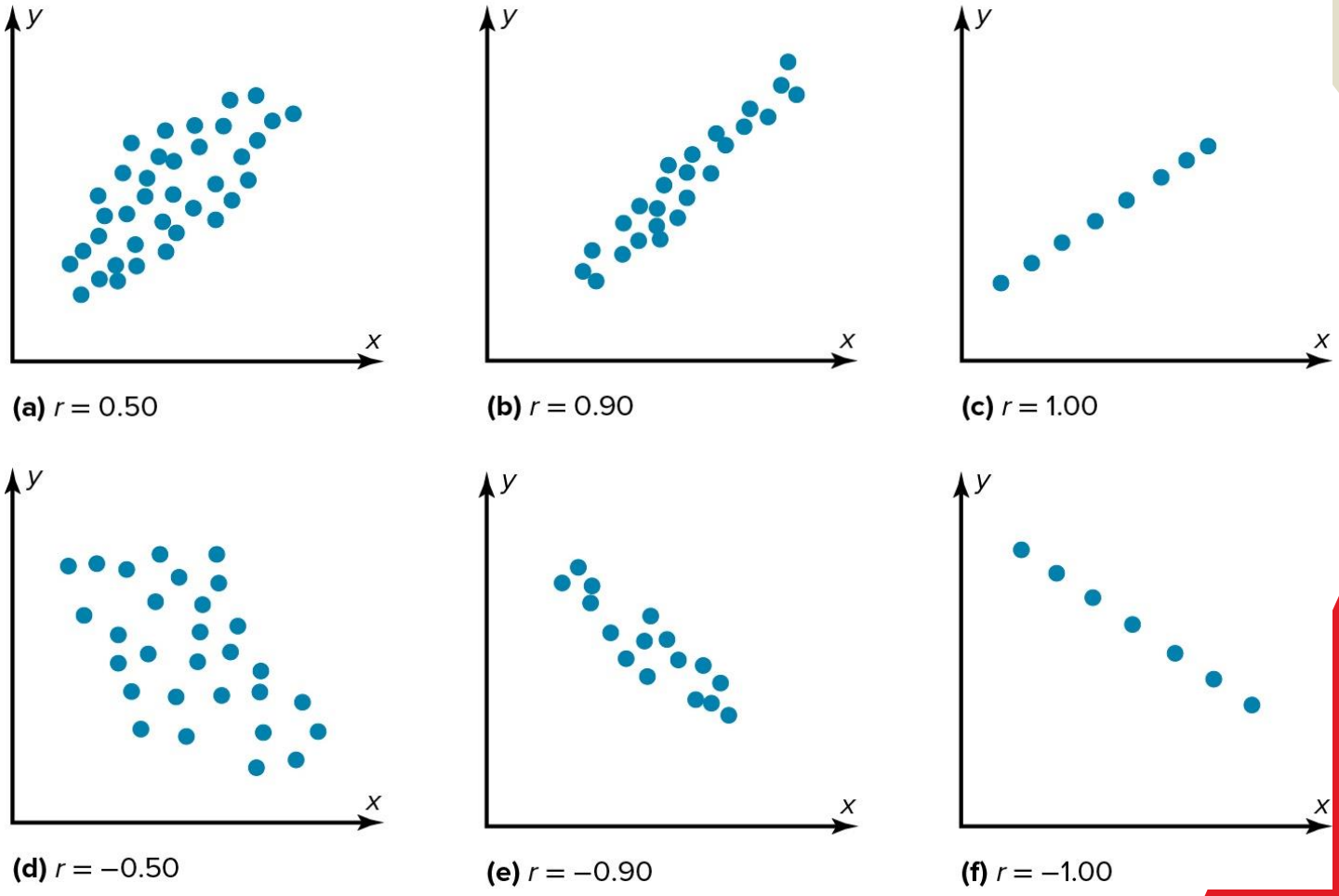
(d) No relationship

Correlation Coefficients 1

- The **correlation coefficient r** is a numerical measure of the strength of the linear relationship between two variables.
  - Linear relationship between two quantitative variables
- **Pearson product moment correlation coefficient** (PPMC).
- The correlation coefficient ranges from −1 to 1.
  - Values close to −1: A strong negative linear relationship
  - Values close to 1: A strong positive linear relationship.
  - Values close to 0: No relationship between the variables, or a non-linear relationship.



