

K6 Korrelation und Regression

Korrelation mißt Assoziationen zwischen stetigen ZV.

In Regression geht es um kausale Modelle — wir versuchen die Variabilität einer ZV durch andere erklärende ZV zu reduzieren.

Hier geht es um stetige ZV und die Abhängigkeiten zwischen ihnen.

Beispiele — Streudiagramme und Korrelation

1. Tiere $\log(\text{Länge})$ gegen $\log(\text{Gebärzeit})$
2. Bankdaten (Profit und Umsatz)
3. Deutsche demographische Daten

6.1 Der Korrelationskoeffizient

ρ ist der unbekannte Korrelationskoeffizient der Grundgesamtheit

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

σ_X, σ_Y sind die Standardabweichungen von X, Y

ρ misst lineare Abhängigkeit.

r ist der Stichproben-Korrelationskoeffizient

$$r =$$

s_X, s_Y sind die Stichproben-Standardabweichungen

Es gilt

6.1.1 Korrelationskoeffizient für Normalverteilte ZV

Seien X und Y bivariat normalverteilt:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}H}$$

$$H =$$

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Nach den Transformationen

$$u = \frac{x - \mu_X}{\sigma_X} \quad \text{und} \quad v = \frac{y - \mu_Y}{\sigma_Y}$$

$$Cov(X, Y) = \frac{\sigma_X\sigma_Y}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) du dv$$

$$g(u, v) =$$

und daraus folgt

$$Cov(X, Y) = \rho\sigma_X\sigma_Y$$

6.1.2 Verteilung des Stichproben-Korrelationskoeffizients (für Normalverteilte ZV)

Für X und Y normalverteilt mit $\rho = 0$

Für X und Y normalverteilt mit $\rho \neq 0$ hat Fisher gezeigt dass

$$z = \frac{1}{2} \log \frac{1+r}{1-r}$$

ungefähr normalverteilt ist mit

$$E[z] = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$$

und

$$V[z] = \frac{1}{n-3}$$

so dass die Varianz unabhängig von ρ ist.

6.2 Assoziation zwischen stetigen Variablen

$X \rightarrow Y$ X beeinflusst Y

z.B. $X =$ Arbeitsstunden $Y =$ Gehalt

$Y \rightarrow X$ Y beeinflusst X

z.B. $X =$ Herzinfarktrate $Y =$ Weinkonsum

Aber in welcher Richtung?

z.B. $X =$ Werbung $Y =$ Umsatz

X und Y assoziiert

z.B. $X =$ Umweltverschmutzung $Y =$ Gesundheit

$Z \rightarrow Y, Z \rightarrow X$ $Z =$ Armut

X und Y zufällig assoziiert (“spurious correlation”)

z.B. $X =$ Selbstmordrate $Y =$ Anzahl neuer Priester

6.3 Regression

z. B. Größe des Sohns gegen Größe des Vaters (Galton)

Für eine lineare Assoziation haben wir das Modell

$$Y = a + bX$$

Gegeben Daten $\{(y_i, x_i)\}$, wie schätzen wir die Parameter a und b ?

6.3.1 Kleinstquadrate

$$\min_{a,b} C =$$

$$\frac{dC}{da} = 0 = -2 \sum (y_i - a - bx_i) \Rightarrow \bar{y} = a + b\bar{x}$$

$$\frac{dC}{db} = 0 \Rightarrow \sum y_i x_i = an\bar{x} + b \sum x_i^2$$

$$\hat{b} =$$

wo r ist die Stichprobenkorrelation zwischen Y und X .

$$\hat{a} =$$

$$\min C = C(\hat{a}, \hat{b}) = (n - 1)(1 - r^2)s_y^2$$

$$(n - 1)s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

die Variabilität von Y in der Stichprobe

$$\Rightarrow C(\hat{a}, \hat{b}) =$$

Das Modell $Y = a + bX$ hat r^2 von der Variabilität von Y "erklärt". Wir setzen

$$R^2 = 1 - \frac{C(\hat{a}, \hat{b})}{\sum (y_i - \bar{y})^2}$$

R^2 ist ein Gütekriterium für das Modell, das Bestimmtheitsmaß. (Für einfache lineare Regression gilt $R^2 = r^2$, sonst nicht.) Im allgemeinen gilt

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{Modell Variabilität}}{\text{Gesamt-Variabilität}}$$

wo \hat{y}_i = der Modellwert für Fall i .

6.3.2 ML-Schätzer für Regression

X gegeben, Y beobachtet

$$Y = a + bX + \epsilon$$

$$\epsilon \sim N(0, \sigma^2) \quad u.i.v.$$

$$\Rightarrow Y \sim$$

$Y|X$ ist normalverteilt, Y ist nicht unbedingt normalverteilt.

$$L(a, b; \{x_i, y_i\}, \sigma) =$$

$$\log L =$$

$\Rightarrow (\hat{a}, \hat{b})$, die ML-Schätzer, sind hier die KQ-Schätzer (angenommen σ bekannt). Unter der Annahme der unabhängigen normalverteilten "Fehler" können wir σ^2 schätzen mit

$$\begin{aligned} s^2 = \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n e_i^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2 \end{aligned}$$

$n - 2$, weil zwei Parameter geschätzt werden.

6.3.3 Eigenschaften der ML-Parameterschätzer

$$\begin{aligned}\hat{b} &= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \\ &= \sum d_i y_i\end{aligned}$$

wo

$$d_i =$$

d.h. die ZV \hat{b} , die wir mit

$$\hat{b} = \sum d_i Y_i$$

definieren, ist eine lineare Funktion von n normalverteilten ZV

$$\Rightarrow \hat{b} \sim N$$

$$E[\hat{b}] = b$$

$$V[\hat{b}] = V[\sum d_i Y_i]$$

$$= \sum d_i^2 V[Y_i]$$

(wegen der Unabhängigkeit)

$$= \sigma^2 \sum d_i^2$$

=

$$\Rightarrow s_b^2 =$$

und ähnlicherweise

$$s_a^2 =$$

aber

$$\text{Cov}[\hat{a}, \hat{b}] =$$

so dass die Schätzer für a und b nicht unabhängig sind.