Lecture 20: Covariance, correlation, and linear regression

Announcements:

- Reading: Chapter 7 in Vasilj.

Lecture 20: Covariance, correlation, and linear regression

Announcements:

- Reading: Chapter 7 in Vasilj.
- Problem set 11 due today.

Lecture 20: Covariance, correlation, and linear regression

Announcements:

- Reading: Chapter 7 in Vasilj.
- Problem set 11 due today.
- Today: Covariance, correlation, and linear regression

Problem. Two continuous variables: are they related?

Examples

- Parents versus offspring.

Problem. Two continuous variables: are they related?

Examples

- Parents versus offspring.
- Growth curves (organs, organisms, populations).

Problem. Two continuous variables: are they related?

Examples

- Parents versus offspring.
- Growth curves (organs, organisms, populations).
- Allometric relationships.

Problem. Two continuous variables: are they related?

Examples

- Parents versus offspring.
- Growth curves (organs, organisms, populations).
- Allometric relationships.
- Physiological relationships.

Problem. Two continuous variables: are they related?

Examples

- Parents versus offspring.
- Growth curves (organs, organisms, populations).
- Allometric relationships.
- Physiological relationships.
- Ecological relationships.

Correlation versus causation

Storks

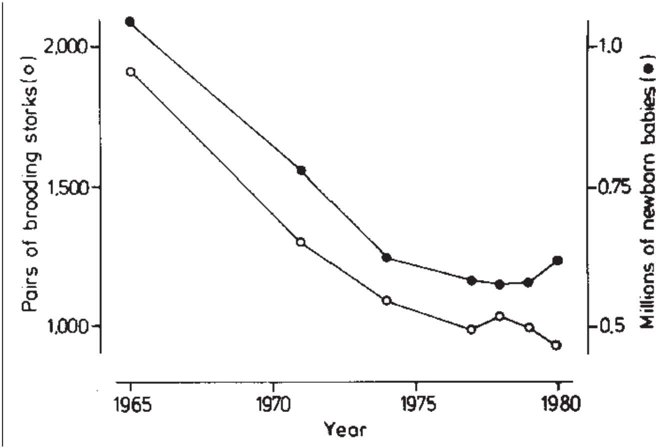
Correlation versus causation

Babies 

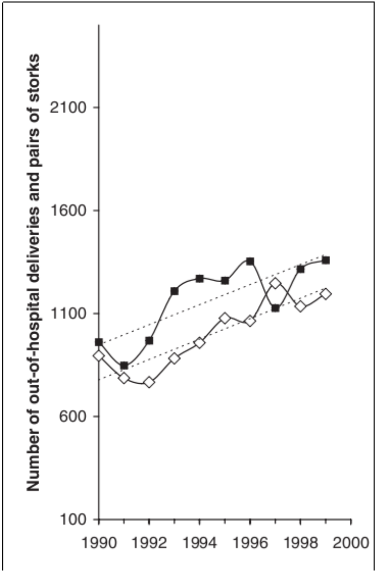
Correlation versus causation

Storks and babies?

Correlation versus causation



Correlation versus causation



Covariance

1. $Cov(x, y) = E[(x - \overline{x})(y - \overline{y})]$

Covariance

1. $Cov(x,y) = E[(x-\overline{x})(y-\overline{y})]$
2. $$Cov(x,y) = \frac{\sum_i [(x_i-\overline{x})(y_i-\overline{y})]}{n-1}$$

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.
2. $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$, for any variable x.

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.
2. $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$, for any variable x.
3. $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$, for any variables x and y.

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.
2. $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$, for any variable x.
3. $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$, for any variables x and y.
4. $\mathrm{Cov}(x, bx) = b\mathrm{Var}(x)$, for any variables x and y and constant b.

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.
2. $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$, for any variable x.
3. $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$, for any variables x and y.
4. $\mathrm{Cov}(x, bx) = b\mathrm{Var}(x)$, for any variables x and y and constant b.
5. $\mathrm{Cov}(ax, by) = ab\mathrm{Cov}(x, y)$, for any constants a and b.