Correlation Coefficients 2



Properties of the Linear Correlation Coefficient

- The correlation coefficient is a unitless measure.
- The value of r is always between −1 and 1.
- The value of r is sensitive to outliers and can change dramatically in their presence.
- For the correlation coefficient to be valid, three things must be true:
  - The sample must be random.
  - The data pairs fall approximately on a straight line, measured at the interval or ratio level.
  - The variables should have a **bivariate normal distribution**. This means that if x is fixed, y should be normally distributed, and if y is fixed, then x is normally distributed.

Significance of the Linear Correlation Coefficient 1

- Hypothesis test to make this decision,
  - As long as data are quantitative AND,
  - Come from a random sample,
  - Scatter plot shows a rough linear trend, there are no outliers, and the variables *x* and *y* both come from normally distributed populations.

- Our hypotheses are

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0.$$

- There are three methods of testing the significance of the correlation coefficient.
- 1. We can use a *t* test with

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

This test has $n - 2$ degrees of freedom, where n is the number of ordered pairs (x, y).

- 2. Perform this test by finding the *P*-value of the test statistic.
- 3. You can look at a significance table.

Correlation vs. Causation

If the null hypothesis has been rejected, one of the following five things could be true:

- The variable *x* causes the change in the variable *y*.
- The variable *y* causes the change in the variable *x.* In this situation, the researcher has confused cause and effect.
- The relationship between *x* and *y* is caused by a third **lurking variable** that was not studied.
- There could be a complex interrelationship between *x*, *y*, and many other variables.
- The relationship is coincidental. In this case we have made a type I error.

Remember, correlation does not necessarily imply causation. But, with careful research design, and elimination of lurking variables and other alternate explanations, researchers can find good evidence to support causation, if it is there.

Regression Lines

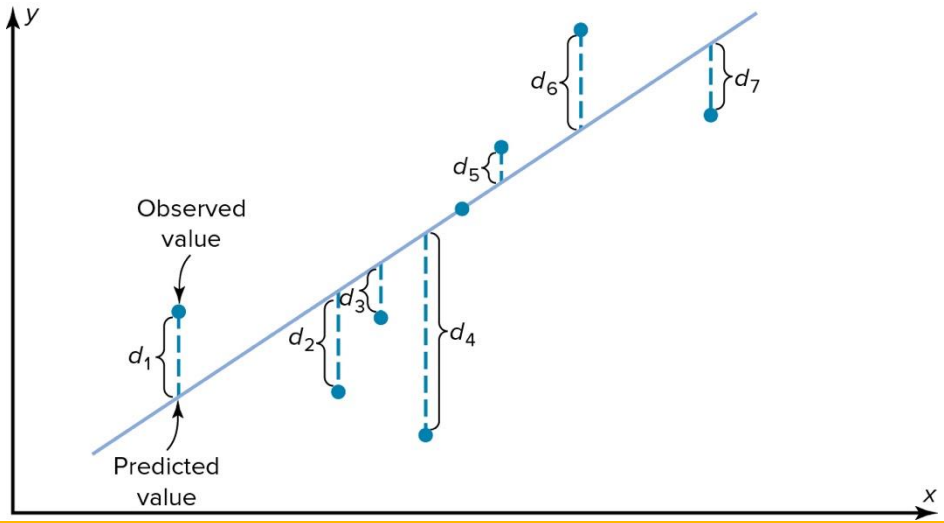
- The **regression line** provides the data's line of best fit to predict the value of the dependent variable from the value of the independent variable.

- The regression line has the form

$$y' = a + bx$$

where b is the slope, a is the y-intercept, x is a value of the independent variable, and y' is the **predicted value** of y for the given value of x.

- The **residual** is the difference (y - y') between the predicted and actual values.

