Vorzeichentest, Binomialtest

Test für den Median
Modell: $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$.
F besitze den eindeutigen Median $m = F^{-1}(0.5)$.
Testproblem
Einseitig

$$H_0 : m \geq m_0 \qquad \text{versus} \qquad H_1 : m < m_0$$

bzw.

$$H_0 : m \leq m_0 \qquad \text{versus} \qquad H_1 : m > m_0$$

Rückführung auf Binomialtest
Zähle Anzahl der Beobachtungen, die größer als m_0 sind $\rightarrow Y$

$$Y \sim \text{Bin}(n, p), \qquad p = P(Y_1 > m_0).$$

Für $m \overset{\leq}{\geqq} m_0$ gilt $p \overset{\leq}{\geqq} p_0 = 1/2$.

Exakter Binomialtest

Modell: Sei $Y \sim \text{Bin}(n, p)$.
Testproblem:

$$H_0 : p = p_0 \qquad \text{versus} \qquad H_1 : p \neq p_0.$$

Einseitig, z.B.

$$H_0 : p \leq p_0 \qquad \text{versus} \qquad H_1 : p > p_0.$$

Exakter Binomialtest

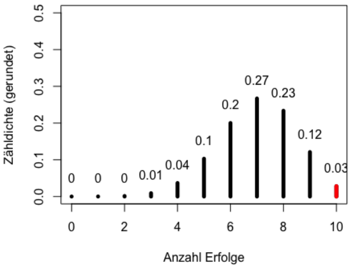
Lehne $H_0 : p \leq p_0$ zugunsten von $H_1 : p > p_0$, wenn $Y > c_{\text{krit}}$ ist. Hierbei ist c_{krit} die kleinste ganze Zahl, so dass

$$\sum_{k=c_{\text{krit}}+1}^{n} \binom{n}{k} p_0^k (1 - p_0)^{n-k} \leq \alpha.$$

Exakter Binomialtest

Beispiel: $Y \sim \text{Bin}(10, p)$, $H_0 : p \leq 0.7$ versus $H_1 : p > 0.7$

Nominales Signifikanzniveau $\alpha = 0.05$



Kritischer Wert: $c_{\text{krit}} = 9$. H_0 wird abgelehnt, falls $Y > 9$.

Reales Signifikanzfikanzniveau: $P_{0.8}(Y > 9) = 0.03$.

$\Rightarrow$ Konservativer Test: Das Signifikanzniveau wird nicht ausgeschöpft.

Asymptotischer Binomialtest

Asymptotischer Test

1. Lehne $H_0 : p \leq p_0$ auf dem Niveau α zugunsten von $H_1 : p > p_0$ ab, wenn
$$T = \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}} > z_{1-\alpha}.$$
Äquivalent zu: $Y > np_0 + z_{1-\alpha}\sqrt{np_0(1 - p_0)}$.
2. Lehne $H_0 : p \geq p_0$ zugunsten $H_1 : p < p_0$ ab, wenn $T < -z_{1-\alpha}$.
3. Lehne $H_0 : p = p_0$ zugunsten von $H_1 : p \neq p_0$ ab, wenn $|T| > z_{1-\alpha/2}$.

$z_{1-\alpha}$: $(1 - \alpha)$-Quantil der $N(0, 1)$-Verteilung.

Spezialfall $p_0 = 1/2$: Teststatistik vereinfacht sich zu
$$T = \frac{Y - n/2}{\sqrt{n/4}} = 2\frac{Y - n/2}{\sqrt{n}}.$$

2-Stichproben-Binomialtest

Modell: $Y_1 \sim \text{Bin}(n_1, p_1)$, $Y_2 \sim \text{Bin}(n_2, p_2)$ stochastisch unabhängig.
Testproblem:

$$H_0 : p_1 = p_2 \qquad \text{versus} \qquad H_1 : p_1 \neq p_2.$$

Schätzer: (Erfolgsraten in den Stichproben)

$$\widehat{p}_1 = Y_1/n_1, \qquad \widehat{p}_2 = Y_2/n_2$$

Teststatistik:

$$T = \frac{\widehat{p}_2 - \widehat{p}_1}{\sqrt{\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2} + \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}}}$$

Lehne H_0 ab, falls $|T| > z_{1-\alpha/2}$.

