Vorlesung vom 15.11.2012:

- Lineare Regression = Ausgleichsgerade
- Einstieg in die beurteilende Statistik $\Rightarrow$ naive Mengenlehre!

---

1) Zur Berechnungsformel für die Kovarianz:

Im Skript taucht als alternative Formel zur Berechnung der Kovarianz auf:

$$\operatorname{cov}(\underline{X},\underline{Y}) = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\cdot\left(\sum_{i=1}^{n}x_i y_i\right)-\bar{x}\cdot\bar{y}$$

Achtung: Für $\underline{X}=\underline{Y}$ erhält man dann die Varianzformel

$$\operatorname{var}(\underline{X}) = \operatorname{cov}(\underline{X},\underline{X}) = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right)-\bar{x}^2$$

Dabei sind $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ und $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ die arithmetischen Mittelwerte für $\underline{X}$ und $\underline{Y}$.

Beispielrechnung mit $n=4$ Datensätzen:

Seien $z_1=(x_1,y_1)=(1.0,3.0)$, $z_2=(2.0,2.0)$, $z_3=(2.0,3.5)$, $z_4=(4.0,2.5)$

Tabelle:

i	1	2	3	4	Σ	$\frac{1}{n}\cdot\Sigma$	
x_i	1.0	2.0	2.0	4.0	9.0	2.25	$=\bar{x}$
y_i	3.0	2.0	3.5	2.5	11.0	2.75	$=\bar{y}$
x_i^2	1.0	4.0	4.0	16.0	25.0	6.25	
y_i^2	9.0	4.0	12.25	6.25	31.5	7.875	
$x_i y_i$	3.0	4.0	7.0	10.0	24.0	6.0	

Also: $\bar{x}=2.25$, $\bar{y}=2.75$,

$$\operatorname{var}(\underline{X}) = \frac{1}{4}\sum_{i=1}^{4}x_i^2-\bar{x}^2 = 6.25-2.25^2 = 1.1875$$

$$\operatorname{var}(\underline{Y}) = \frac{1}{4}\sum_{i=1}^{4}y_i^2-\bar{y}^2 = 7.875-2.75^2 = 0.3125$$

$$\operatorname{cov}(\underline{X},\underline{Y}) = \frac{1}{4}\cdot\sum_{i=1}^{4}x_i y_i-\bar{x}\bar{y} = 6.0-2.25\cdot 2.75 = -0.1875 < 0$$

Für die empirische Korrelation folgt dann:

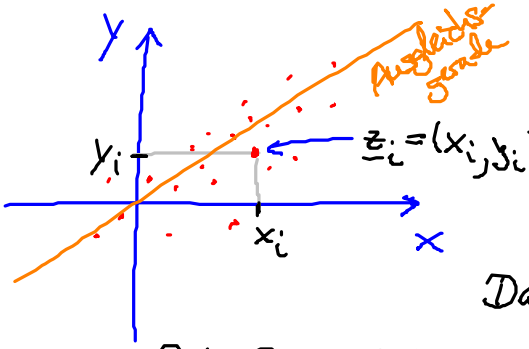
$$r(\underline{x}^*, \underline{y}^*) = \frac{\mathrm{cov}(\underline{X}, \underline{Y})}{\sqrt{\mathrm{var}(\underline{X})} \cdot \sqrt{\mathrm{var}(\underline{Y})}} = -\frac{0.1875}{\sqrt{1.1875} \cdot \sqrt{0.3125}} \approx -0{,}9733 \approx -1.0$$

Also haben wir es hier mit einer sehr starken negativen Korrelation dicht an -1 zu tun.

---

2) Zur linearen Regression = Ausgleichsgerade !!

Gegeben: eine Punktwolke $\underline{z} = (\underline{z}_1, \dots, \underline{z}_n)$ mit Punkten $\underline{z}_i = (x_i, y_i)$



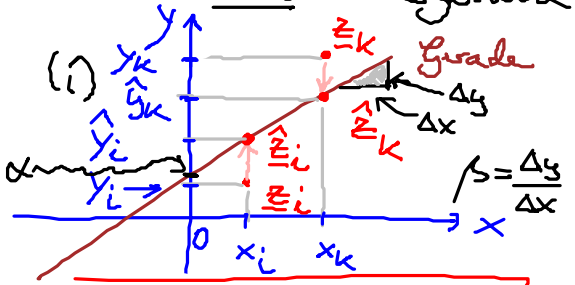
Frage: Lässt sich – zumindest bei hoher Korrelation zwischen den Merkmalen X und Y – eine relativ „gute" Ausgleichsgerade legen?