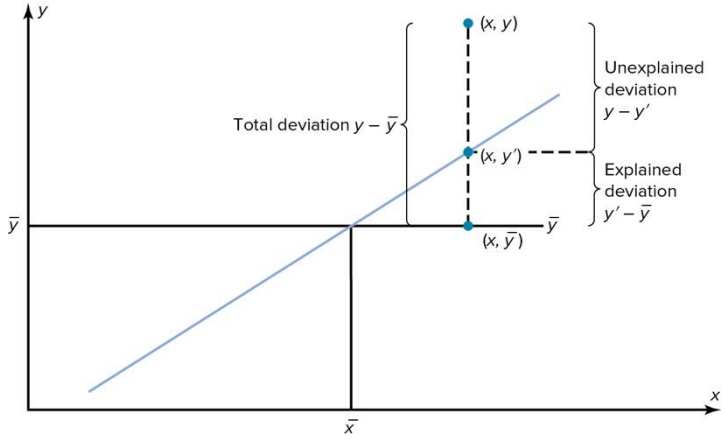


Explained and Unexplained Deviation 1

- Consider the following regression model:

X	1	2	3	4	5
Y	10	8	12	16	20

- The regression line for this data is $y' = 4.8 + 2.8x$, and $r = 0.919$.
- The **total variation** $\sum(y-\bar{y})^2$ is the sum of the squares of the vertical distances each point is from the mean of *y*.
- The total variation consists of the **explained variation** $\sum(y'-\bar{y})^2$, which is the sum of the squares of the differences between the mean of *y* and the predicted values of *y*, and the **unexplained variation** $\sum(y-y')^2$, which is the sum of the squares of the differences between the actual values of *y* and the predicted values of *y*.



Explained and Unexplained Deviation 2

- Total variation $\sum(y-\bar{y})^2 = \sum(y'-\bar{y})^2 + \sum(y-y')^2$ (Explained + Unexplained variations)

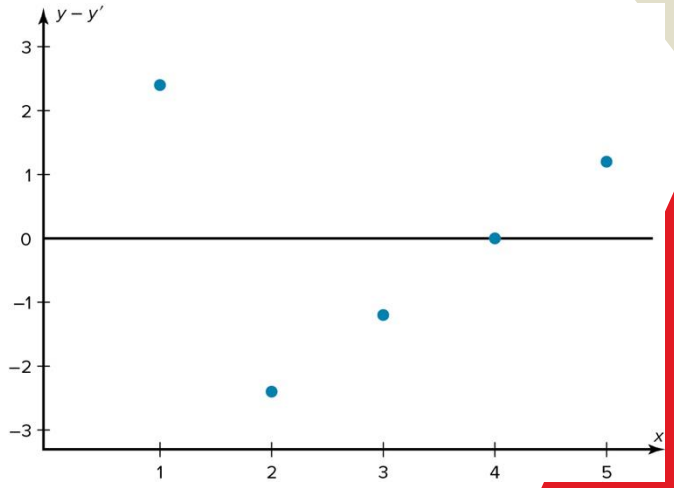
- When high correlation coefficient (+/-1):
  - The unexplained variation is small,
  - Most of the total variation is the explained variation.

- When low correlation coefficient (close to 0):
  - The explained variation very small
  - Most of the total variation is the unexplained variation.