Test auf Korrelation

Modell: $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} N(\mu, \mathbf{\Sigma})$ bivariat normalverteilt mit

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

und Kovarianzmatrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \gamma \\ \gamma & \sigma_Y^2 \end{pmatrix}$$

Kovarianz

$$\gamma = \rho_{XY} = \mathrm{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y)$$

Korrelation

$$\rho = \mathrm{Cor}(X, Y) = \frac{\gamma_{XY}}{\sigma_X \sigma_Y}$$

Test auf Korrelation

Testproblem:

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_1 : \rho \neq 0$$

(Empirischer) Korrelationskoeffizient:

$$\widehat{\rho} = r_{XY} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}},$$

Teststatistik:

$$T = \frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}}$$

Unter H_0 gilt:

$$T \sim t(n-2)$$

1. Lehne $H_0 : \rho = 0$ ab, falls $|T| > t(n-2)_{1-\alpha/2}$.
2. Lehne $H_0 : \rho \geq 0$ ab, falls $T < -t(n-2)_{1-\alpha}$.
3. Lehne $H_0 : \rho \leq 0$ ab, falls $T > t(n-2)_{1-\alpha}$.

Beispiel: Die erwarteten Umsatzerlöse (Y) des Online-Shop SuperBuy4U eines Startups seien linear abhängig von der kumulierten Betrachtungszeit X der Werbespots auf Youtube (X = Viewer Retention · Clicks). Basierend auf den Daten der letzten $n = 25$ Wochen erhielt man folgende Statisiken:

$$\sum_{i=1}^{n} x_i y_i = 2603.316, \quad \sum_{i=1}^{n} x_i^2 = 16256.15, \quad \sum_{i=1}^{n} y_i^2 = 420.4859$$

sowie $\overline{x} = 24.96432$ und $\overline{y} = 4.049282$.

1. Formulieren Sie das zugehörige lineare Regressionsmodell unter Normalverteilungsannahme.
2. Berechnen Sie die Ausgleichsgerade.

Regression

Regressionsproblem

Gegeben: Punkte $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ **Punktewolke**.
Modell: Daten streuen um Gerade

$$f(x) = a + b \cdot x, \qquad x \in \mathbb{R}.$$

- Finde diejenige Gerade, die den Datensatz optimal approximiert.
- y_i : Zielwert (target, response, output)
- x_i : Regressor (erklärende Variable, input)

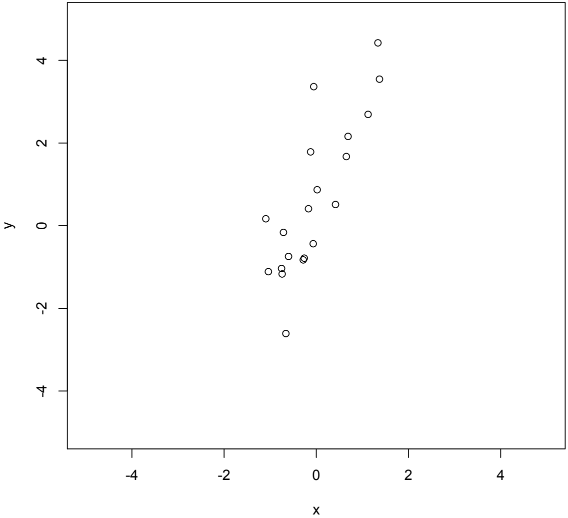
Regression

Idee

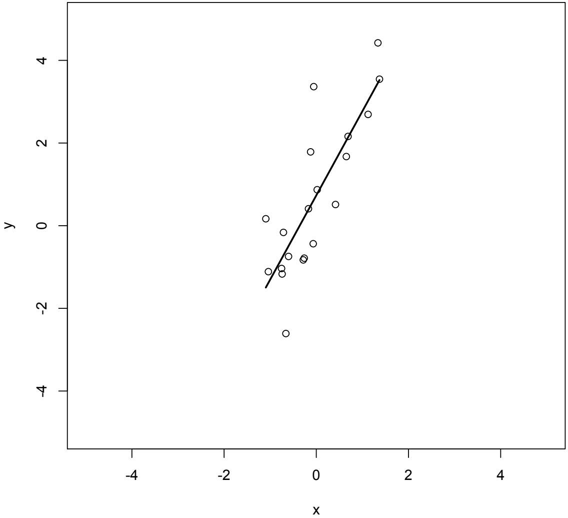
n Abstände der Punkte (x_i, y_i) zur Gerade in y-Richtung:

