Regression

Covariance and Correlation, the Bivariate Normal Distribution

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Covariance and correlation

- Recall the *covariance* of random variables X and Y

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

- The covariance depends on the units of the variables.
- Make independent of the units by dividing by σ_X and σ_Y, to obtain the *correlation coefficient*,

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

- This also has the effect of ensuring $\rho(X,Y) \in [-1,+1]$.

Proof that $\rho(X, Y) \in [-1, +1]$

- Define the standardized r.v.s, $X^* = \frac{X-\mu_X}{\sigma_X}$ and $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$
- Hence $E(X^*) = E(Y^*) = 0$ and $\mathrm{Var}(X^*) = \mathrm{Var}(Y^*) = 1$.
- Now consider

$$
\begin{aligned}
0 \leq \mathrm{Var}(X^* \pm Y^*) =& E\left[(X^*)^2\right] + 2E(X^* Y^*) + E\left[(Y^*)^2\right] \\
=& \mathrm{Var}(X^*) \pm 2\mathrm{Cov}(X^*, Y^*) + \mathrm{Var}(Y^*) \\
=& 2 \pm 2\rho(X, Y).
\end{aligned}
$$

- It follows that $-1 \leq \rho(X, Y) \leq +1$. $\square$

- We proved above earlier using Cauchy-Schwarz inequality.

The Pearson correlation coefficient

- We have defined the correlation

$$\rho(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

- Define the *sample correlation coefficient* by replacing expectation values of moments with sample moments

$$R = \frac{\frac{1}{n}\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\frac{1}{n}\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\frac{1}{n}\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$

- In terms of sampled data points (x_j, y_j), this is written

$$r = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_i^n (x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_i^n (y_i - \overline{y})^2}}$$

Relating $\hat{\beta}_1$ and r

- At this point, we have

$$r = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_i^n (x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_i^n (y_i - \overline{y})^2}}$$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_i^n (x_i - \overline{x})^2}.$$

- Eliminating numerators yields relation between $\hat{\beta}_1$ and r,

$$\hat{\beta}_1 = r\sqrt{\frac{\sum_i^n (y_i - \overline{y})^2}{\sum_i^n (x_i - \overline{x})^2}}$$

Interpreting r

- Mean square error due to lack of linearity

$$\begin{aligned}\sum_i^n (y_i - \hat{y}_i)^2 &= \sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\ &= \sum_i^n \left[y_i - \left(\overline{y} - \hat{\beta}_1 \overline{x}\right) - \hat{\beta}_1 x_i\right]^2 \\ &= \sum_i^n \left[(y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})\right]^2,\end{aligned}$$

where we used $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$.

- Expand the above and use $\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2}$ to find

$$\begin{aligned}\sum_i^n (y_i - \hat{y}_i)^2 &= \sum_i^n (y_i - \overline{y})^2 - 2\hat{\beta}_1 \sum_i^n (x_i - \overline{x})(y_i - \overline{y}) + \hat{\beta}_1^2 \sum_i^n (x_i - \overline{x})^2 \\ &= \sum_i^n (y_i - \overline{y})^2 - \hat{\beta}_1^2 \sum_i^n (x_i - \overline{x})^2\end{aligned}$$

Interpreting r

- Relationship between $\hat{\beta}_1$ and r

$$\hat{\beta}_1 = r\sqrt{\frac{\sum_i^n (y_i - \overline{y})^2}{\sum_i^n (x_i - \overline{x})^2}}$$

- Mean square error due to lack of linearity

$$\sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n (y_i - \overline{y})^2 - \hat{\beta}_1^2 \sum_i^n (x_i - \overline{x})^2$$

- Eliminating $\hat{\beta}_1$ and solving for r yields

$$r^2 = \frac{\sum_i^n (y_i - \overline{y})^2 - \sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\sum_i^n (y_i - \overline{y})^2}.$$

Interpreting r

- Define the *coefficient of determination*

$$r^2 = \frac{\sum_i^n (y_i - \overline{y})^2 - \sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\sum_i^n (y_i - \overline{y})^2}.$$

- This admits a simple interpretation
  - $\sum_i^n (y_i - \overline{y})^2$ is the *total variability* in y.
  - $\sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ is the variability that can not be explained by linear regression.
  - The numerator of r^2 is the variability that *can* be explained by linear regression.
  - The quantity r^2 is the fraction of the variability that can be explained by regression.