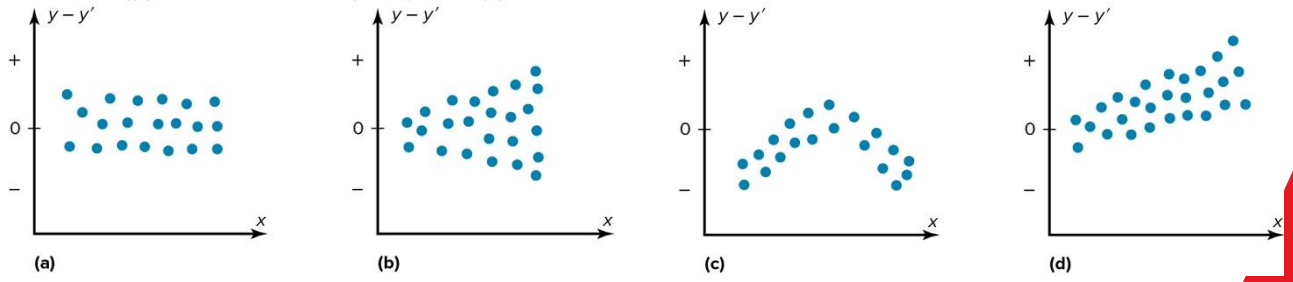
Residual Plots 1

- We can plot the residuals of each y value against the corresponding x value. The resulting plot is called a **residual plot**.
- In the previous example we had the following.

x	y	y'	$y - y'$
1	10	7.6	2.4
2	8	10.4	-2.4
3	12	13.2	-1.2
4	16	16	0
5	20	18.8	1.2



Any pattern in the residual plot suggests that the relationship between x and y is not linear and that the regression line is a bad model for making predictions.



Coefficient of Determination

- The **coefficient of determination** is the ratio of explained variation to total variation.

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

- If close to 0:
  - Model cannot explain the variation and is not useful for predictions.

- If close to 1:
  - Most of the variation is explained variation, and the model is useful for predictions.

$$r^2 = \frac{78.4}{92.8} = 0.845$$

- This means 84.5% of the total variation is explained by the regression line.

- Notice that the coefficient of determination is the square of the correlation coefficient

Multiple Regression

- In multiple regression:
  - Several independent variables and one dependent variable

$$y' = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- Suppose you wanted to predict the score y of students on a state nursing certification exam based on their GPA x_1 and age x_2.

GPA	3.2	2.7	2.5	3.4	2.2
Age	22	27	24	28	23
Board Score	550	570	525	670	490

$$y' = -44.81 + 87.64 x_1 + 14.533 x_2$$

- Predict score of a 25 year old student with a 3.0 GPA:

$$y' = -44.81 + 87.64(3.0) + 14.533(25) = 581.44$$

Five Assumptions for the Multiple Regression Equation

- For any specific values of the independent variables, the values of the *y* variable are normally distributed.
- The variances of the *y* variables are the same for each value of the independent variables.
- There is a linear relationship between the independent variables and the dependent variable.
- The independent variables are not correlated.
- The values for the *y* variable are independent.

Multiple Correlation Coefficient 1

Suppose we have two independent variables x_1 and x_2, and a dependent variable y

- r_{yx_1} is the correlation coefficient for x_1 and y.
- r_{yx_2} is the correlation coefficient for x_2 and y.
- $r_{x_1x_2}$ is the correlation coefficient for x_1 and x_2.

The **multiple correlation coefficient *R*** is then defined as

$$R = \sqrt{\frac{r_{yx_1}{}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

The multiple correlation coefficient is always between 0 and 1.

R^2 is the **coefficient of multiple determination** and it is the amount of variation explained by the regression model.

- The **adjusted R^2** is smaller
- Smaller because it corrects for n and k

- The formula for adjusted

$$R_{\text{adj}}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}$$

Dealing With Missing Values

- The rule of thumb is that above 25-33% missing values
  - Remove the variable
  - Remove missing entry

 **However, be extremely careful** 

- There is no one-size-fits-all approach; it is context-dependent
- If you impute, you are entering noise into your data set
  - Verify the effect on the distribution
- If you don't impute, you might reduce your statistical power

- For numerical data:
  - Impute with the mean
  - Impute with mode
  - Impute with median
  - Impute with regression/predictive model

- For categorical/factorial variable:
  - Impute with the mode
  - Impute with logistic/predictive model

Activity

- This dataset tracks how users interact with a wellness subscription platform from signup through trial and, potentially, the purchase of a 3-month plan.
- It captures user authenticity, registration method, demographics, and detailed engagement metrics across key marketing pages.
- The data is designed to analyze how browsing behavior and trials influence conversion to paid subscriptions.

