

## Präsidentschaftswahl in Iran

Die Wahl fand am 12. Juni 2009 statt. Einige haben die Resultate bestritten, z.B.:

“The Devil Is in the Digits”

Bernd Beber and Alexandra Scacco

Washington Post 20. Juni 2009

[www.washingtonpost.com/wp-dyn/content/article/2009/06/20/AR2009062000004.html?hpid=opinionsbox1](http://www.washingtonpost.com/wp-dyn/content/article/2009/06/20/AR2009062000004.html?hpid=opinionsbox1)

Sie untersuchten zwei Eigenschaften aus den Ergebnissen für die 29 Provinzen. Es gab vier Kandidaten und sie analysierten zwei Aspekte der 116 Stimmzahlen in einem Datensatz.

1. Die Verteilung der letzten Ziffer.
2. Die Anzahl der nichtbenachbarten Zahlen unter den letzten zwei Ziffern.

## 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

$\rho$  ist der unbekannte Korrelationskoeffizient der Grundgesamtheit

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

$\sigma_X, \sigma_Y$  sind die Standardabweichungen von  $X, Y$

$\rho$  misst lineare Abhängigkeit.

$r$  ist der Stichproben-Korrelationskoeffizient

$$r = \frac{\frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$$

$s_X, s_Y$  sind die Stichproben-Standardabweichungen

Es gilt

$$-1 \leq \rho \leq 1 \quad \text{und} \quad -1 \leq r \leq 1$$

## 6.1.1 Korrelationskoeffizient für Normalverteilte ZV

Seien  $X$  und  $Y$  bivariat normalverteilt:

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

$$H = \left[ \frac{(x - \mu_X)^2}{\sigma_X^2} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x - \mu_X)(y - \mu_Y)}{\sigma_X\sigma_Y} \right]$$

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

Nach den Transformationen

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

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

$$g(u, v) = uv \exp \left[ -\frac{1}{2(1-\rho^2)}(u^2 + v^2 - 2\rho uv) \right]$$

und daraus folgt

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

## 6.1.2 Verteilung des Stichproben-Korrelationskoeffizients (für Normalverteilte ZV)

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

$$\sqrt{n-2} \frac{r}{1-r^2} \sim t_{n-2}$$

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  $\rho$  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

$$\min_{a,b} C = \sum (y_i - a - bx_i)^2$$

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

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

$$\hat{b} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{s_y}{s_x} r$$

wo  $r$  ist die Stichprobenkorrelation zwischen  $Y$  und  $X$ .

$$\hat{a} = \bar{y} - \frac{s_y}{s_x} r \bar{x}$$

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

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

die Variabilität von  $Y$  in der Stichprobe

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

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

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

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

$$\Rightarrow Y \sim N(a + bX, \sigma^2)$$

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

$$L(a, b; \{x_i, y_i\}, \sigma) = \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^n \prod e^{-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2}$$

$$\log L = -n \log \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2} \sum (y_i - a - bx_i)^2$$

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

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

$$= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2$$

$n - 2$ , 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

$$d_i = \frac{x_i - \bar{x}}{\sum x_i^2 - n \bar{x}^2}$$

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

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

definieren, ist eine lineare Funktion von u. normalverteilten ZV

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

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

$$\begin{aligned}V[\hat{b}] &= V[\sum d_i Y_i] \\ &= \sum d_i^2 V[Y_i]\end{aligned}$$

(wegen der Unabhängigkeit)

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

$$= \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$$

$$\Rightarrow s_b^2 = \frac{s^2}{\sum x_i^2 - n\bar{x}^2}$$

und ähnlicherweise

$$s_a^2 = s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2} \right)$$

aber

$$Cov[\hat{a}, \hat{b}] = \frac{\bar{x}s^2}{\sum x_i^2 - n\bar{x}^2}$$

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

### 6.3.4 Konfidenz und Vorhersage Intervalle

Da  $\hat{b}$  und  $\hat{a}$  normalverteilt sind, können wir KIs für  $a$  und  $b$  mit der  $t$ -Verteilung erstellen. Für  $b$ :

$$\hat{b} \pm t_{(n-2), 1-\frac{\alpha}{2}} \frac{s}{\sqrt{\sum x_i^2 - n\bar{x}^2}}$$

Um unsere Unsicherheit um des Modells auszudrücken, brauchen wir die Varianz von  $\hat{a} + \hat{b}x_0$ :

$$\begin{aligned} V[\hat{a} + \hat{b}x_0] &= V[\hat{a}] + x_0^2 V[\hat{b}] + 2x_0 \text{Cov}[\hat{a}, \hat{b}] \\ &= s^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2} \right) \end{aligned}$$

$\frac{1}{n}$  kommt von der Stichprobengröße,  $\frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}$  hängt von der Entfernung von  $x_0$  von  $\bar{x}$  ab.