The bivariate normal distribution

- Two-dimensional integral of the exponential of a negative-definite quadratic form

$$I(a,b,c,\mu_X,\mu_Y) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \ \exp\left[-\left(ax^2+2bxy+cy^2\right)\right]$$

- Try to demand that

$$\begin{aligned} ax^2+2bxy+cy^2 =& \mu\left[(x+\kappa y)^2+(y+\lambda x)^2\right] \\ =& \mu\left[\left(1+\lambda^2\right)x^2+2\left(\kappa+\lambda\right)xy+\left(1+\kappa^2\right)y^2\right] \end{aligned}$$

- If true for all $x, y \in \mathbb{R}$, we must have

$$\begin{aligned} a =& \mu\left(1+\lambda^2\right) & \lambda =& \frac{a \pm \sqrt{ac-b^2}}{b} \\ b =& \mu\left(\kappa+\lambda\right) & \kappa =& \frac{c \pm \sqrt{ac-b^2}}{b} \\ c =& \mu\left(1+\kappa^2\right) & \mu =& \frac{b^2}{a+c \pm 2\sqrt{ac-b^2}} \end{aligned}$$

Bruce M. Boghosian

- We have demonstrated that

$$ax^2 + 2bxy + cy^2 = \mu\left[(x+\kappa y)^2 + (y+\lambda x)^2\right] = \mu\left(\xi^2 + \eta^2\right)$$

where the new variables are $\xi = x + \kappa y$ and $\eta = y + \lambda x$.

- The old and new constants are related as follows

$$\begin{aligned} a &= \mu\left(1+\lambda^2\right) & \lambda &= \frac{a \pm \sqrt{ac-b^2}}{b} \\ b &= \mu\left(\kappa+\lambda\right) & \kappa &= \frac{c \pm \sqrt{ac-b^2}}{b} \\ c &= \mu\left(1+\kappa^2\right) & \mu &= \frac{b^2}{a + c \pm 2\sqrt{ac-b^2}} > 0 \end{aligned}$$

- Jacobian is (after a bit of algebra)

$$\frac{\partial(\xi,\eta)}{\partial(x,y)} = \left|\begin{bmatrix} 1 & \kappa \\ \lambda & 1 \end{bmatrix}\right| = |1-\kappa\lambda| = \frac{\sqrt{ac-b^2}}{\mu}$$

The bivariate normal distribution

- It follows that

$$\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy\ e^{-\left(ax^2+2bxy+cy^2\right)}$$

$$=\frac{\mu}{\sqrt{ac-b^2}} \int_{-\infty}^{+\infty} d\xi \int_{-\infty}^{+\infty} d\eta\ e^{-\mu\left(\xi^2+\eta^2\right)} = \frac{\pi}{\sqrt{ac-b^2}}$$

- From this we see that

$$f_{X,Y}(x,y) = \frac{\sqrt{ac-b^2}}{\pi} e^{-\left(ax^2+2bxy+cy^2\right)}$$

is a normalized bivariate pdf for X and Y.

The bivariate normal distribution

- We have shown that

$$f_{X,Y}(x,y) = \frac{\sqrt{ac-b^2}}{\pi} e^{-\left(ax^2+2bxy+cy^2\right)}$$

is a normalized bivariate pdf for X and Y.

- To obtain form given in textbook, rename the constants

$$a = \frac{1}{2(1-\rho^2)\sigma_X^2}, \qquad b = -\frac{\rho}{2(1-\rho^2)\sigma_X\sigma_Y}, \qquad c = \frac{1}{2(1-\rho^2)\sigma_Y^2}$$

- The above becomes

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{1-\rho^2}\right)\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$

- In the above, we have shifted x and y by means, μ_X and μ_Y, which will not affect normalization.