user id	signup date	user type	trial date	phone number	registration type	age	three month subscription	homepage_visits	testimonial page visits	benefit page visits	homepage minutes	testimonial page minutes	benefit page minutes
95822412	2024-02-27	real_customer		4163119785	email	26	2024-03-06	22	17	2	43.5	9.9	1.5
77827638	2024-01-14	real_customer		4733616459	single_sign_on	55	2024-01-29	25	5	22	38.9	6	22.8
23718431	2024-02-17	disguised_bot	2024-02-23	9776552803	single_sign_on	56		15	17	3	44	24.6	5.6
87490893	2024-04-08	real_customer	2024-04-19	5231494220	email	36	2024-05-14	28	7	3	40.6	11.4	7.8
59684848	2024-06-30	real_customer		4131575764	email	58	2024-07-20	6	14	12	7	39.4	26.6
53524491	2024-01-29	real_customer	2024-02-11	2149938334	single_sign_on	43		7	18	22	9	38.5	32.8
29175900	2024-05-15	real_customer	2024-05-27	7805745017	single_sign_on	65		29	18	12	40.7	53.9	10.1
22201654	2024-01-25	real_customer	2024-02-05	5567816720	email	61	2024-03-02	13	19	14	23.7	55.6	37.1
25374874	2024-10-01	disguised_bot		5974249674	single_sign_on	45	2024-10-11	1	8	16	2.4	14.2	12.3

Activity

user id	signup date	user type	trial date	phone number	registration type	age
95822412	2024-02-27	real_customer		4163119785	email	26
77827638	2024-01-14	real_customer		4733616459	single_sign_on	55
23718431	2024-02-17	disguised_bot	2024-02-23	9776552803	single_sign_on	56
87490893	2024-04-08	real_customer	2024-04-19	5231494220	email	36
59684848	2024-06-30	real_customer		4131575764	email	58
53524491	2024-01-29	real_customer	2024-02-11	2149938334	single_sign_on	43
29175900	2024-05-15	real_customer	2024-05-27	7805745017	single_sign_on	65
22201654	2024-01-25	real_customer	2024-02-05	5567816720	email	61
25374874	2024-10-01	disguised_bot		5974249674	single_sign_on	45

three month subscription	homepage_visits	testimonial page visits	benefit page visits	homepage minutes	testimonial page minutes	benefit page minutes
2024-03-06	22	17	2	43.5	9.9	1.5
2024-01-29	25	5	22	38.9	6	22.8
	15	17	3	44	24.6	5.6
2024-05-14	28	7	3	40.6	11.4	7.8
2024-07-20	6	14	12	7	39.4	26.6
	7	18	22	9	38.5	32.8
	29	18	12	40.7	53.9	10.1
2024-03-02	13	19	14	23.7	55.6	37.1
2024-10-11	1	8	16	2.4	14.2	12.3

Teams:

Mandatory:

- Please send me an email with your 3-4 team members by Monday, 6 PM (1 email per team)
  - 6 teams of 3; 2 teams of 4
  - After that, I will assign you to teams
  - I will send each team an email on Tuesday, January 13th, confirming the teams

Datasets to avoid:

- Sleep Health and Lifestyle
- Screen Time vs Mental Wellness vs Stress
- Student Exam Performance
- Academic Stress

Dataset should have:

- At least 1000 rows
- Minimum of 5 numerical and 5 categorical variables
- If a dataset has +5000 rows, feel free to take a subset for computational power

Questions

ALY6015 Course Tutor – Melika Zandi

Course Tutor: Melika Zandi
Course: ALY6015 – Spring 2025 Term A
Availability:

Sunday: 10:00 Am – 11:00 Am (PT) - Virtual

Tuesdays: 11:00 AM – 3:00 PM (PT) - Virtual

Wednesday: 11:00 AM – 2:00 PM (PT) – in-person (Vancouver Campus) / Virtual

Thursdays: 10:00 AM – 1:00 PM (PT) - in-person(Vancouver Campus) / Virtual

Saturday: 10:00 AM – 11:00 AM (PT) - Virtual
Contact: melika.zandi@northeastern.edu or via Microsoft Teams