Lab 13 (Last lab)

Correlation and Covariance II

Applied Statistics and Experiments

November, 2024

Agenda

1. Properties of expected value of random variables
2. Properties of covariance and correlation
3. Correlation coefficient vs Coefficient of determination

Lecture Recap

https://quizizz.com/join

Join and enter
game code

Properties of Expected value

- Expected value of a constant:
  - $E(E(X)) = E(X)$
- Linearity:
  - $E(X + Y) = E(X) + E(Y)$
  - $E(aX) = aE(X)$
- For independent X and Y
  - $E(XY) = E(X)E(Y)$

Properties of Variance and Covariance

- $Var(X) = E\left(X^2\right) - E(X)^2$
- $Var(cX) = c^2 Var(X)$
- $Var(X) = 0 \; if \; X = c, \; otherwise \; Var(X) \geqslant 0$
- $Var(X + c) = Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$

Properties of Variance and Covariance

- $Cov(X, X) = Var(X)$
- $If\ X\ and\ Y\ are\ independent\ then\ Cov(X, Y) = 0$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(aX, Y) = aCov(X, Y)$ for any constant a
- $Cov(X + c, Y) = Cov(X, Y)$ for any constant c
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} bjY_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j)$

Properties of Variance and Covariance

Show that $Cov(X,Y) = E(XY) - E(X) E(Y)$

Properties of Variance and Covariance

Show that Cov(X,Y) = E(XY) - E(X) E(Y)

$$Cov(X, Y) = E(XY) - E(X)\,E(Y)$$

$$
\begin{aligned}
Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] && (from\ definition\ of\ covariance)\\
&= E[XY - XE(Y) - YE(X) + E(X)E(Y)] && (from\ Binomial\ theorem)\\
&= E(XY) - E(X\,.\,E(Y)) - E(Y.\,E(X)) + E(E(X)E(Y))\\
&= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y)\\
&= E(XY) - E(X)\,E(Y)
\end{aligned}
$$

where

$$E(X + Y) = E(X) + E(Y)$$

$$E(X.\,E(Y)) = E(X).\,E(Y)$$

$$E(E(X).\,E(Y)) = E(X).\,E(Y)$$

Properties of Variance and Covariance

Show that $Var(X, Y) = Var(X)+Var(Y)+2Cov(X, Y)$

Properties of Variance and Covariance

Show that Var(X+Y) = Var(X)+Var(Y)+2Cov(X, Y)

$$
\begin{aligned}
Var(X+Y) &= E((X+Y)^2 - E(X+Y)^2 \\
&= E\left(X^2+Y^2+2XY\right) - (E(X)+E(Y))^2 \\
&= E\left(X^2\right)+E\left(Y^2\right)+2E(XY) - E(X)^2 - E(Y)^2 - 2E(X)E(Y) \\
&= \left[E\left(X^2\right)-E(X)^2\right] + \left[E\left(Y^2\right)-E(Y)^2\right] + 2[E(XY)-E(X)E(Y)] \\
&= Var(X)+Var(Y)+2Cov(X,Y)
\end{aligned}
$$

Properties of Variance and Covariance

Show that $Var(X+Y) + Var(X-Y) = 2*(Var(X) + Var(Y))$

Properties of Variance and Covariance

Show that Var(X+Y) + Var(X-Y) = 2*(Var(X) + Var(Y))

$$
\begin{aligned}
Var(X+Y) + Var(X-Y) &= [Var(X) + Var(Y) + 2Cov(X,Y)] + [Var(X) + Var(-Y) + 2Cov(X,-Y)] \\
&= Var(X) + Var(Y) + 2Cov(X,Y) + Var(X) + Var(Y) - 2Cov(X,Y) \\
&= 2*(Var(X) + Var(Y))
\end{aligned}
$$

where

$$Cov(X,-Y) = -Cov(X,Y)$$

$$Var(-Y) = Var(Y)$$

Properties of Variance and Covariance

Show that Show that $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$

Properties of Variance and Covariance

Show that Show that Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)

$$
\begin{aligned}
& = E[(X+Y)*Z] - E(X+Y)*E(Z) \\
& = E(XZ+YZ) - [E(X)+E(Y)]*E(Z) \\
Cov(X+Y,Z) & = E(XZ)+E(YZ) - E(X)*E(Z) - E(Y)*E(Z) \\
& = [E(XZ)-E(X)E(Z)] + [E(YZ)-E(Y)E(Z)] \\
& = Cov(X,Z)+Cov(Y,Z)
\end{aligned}
$$

Properties of Variance and Covariance

Show that Show that $Cov(aX, Y) = aCov(X, Y)$

Home assignment.