The bivariate normal distribution

- Use the bivariate normal distribution in the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{1-\rho^2}\right)\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$

- Either in (x, y) or (ξ, η) coordinates, we can then calculate

$$E(X) = \mu_X \qquad E(X^2) = \mu_X^2 + \sigma_X^2 \qquad \text{Var}(X) = \sigma_X^2$$

$$E(Y) = \mu_Y \qquad E(Y^2) = \mu_Y^2 + \sigma_Y^2 \qquad \text{Var}(Y) = \sigma_Y^2$$

$$E(XY) = \mu_X\mu_Y + \rho\sigma_X\sigma_Y \qquad \text{Cov}(X,Y) = \rho\sigma_X\sigma_Y \qquad \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \rho$$

- We also have

$$E(Y \mid x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$$

$$\text{Var}(Y \mid x) = (1-\rho^2)\sigma_Y^2$$

Bivariate normal distribution parameter estimation

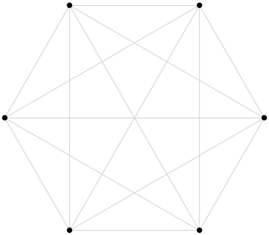
- The MLEs for μ_X, μ_Y, σ_X^2, σ_Y^2 and ρ, assuming that all five of them are unknown, are $\overline{X}$, $\overline{Y}$, $\frac{1}{n}\sum_i^n \left(X_i - \overline{X}\right)^2$, $\frac{1}{n}\sum_i^n \left(Y_i - \overline{Y}\right)^2$, and R, respectively.
- It is also possible to test the null hypothesis $H_0:\ \rho = 0$, in order to test for the presence or absence of correlation, using a T test.

Correlation and causation

- Correlation does not indicate causation.
- Two things can be correlated only because they are both correlated with a third thing that is not observed.
- Correlations can be due to the structure of what is observed.

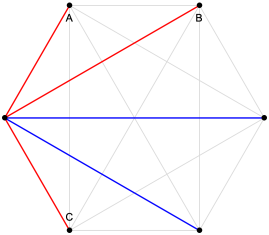
An amazing discovery?

- Cristian Calude & Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* **22**/3 (2017) 595-612.
- It turns out that in *any* party of six people, there must either be three who all know each other, or three who are all strangers.



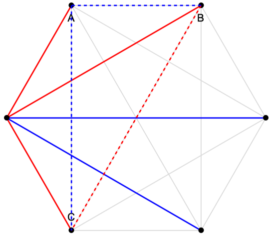
Perhaps not so amazing a discovery. . .

- Cristian Calude & Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* **22**/3 (2017) 595-612.
- It turns out that in *any* party of six people, there must either be three who all know each other, or three who are all strangers.



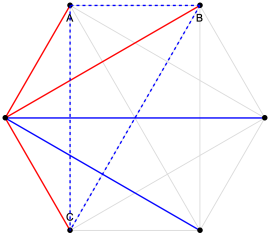
Perhaps not so amazing a discovery...

- Cristian Calude & Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* **22**/3 (2017) 595-612.
- It turns out that in *any* party of six people, there must either be three who all know each other, or three who are all strangers.



Perhaps not so amazing a discovery...

- Cristian Calude & Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* **22**/3 (2017) 595-612.
- It turns out that in *any* party of six people, there must either be three who all know each other, or three who are all strangers.



Spurious correlations!

- Cristian Calude & Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* **22**/3 (2017) 595-612.
- Ramsey theory is about finding structure and organization in sets of data.
- Ramsey numbers indicate how big a set must be to guarantee the existence of certain minimal structures:
  - $R(3,3) = 6$ (example on previous slide)
  - $R(4,5) = 25$
  - $R(3,3,3) = 17$
  - $43 \leq R(5,5) \leq 49$
- Ramsey theory explains why we tend to find structure in seemingly random sets.

Spurious correlations

- Cristian Calude & Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* **22**/3 (2017) 595-612.
- As data sets increase in size, the ratio of the number of meaningful correlations to the number of spurious correlations will tend to zero.

Spurious correlations

- Correlations can be spurious.
- Tyler Vigen web page

Summary

- We defined the *correlation* $\rho(X, Y) \in [-1, +1]$.
- We presented a method of estimating $\rho(X, Y)$ using sample moments.
- We constructed the *Pearson correlation coefficient* R.
- We have made an interpretation of the r^2 as the *coefficient of determination*.
- We have studied bivariate normal distributions.
- We have seen how to parametrize bivariate normal distributions using five parameters – two means, two standard deviations, and the correlation.
- We have discussed some of the problems associated with naive hunting for correlations.