$$|y_i - (a + b \cdot x_i)|, \qquad i = 1, \ldots, n.$$

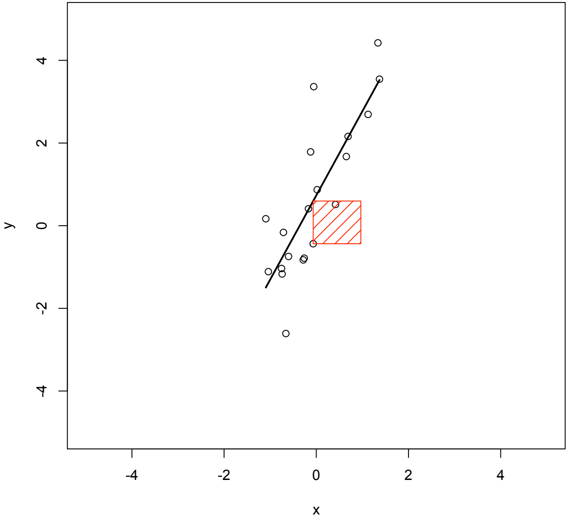
Alle n quadrierten Abstände $(y_i - (a + b \cdot x_i))^2$ sollen gleichmäßig klein sein.


y
x
4
2
0
-2
-4
-4
-2
0
2
4

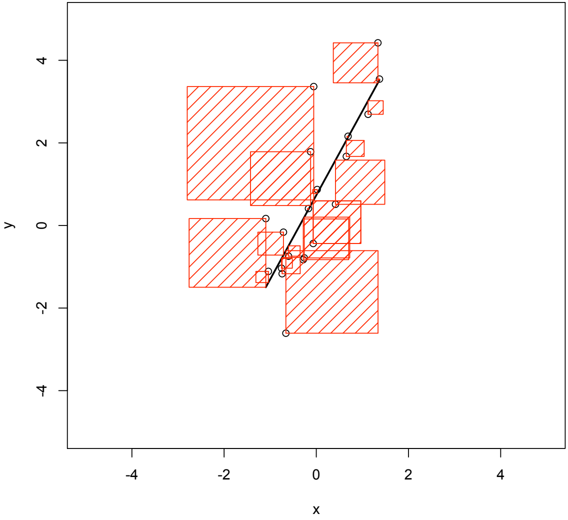
Regression. KQ-Schätzung



Regression. KQ-Schätzung



Regression. KQ-Schätzung



Methode der kleinsten Quadrate

KQ-Methode

Minimiere

$$Q(a,b) = \sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2, \qquad (a,b) \in \mathbb{R}^2.$$

Lösungen $(\widehat{a}, \widehat{b})$:

$$\widehat{b} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

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

Methode der kleinsten Quadrate

$Q(a,b)$ stetig partiell differenzierbar mit $\lim_{|a|\to\infty} Q(a,b) = \lim_{|b|\to\infty} Q(a,b) = \infty$.

$$\frac{\partial Q(a,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i), \qquad \frac{\partial Q(a,b)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i\,.$$

Ist $(\widehat{a}, \widehat{b})$ eine Minimalstelle, dann gilt nach dem notwendigen Kriterium 1. Ordnung:

$$0 = -\sum_{i=1}^{n} y_i + n\widehat{a} + \widehat{b}\sum_{i=1}^{n} x_i\,,$$

$$0 = -\sum_{i=1}^{n} y_i x_i + \widehat{a}\sum_{i=1}^{n} x_i + \widehat{b}\sum_{i=1}^{n} x_i^2\,.$$

Dies ist ein lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten. Division der ersten Gleichung durch $n > 1$ führt auf:

$$0 = -\overline{y} + \widehat{a} + \widehat{b}\cdot\overline{x}.$$

Löst man diese Gleichung nach $\widehat{a}$ auf, so erhält man $\widehat{a} = \overline{y} - \widehat{b}\,\overline{x}$. Einsetzen in die zweite Gleichung und anschließendes Auflösen nach $\widehat{b}$ ergibt

$$\widehat{b} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2}\,.$$

Berechnet man die Hesse-Matrix, so stellt sich $(\widehat{a}, \widehat{b})$ als Minimalstelle heraus (vgl.

Lineare Regression

Schätzer:

$$\begin{aligned} \widehat{b} &= \frac{\sum_{i=1}^{n} y_i x_i - n \cdot \overline{yx}}{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2} = \frac{s_{xy}}{s_x^2}, \\ \widehat{a} &= \overline{y} - \widehat{b} \cdot x. \end{aligned}$$

Geschätzte Regressionsgleichung (Ausgleichsgerade):

$$\widehat{f}(x) = \widehat{a} + \widehat{b} \cdot x, \qquad \text{für} \quad x \in [x_{\min}, x_{\max}].$$

Vorhersage- oder Prognosewerte:

$$\widehat{y}_i = \widehat{a} + \widehat{b} \cdot x_i, \qquad i = 1, \ldots, n.$$

Geschätzte Residuen:

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i, \qquad i = 1, \ldots, n,$$

Handrechnungen...