Das ist der wesentliche Gedanke der „linearen Regression":

Betrachte anstelle der Originalpunkte $\underline{z}_i = (x_i, y_i)$ entsprechende Projektionspunkte $\hat{\underline{z}}_i = (\hat{x}_i, \hat{y}_i)$ auf der ausgewählten Regressions- bzw. Ausgleichsgeraden !!

Es gibt im Wesentlichen zwei hervorgehobene Ausgleichsgeraden mit zugehöriger Projektion:

(i)



$$\hat{\underline{z}}_i = (\hat{x}_i, \hat{y}_i) \equiv (x_i, \hat{y}_i)$$

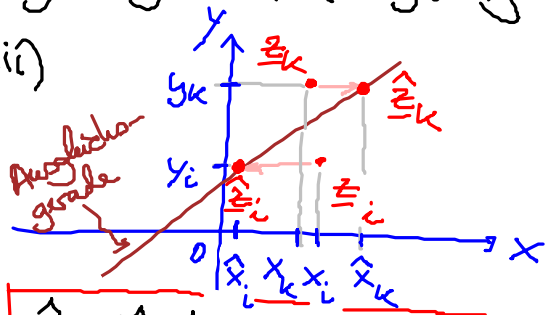
$\hat{x}_i = x_i$

Projektion in y-Richtung (↑↓) auf die Gerade !!

Geradengleichung: $\hat{y} = \alpha + \beta x$

Also: $\hat{y}_i = \alpha + \beta x_i \quad (i = 1, \dots, n)$

(ii)



$$\hat{\underline{z}}_i = (\hat{x}_i, \hat{y}_i) \equiv (\hat{x}_i, y_i)$$

$\hat{y}_i = y_i$

Projektion in x-Richtung (↔) auf die Gerade !!

Geradengleichung: $\hat{x} = \gamma + \delta y$

Also: $\hat{x}_i = \gamma + \delta y_i \quad (i = 1, \dots, n)$

α: y-Achsenabschnitt im $x=0$

$\beta = \frac{\Delta y}{\Delta x} = \tan\varphi$: Steigung der Geraden.

heißt Ausgleichsgerade von y nach x, d.h.: $\hat{y}$ hängt von x ab.

heißt Ausgleichsgerade von x nach y, d.h.: $\hat{x}$ hängt von y ab.

Die Ausgleichsgeradenparameter α, β bzw. γ, δ werden so ermittelt, dass die Gaußsche quadratische Abweichung

$$SSE = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2$$

minimal wird.

1. Fall: $\hat{x}_i = x_i \Rightarrow SSE = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 = \min.$ ($\hat{y}_i$)

2. Fall: $\hat{y}_i = y_i \Rightarrow SSE = \sum_{i=1}^{n} [x_i - (\gamma + \delta y_i)]^2 = \min.$ ($\hat{x}_i$)

Führt auf die Parameter:

$$\beta = \frac{cov(x,y)}{var(x)}, \quad \alpha = \bar{y} - \beta\bar{x}$$

$$\delta = \frac{var(y)}{cov(x,y)}, \quad \gamma = \bar{y} - \delta\bar{x}$$

In beiden Fällen liegt der Datenschwerpunkt $S = (\bar{x}, \bar{y})$ mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

arithmetische Mittelwerte

immer auf der Ausgleichsgeraden, da ja $\alpha, \beta, \gamma, \delta$ so gebildet wurden, dass gilt:

$$\alpha + \beta\bar{x} = \bar{y} \quad \text{bzw.} \quad \gamma + \delta\bar{x} = \bar{y}$$ ✓

---

3) Zum Einstieg in die Wahrscheinlichkeitsrechnung:

Beachten Sie das online gestellte Skript zur Mengenlehre!!

ENDE der heutigen Vorlesung!