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.
2. $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$, for any variable x.
3. $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$, for any variables x and y.
4. $\mathrm{Cov}(x, bx) = b\mathrm{Var}(x)$, for any variables x and y and constant b.
5. $\mathrm{Cov}(ax, by) = ab\mathrm{Cov}(x, y)$, for any constants a and b.
6. $\mathrm{Cov}(x + a, y + b) = \mathrm{Cov}(x, y)$, for any constants a and b.

Properties of the covariance

1. $\mathrm{Cov}(x, a) = 0$, where a is any constant.
2. $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$, for any variable x.
3. $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$, for any variables x and y.
4. $\mathrm{Cov}(x, bx) = b\mathrm{Var}(x)$, for any variables x and y and constant b.
5. $\mathrm{Cov}(ax, by) = ab\mathrm{Cov}(x, y)$, for any constants a and b.
6. $\mathrm{Cov}(x + a, y + b) = \mathrm{Cov}(x, y)$, for any constants a and b.
7. If x and y are independent, then $\mathrm{Cov}(x, y) = 0$.

Calculating the slope and intercept of the best-fit line

The best fit line is the line that minimizes the residual sum of squares.

1.
$$\mathrm{Slope} = b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}.$$

Calculating the slope and intercept of the best-fit line

The best fit line is the line that minimizes the residual sum of squares.

1. $$\text{Slope} = b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}.$$
2. $\text{Intercept} = a = \overline{y} - b\overline{x}.$

Correlation versus covariance

Correlation coefficient is the scaled covariance.

1. The maximum covariance between x and y is $s_x s_y$.

Correlation versus covariance

Correlation coefficient is the scaled covariance.

1. The maximum covariance between x and y is $s_x s_y$.
2. $$\mathrm{Cor}(x, y) = r = \frac{\mathrm{Cov}(x, y)}{s_x s_y}.$$

Correlation versus covariance

Correlation coefficient is the scaled covariance.

1. The maximum covariance between x and y is $s_x s_y$.
2. $$\mathrm{Cor}(x, y) = r = \frac{\mathrm{Cov}(x, y)}{s_x s_y}.$$
3. The correlation coefficient lies between -1 and $+1$.

Strength of the correlation coefficient

korelacijski koeficijent (r)	jačina korelacije
0.00 – 0.10	nema
0.10 – 0.25	vrlo slaba
0.25 – 0.40	slaba
0.40 – 0.50	srednja
0.50 – 0.75	jaka
0.75 – 0.90	vrlo jaka
0.90 – 1.00	potpuna

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.
3. Calculate the line sum of squares $SS_x = \sum(\hat{y} - \overline{y})^2$.

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.
3. Calculate the line sum of squares $SS_x = \sum(\hat{y} - \overline{y})^2$.
4. Calculate the corresponding residual and line mean squares: $MS_{res} = SS_{res}/df_{res}$; and $MS_x = SS_x/df_x$
   $df_{res} = n - 2; df_x = 1$

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.
3. Calculate the line sum of squares $SS_x = \sum(\hat{y} - \overline{y})^2$.
4. Calculate the corresponding residual and line mean squares: $MS_{res} = SS_{res}/df_{res}$; and $MS_x = SS_x/df_x$ $df_{res} = n - 2; df_x = 1$
5. Form the F ratio and calculate its tail probability $F = MS_{res}/MS_x$.

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.
3. Calculate the line sum of squares $SS_x = \sum(\hat{y} - \overline{y})^2$.
4. Calculate the corresponding residual and line mean squares: $MS_{res} = SS_{res}/df_{res}$; and $MS_x = SS_x/df_x$
   $df_{res} = n - 2; df_x = 1$
5. Form the F ratio and calculate its tail probability $F = MS_{res}/MS_x$.
6. Or let R do it all with anova(lm(y ~ x)).

Assumptions of linear regression

1. The x values are measured without any error

Assumptions of linear regression

1. The x values are measured without any error
2. The relationship between x and y is linear

Assumptions of linear regression

1. The x values are measured without any error
2. The relationship between x and y is linear
3. The residuals are normally distributed

Assumptions of linear regression

1. The x values are measured without any error
2. The relationship between x and y is linear
3. The residuals are normally distributed
4. The variance is constant (regardless of the value of x)

Interpretation of correlation