Es gilt:

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n \cdot (\overline{x} \cdot \overline{y}),$$

$$\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n \cdot (\overline{x})^2$$

Daher:

$$\widehat{b} = \frac{\sum_i x_i y_i - n \cdot (\overline{x} \cdot \overline{y})}{\sum_i x_i^2 - n \cdot (\overline{x})^2}$$

Regression

Ausgleichsgerade

Ausgleichs- oder Regressionsgerade:

$$\widehat{f}(x) = \widehat{a} + \widehat{b} \cdot x, \qquad x \in [x_{\min}, x_{\max}]$$

$[x_{\min}, x_{\max}]$: Stützbereich der Regression.

Anpassungsgüte

Streuungszerlegung:

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y}) + \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = SSR + SSE$$

Bestimmtheitsmaß:

$$R^2 = \frac{SSR}{SST} = r_{XY}^2$$

Residuenplot: Plotte Index $i = 1, \ldots, n$ gegen $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$.

Regression: Zahlenbeispiel

Gegeben seien die folgenden Daten:

x	1	2	3	4	5	6	7
y	1.7	2.6	2.0	2.7	3.2	3.6	4.6

Hieraus berechnet man:

$$\sum_{i=1}^{7} x_i = 28, \qquad \sum_{i=1}^{7} x_i^2 = 140, \qquad \overline{x} = 4\,,$$

$$\sum_{i=1}^{7} y_i = 20.4, \qquad \sum_{i=1}^{7} y_i^2 = 65.3, \qquad \overline{y} = 2.91429\,,$$

sowie $\sum_{i=1}^{7} y_i x_i = 93.5$. Die geschätzten Regressionskoeffizienten lauten somit:

$$\begin{aligned} \widehat{b} &= \frac{\sum_{i=1}^{7} y_i x_i - n \cdot \overline{x}\,\overline{y}}{\sum_{i=1}^{7} x_i^2 - n \cdot \overline{x}^2} \\ &\approx \frac{93.5 - 7 \cdot 4 \cdot 2.91}{140 - 7 \cdot (4)^2} = \frac{12.02}{28} \approx \underline{0.4293}. \end{aligned}$$

$$\widehat{a} = \overline{y} - \widehat{b} \cdot \overline{x} = 2.91 - 0.4293 \cdot 4 = \underline{1.1928}.$$

Die Ausgleichsgerade ist somit gegeben durch:

$$\widehat{f}(x) = 1.1928 + 0.4293 \cdot x, \qquad x \in [1, 7].$$

Lineare Regression

Modell: Unabhängige identisch verteilte Zufallsvektoren

$$(Y, x), (Y_1, x_1), \ldots, (Y_n, x_n)$$

mit

- Y_i: gemessener Wert (zufallsbehaftet) der Zielgröße
- x_i: zugehöriger Wert (fest) der erklärenden Variable

Stochastisches Regressionsmodell:

$$Y_i = a + bx_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

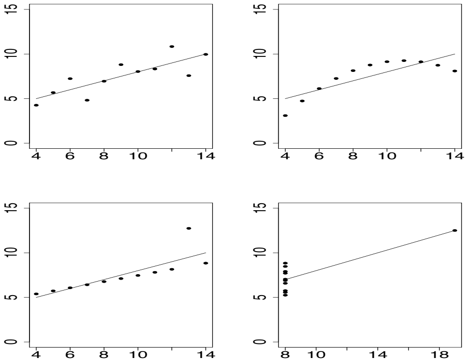
mit Störtermen (Messfehlern, Rauschen (noise)) $\epsilon_1, \ldots, \epsilon_n$,

$$E(\epsilon_i) = 0, \qquad \mathrm{Var}(\epsilon_i) = \sigma^2 \in (0, \infty), \qquad i = 1, \ldots, n.$$

Standardannahmen (klassisch)

1. $\epsilon_1, \ldots, \epsilon_n$ i.i.d $N(0, \sigma^2)$-verteilt.
2. $x_1, \ldots, x_n$ vorgegeben (fest, fixed design).
3. a, b: unbekannte Parameter, **Regressionskoeffizienten**.