Ein  $(100 - \alpha)\%$  KI für  $\hat{a} + \hat{b}x_0 = E[Y|X = x_0]$ :

$$(\hat{a} + \hat{b}x_0) \pm t_{(n-2), 1-\frac{\alpha}{2}} s \sqrt{\left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2} \right)}$$

Ein  $(100 - \alpha)\%$  Vorhersage Intervall für  $Y|X = x_0$  ist:

$$(\hat{a} + \hat{b}x_0) \pm t_{(n-2), 1-\frac{\alpha}{2}} s \sqrt{\left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2} \right)}$$

## 6.4 Residuen

$$\begin{aligned}e_i &= y_i - (\hat{a} + \hat{b}x_i) \\ &= y_i - \hat{y}_i\end{aligned}$$

$\{e_i\}$  sind nicht eine Stichprobe aus  $N(0, \sigma^2)$ , weil wir  $a$  und  $b$  nur geschätzt haben. Wenn wir  $a$  und  $b$  genau wußten, wären  $e_i = y_i - (a + bx_i)$  eine Stichprobe aus  $N(0, \sigma^2)$ .

Die Residuen werden benutzt, um die Güte des Modells eingehender zu überprüfen. Die folgenden Graphiken sind hilfreich:

- (1)  $e_i$  v.  $\hat{y}_i$     Residuen v. Vorhergesagte
- (2)  $e_i$  v.  $i$     Residuen v. Reihenfolge
- (3)  $e_i$  v.  $w_i$     Residuen v. eine andere erklärende Variable
- (4) Histogramm/Boxplot/Dotplot von  $e_i$

## 6.5 Multiple Lineare Regression

Um die Wirkungen von mehreren erklärenden Variablen zu untersuchen:

$$Y = b_0 + \sum_{j=1}^p b_j X_j + \epsilon$$

Das Modell wird als Matrixgleichung geschrieben:

$$Y = Xb + \epsilon$$

$n$  ist die Anzahl Beobachtungen

$p$  ist die Anzahl erklärender Variablen

$Y$  ( $n \times 1$ ) ist die abhängige Variable

$\{X_j\}$  ( $n \times (p + 1)$ ) sind die erklärenden Variablen

$b$  ( $(p + 1) \times 1$ ) ist der Koeffizientenvektor

$\epsilon$  ( $n \times 1$ )  $\sim$  u.i.v.  $N(0, \sigma^2)$  ist der Fehlerterm

### 6.5.1 KQ Schätzung von $b$

$$\min C = \sum \epsilon_i^2 = \min \epsilon' \epsilon = \min (Y - Xb)'(Y - Xb)$$

$$\frac{dC}{db} = 0 \Rightarrow -X'Y + X'Xb = 0$$

$$\Rightarrow \hat{b} = (X'X)^{-1}X'Y$$

$$V[\hat{b}] = V[(X'X)^{-1}X'Y]$$

$$= (X'X)^{-1}X'((X'X)^{-1}X')'V[Y]$$

$$= (X'X)^{-1}\sigma^2$$

weil  $V[Y] = V[\epsilon] = \sigma^2 I$

Als Beispiel kann man die einfache lineare Regression in dieser allgemeinen Form herleiten. N.B. für  $p = 1$  gilt

$$(X'X)^{-1} = \frac{1}{n(\sum x_i^2 - n\bar{x}^2)} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

## 6.5.2 ML Schätzer = KQ Schätzer in Regression

Mit dem Modell

$$Y = Xb + \epsilon$$

wo  $\epsilon \sim N(0, \sigma^2 I)$ , nehmen wir an, dass

$$Y_i \sim N\left(b_0 + \sum_{j=1}^p b_j X_{ij}, \sigma^2\right) \quad \text{u.i.v.}$$

$$L(y, x; b, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} \left(y_i - b_0 - \sum_{j=1}^p b_j x_{ij}\right)^2}$$

$$l = \log L$$

$$= -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(y_i - b_0 - \sum_{j=1}^p b_j x_{ij}\right)^2$$

Für  $\sigma$  bekannt wird das Maximum von  $l$  beim Minimum von  $\epsilon' \epsilon$  erreicht.

### 6.5.3 Gütekriterien

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

Je mehr Variablen einbezogen werden, desto größer wird  $R^2$ . Ein Vorschlag ist das adjustierte  $R^2$ :

$$R^2 = 1 - \frac{RSS}{TSS}$$

wo  $RSS$  = Rest-Variabilität

$$R_{adj}^2 = 1 - \frac{RSS/(n - p - 1)}{TSS/(n - 1)}$$

$RSS/(n - p - 1) = s^2$  schätzt  $\sigma^2$

$TSS/(n - 1)$  schätzt  $\sigma_Y^2$

$$R_{adj}^2 = 1 - \frac{(n - 1)}{(n - p - 1)}(1 - R^2)$$

Beim kleinen  $R^2$  kann  $R_{adj}^2$  sogar  $< 0$  sein, aber man sollte Modelle mit solchen Werten sowieso nicht in Betracht ziehen. Bei größeren Datensätzen ist  $R_{adj}^2 \approx R^2$ .