Properties of the correlation coefficient

- $-1 \leqslant \rho(X, Y) \leqslant 1$
- $if\ \rho(X, Y) = 1,\ then\ Y = aX + b,\ where\ a > 0$
- $if\ \rho(X, Y) = -1,\ then\ Y = aX + b,\ where\ a < 0$
- $\rho(aX, Y) = \rho(X, Y)\ where\ a > 0$
- $\rho(aX, Y) = -\rho(X, Y)\ where\ a < 0$
- $\rho(aX + b, cY + d) = \rho(X, Y),\ where\ a,\ c > 0$

Properties of the correlation coefficient

Show that $\rho(aX + b, cY + d) = \rho(X, Y),\ where\ a,\ c > 0$

Properties of the correlation coefficient

$$\rho(aX+b, cY+d) = \frac{Cov(aX+b, cY+d)}{\sqrt{Var(aX+b)*Var(cY+d)}} = \frac{Cov(aX,\ cY)}{\sqrt{a^2Var(X)*c^2Var(Y)}} = \frac{a*c*Cov(X,Y)}{a*c*\sqrt{Var(X)*Var(Y)}} = \frac{Cov(X,Y)}{\sqrt{Var(X)*Var(Y)}} = \rho(X,Y) \qquad (a, c > 0)$$

Coefficient of Determination R^2 vs. Coefficient of Correlation r

Simple linear regression

Coefficient of Determination R^2 vs. Coefficient of Correlation r

$$y = \widehat{y} + e$$
$$\widehat{y} = ax + b \text{ (simple linear regression)}$$

$$\overline{y} = \frac{\sum y_i}{n}$$

$$a = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \overline{x})(x_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

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

Simple linear regression

Coefficient of Determination R² vs. Coefficient of Correlation r

$$R^2 = \frac{ESS}{TSS} = \frac{1 - SS_{res}}{SS_{tot}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(ax_i + b - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(ax_i + \bar{y} - a\bar{x} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

$$= a^2\frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = a^2\frac{SS_{xx}}{SS_{yy}} = \left(\frac{SS_{xy}}{SS_{xx}}\right)^2 \cdot \frac{SS_{xx}}{SS_{yy}} = \frac{(SS_{xy})^2}{SS_{yy} * SS_{xx}} = \frac{\left(\sum(x_i - \bar{x})(x_i - \bar{y})\right)^2}{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}$$

$$= \frac{\left(\sum(x_i - \bar{x})(x_i - \bar{y})\right)^2}{\left(\sqrt{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}\right)^2} = \left(\frac{\sum(x_i - \bar{x})(x_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}}\right)^2 = \left(\frac{Cov(X, Y)}{\sqrt{Var(X) * Var(Y)}}\right)^2$$

$$= r^2$$

Q. What is the coefficient of correlation (r)?

A. The coefficient of correlation measures the direction and strength of the linear relationship between two variables, ranging from -1 to 1.

Q. What is the coefficient of determination (R^2)?

A. The coefficient of determination represents the variance proportion in a dependent variable explained by an independent variable, ranging from 0 to 1.

Q. How do you calculate the coefficient of determination?

A. The determination coefficient is the correlation coefficient squared: $R^2 = r^2$.

Q. Can correlation imply causation?

A. No, correlation does not necessarily mean causation, as confounding factors may be involved.

Q. Does a low correlation coefficient always indicate no relationship between variables?

A. No, a low correlation coefficient could indicate a nonlinear relationship rather than the absence of a relationship

Q. What do positive and negative r values signify?

A. Positive r values indicate a direct relationship, while negative values represent an inverse relationship between variables.

Q. What do R^2 values closer to 1 and 0 mean?

A. R^2 values closer to 1 indicate stronger model explanatory power; values closer to 0 suggest weaker explanatory power.

Q. Are R^2 and r interchangeable?

A. No, R^2 and r serve different purposes and should not be used interchangeably.

   a. Pearson correlation coefficient (r) is used to identify patterns in things whereas the coefficient of determination (R^2) is used to identify the strength of a model.

   b. r values ranges from -1 to +1 while R^2 ranges between 0 to +1.

Q. When should I use the coefficient of correlation and coefficient of determination?

A. Use these coefficients to assess the relationship between variables, determine model effectiveness, and inform data-driven decision-making.

References

- https://www.probabilitycourse.com/chapter5/5_3_1_covariance_correlation.php
- https://www.colorado.edu/amath/sites/default/files/attached-files/ch5_covariance_0.pdf
- https://online.stat.psu.edu/stat414/book/export/html/728
- https://www.wolframalpha.com/input?i=Covariance
- https://statisticseasily.com/coefficient-of-determination-vs-coefficient-of-correlation/

Attendance

https://baam.duckdns.org

Questions?