Beispiel



4 Datensätze mit identischen Ausgleichsgeraden!

Schätzer:

$$
\begin{aligned}
\widehat{b} &= \frac{\sum_{i=1}^{n} Y_i x_i - n \cdot \overline{Y}\overline{x}}{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2} = \frac{s_{xy}}{s_x^2}, \\
\widehat{a} &= \overline{Y} - \widehat{b} \cdot x.
\end{aligned}
$$

Die Schätzer sind zufällig (Zufallsvariablen/Statistiken).
Geschätzte Regressionsgleichung (Ausgleichsgerade):

$$\widehat{f}(x) = \widehat{a} + \widehat{b} \cdot x, \qquad \text{für} \quad x \in [x_{\min}, x_{\max}].$$

Vorhersage- oder Prognosewerte:

$$\widehat{Y}_i = \widehat{a} + \widehat{b} \cdot x_i, \qquad i = 1, \ldots, n.$$

Geschätzte Residuen:

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i, \qquad i = 1, \ldots, n,$$

Statistische Eigenschaften

Eigenschaften

Die Schätzer $\widehat{a}$ und $\widehat{b}$ sind erwartungstreu und konsistent für die Regressionskoeffizienten a bzw. b. Ihre Varianzen können durch

$$\widehat{\sigma}_b^2 = \frac{\widehat{\sigma}^2}{n \cdot s_x^2} \qquad \text{sowie} \qquad \widehat{\sigma}_a^2 = \frac{\sum_{i=1}^n x_i^2}{n \cdot s_x^2} \cdot \widehat{\sigma}^2$$

geschätzt werden. Hierbei ist

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

eine erwartungstreue und konsistente Schätzung des Modellfehlers σ^2.

Statistische Eigenschaften

Verteilung

Sind $\epsilon_1, \dots, \epsilon_n$ i.i.d. $N(0, \sigma^2)$-verteilt, dann gilt:

$$T_b = \frac{\widehat{b} - b}{\widehat{\sigma}_b} \sim t(n-2), \quad T_a = \frac{\widehat{a} - a}{\widehat{\sigma}_a} \sim t(n-2),$$

und

$$Q = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Konfidenzintervall

$(1-\alpha)$-Konfidenzintervall für b:

$$\widehat{b} \pm t(n-2)_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$(1-\alpha)$-Konfidenzintervall für σ^2:

$$\left[\frac{(n-2)\widehat{\sigma}^2}{\chi^2(n-2)_{1-\alpha/2}}, \frac{(n-2)\widehat{\sigma}^2}{\chi^2(n-2)_{\alpha/2}} \right]$$

Test des Steigungsmaßes b und Intercepts a:

Teststatistiken mit H_0-Wert eingesetzt: $T_a = \frac{\widehat{a}-a_0}{\widehat{\sigma}_a}$, $T_b = \frac{\widehat{b}-b_0}{\widehat{\sigma}_b}$.

1. $H_0 : b = b_0$ gegen $H_1 : b \neq b_0$.
   H_0 ablehnen, wenn $|T_b| > t(n-2)_{1-\alpha/2}$.
2. $H_0 : b \leq b_0$ gegen $H_1 : b > b_0$.
   H_0 ablehnen, falls $T_b > t(n-2)_{1-\alpha}$.
3. $H_0 : b \geq b_0$ gegen $H_1 : b < b_0$.
   H_0 ablehnen, falls $T_b < -t(n-2)_{1-\alpha}$.

Die entsprechenden Tests für den Parameter a erhält man durch Ersetzen von b durch a in den Hypothesen und Ersetzen von T_b durch T_a.

Test des Modellfehlers σ^2:

1. $H_0 : \sigma^2 = \sigma_0^2$ gegen $H_1 : \sigma^2 \neq \sigma_0^2$.
   H_0 ablehnen, wenn $Q < \chi^2(n-2)_{\alpha/2}$ oder $Q > \chi^2(n-2)_{1-\alpha/2}$.
2. $H_0 : \sigma^2 \leq \sigma_0^2$ gegen $H_1 : \sigma^2 > \sigma_0^2$.
   H_0 ablehnen, falls $Q > \chi^2(n-2)_{1-\alpha}$.
3. $H_0 : \sigma^2 \geq \sigma_0^2$ gegen $H_1 : \sigma^2 < \sigma_0^2$.
   H_0 ablehnen, falls $Q < \chi^2(n-2)_{\alpha}$.