

Korrelation och regression

Mathias Johansson

Kristofer Söderström

2025-12-15

Innehållsförteckning

1	Introduktion	1
2	Korrelation	2
2.1	Korrelationskoefficienten (Pearson's r)	2
2.2	Beräkning	2
2.3	Tolkning	2
2.4	Exempel	2
3	Enkel linjär regression	2
3.1	Regressionsekvationen	3
3.2	Beräkna koefficienter	3
3.3	Exempel (fortsättning)	3
4	Residualer (fel)	3
5	Determinationskoefficienten R^2	3
6	Inferentiell statistik	4
6.1	Hypotestest för korrelation	4
6.2	Konfidensintervall för lutningen	4
7	Antaganden	4
8	Korrelation vs kausalitet	4
8.1	Klassiska exempel	5
9	Viktiga begrepp	5
10	Interaktivt verktyg	5
11	Referenser	5

1 Introduktion

Denna lektion behandlar två centrala statistiska metoder för att analysera samband mellan variabler: korrelation och regression. Vi går igenom beräkningar, tolkning och inferentiell statistik.

2 Korrelation

Korrelation mäter styrkan och riktningen på ett linjärt samband mellan två variabler.

2.1 Korrelationskoefficienten (Pearson's r)

Korrelationskoefficienten r varierar mellan -1 och +1:

- $r = +1$: Perfekt positivt linjärt samband
- $r = 0$: Inget linjärt samband
- $r = -1$: Perfekt negativt linjärt samband

2.2 Beräkning

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

2.3 Tolkning

Värde på r	Tolkning
0.7 till 1.0	Stark positiv korrelation
0.3 till 0.7	Måttlig positiv korrelation
-0.3 till 0.3	Svag eller ingen korrelation
-0.7 till -0.3	Måttlig negativ korrelation
-1.0 till -0.7	Stark negativ korrelation

2.4 Exempel

Anta att vi har data för studietid (X) och tentamensresultat (Y):

Student	Studietid (h)	Resultat (%)
1	2	55
2	4	65
3	6	75
4	8	85
5	10	90

Med dessa data får vi $r = 0.98$, vilket indikerar en stark positiv korrelation.

3 Enkel linjär regression

Regression används för att förutsäga värdet på en beroende variabel (Y) baserat på en oberoende variabel (X).

3.1 Regressionsekvationen

$$\hat{y} = a + bx$$

- \hat{y} = förutsagt värde
- a = intercept (y-värde när $x = 0$)
- b = lutning (förändring i y per enhet x)

3.2 Beräkna koefficienter

Lutning (b):

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Intercept (a):

$$a = \bar{y} - b\bar{x}$$

3.3 Exempel (fortsättning)

Med data från exemplet ovan:

- Medelvärde X: $\bar{x} = 6$
- Medelvärde Y: $\bar{y} = 74$

Beräkningar ger: - b 4.5 - a 47

Regressionsekvation: $\hat{y} = 47 + 4.5x$

Detta betyder att för varje extra timme studietid förväntas resultatet öka med 4.5 procentenheter.

4 Residualer (fel)

Residualen är skillnaden mellan det observerade värdet och det värde som regressionen förutsäger:

$$e_i = y_i - \hat{y}_i$$

- Positiv residual: Observerat värde är högre än förutsagt
- Negativ residual: Observerat värde är lägre än förutsagt

Minsta-kvadratmetoden minimerar summan av de kvadrerade residualerna.

5 Determinationskoefficienten R^2

$$R^2 = r^2$$

R^2 anger andelen av variansen i Y som förklaras av X:

- $R^2 = 0.81$ betyder att 81% av variationen i Y förklaras av X
- Värdet varierar från 0 (ingen förklaring) till 1 (perfekt förklaring)

6 Inferentiell statistik

6.1 Hypotestest för korrelation

Nollhypotes: $H_0 : \rho = 0$ (ingen korrelation i populationen)

Teststatistika:

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

- Frihetsgrader: $df = n - 2$
- Om $p < 0.05$ förkastar vi nollhypotesen och säger att korrelationen är statistiskt signifikant

6.2 Konfidensintervall för lutningen

$$b \pm t_{\alpha/2} \cdot SE_b$$

Standardfelet för b beräknas som:

$$SE_b = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

där s_e är standardfelet för residualerna.

Ett 95% konfidensintervall som inte innehåller 0 indikerar ett signifikant samband.

7 Antaganden

För att regressionsanalysen ska vara giltig bör följande antaganden uppfyllas:

1. **Linjäritet** – Sambandet mellan X och Y är linjärt
2. **Oberoende** – Observationerna är oberoende av varandra
3. **Normalfördelning** – Residualerna är normalfördelade
4. **Homoskedasticitet** – Variansen i residualerna är konstant

8 Korrelation vs kausalitet

Viktigt: Korrelation innebär inte orsakssamband (kausalitet).

Ett observerat samband kan bero på:

- **Omvänd kausalitet:** Y orsakar X, inte tvärtom
- **Confounding:** En tredje variabel påverkar både X och Y
- **Slump:** Särskilt vid små urval

8.1 Klassiska exempel

- Glasskonsumtion korrelerar med drunkningsolyckor
 - Bakomliggande variabel: Sommarväder
- Antalet brandmän korrelerar med brandens storlek
 - Kausaliteten går åt andra hållet

9 Viktiga begrepp

Svenska	Engelska
Korrelation	Correlation
Regression	Regression
Lutning	Slope
Intercept	Intercept
Residual	Residual
Determinationskoefficient	Coefficient of determination
Standardfel	Standard error
Konfidensintervall	Confidence interval
Signifikans	Significance

10 Interaktivt verktyg

Använd det [interaktiva regressionsverktyget](#) för att:

- Skapa egna datapunkter genom att klicka
- Se regressionslinjen uppdateras i realtid
- Visualisera residualer (fel)
- Utforska hur olika punktmönster påverkar korrelation och R^2

11 Referenser

- Wikipedia: [Pearson correlation coefficient](#)
- Wikipedia: [Simple linear regression](#)
- Wikipedia: [Least squares](#)
- Wikipedia: [Coefficient of determination \(\$R^2\$ \)](#)
- Wikipedia: [Errors and residuals](#)
- Wikipedia: [Regression analysis](#)
- Wikipedia: [Correlation does not imply causation](#)
- Wikipedia: [Standard error](#)
- Wikipedia: [Confounding](#)