Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Korrelationskoeffizient

Für den (empirischen) Korrelationskoeffizienten r erhält man mit

$$\hat{b} = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} \quad \text{und} \quad r = \frac{\tilde{s}_{xy}}{\tilde{s}_x \, \tilde{s}_y}$$

die Beziehung

$$\hat{b} = \frac{\tilde{s}_y}{\tilde{s}_x} \, r$$

z-transformierte Werte

Wird die Regression mit z-transformierten Stichprobenwerten x_i, y_i $(i = 1, \ldots, n)$ durchgeführt, gilt also

$$\bar{x} = \bar{y} = 0 \quad \text{und} \quad \tilde{s}_x = \tilde{s}_y = 1,$$

so verschwindet der Ordinatenabschnitt und der Regressionskoeffizient ist identisch mit dem Korrelationskoeffizienten, d. h.

$$\hat{a} = 0 \quad \text{und} \quad \hat{b} = r$$

Determinationskoeffizient

Wir betrachten die Varianz $\tilde{s}_e^2$ der Residuen e_i

$$\tilde{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Im Folgenden bestimmen wir diese Varianz unter Verwendung der Beziehung

$$\hat{y}_i = \bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\,(x_i - \bar{x})$$

Determinationskoeffizient

Es gilt

$$\begin{aligned}\tilde{s}_e^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i - (\bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\,(x_i - \bar{x}))\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \bar{y}) - \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\,(x_i - \bar{x})\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\,\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\,\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \left(\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\right)^2\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\end{aligned}$$

Mit den Definitionen von $\tilde{s}_y^2$, $\tilde{s}_x^2$ und $\tilde{s}_{xy}$ erhält man weiter

$$\tilde{s}_e^2 = \tilde{s}_y^2 - 2\,\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2} + \left(\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\right)^2 \cdot \tilde{s}_x^2 = \tilde{s}_y^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$$

Determinationskoeffizient

Nun nutzt man die Beziehung

$$b = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} = \frac{\tilde{s}_y}{\tilde{s}_x} \cdot r$$

Eingesetzt ergibt sich

$$\begin{aligned}
\tilde{s}_e^2 &= \tilde{s}_y^2 - \frac{\tilde{s}_y}{\tilde{s}_x}\, \tilde{s}_{xy}\, r \\
&= \tilde{s}_y^2 - \tilde{s}_y^2 \frac{\tilde{s}_{xy}}{\tilde{s}_x\, \tilde{s}_y}\, r \\
&= \tilde{s}_y^2 - \tilde{s}_y^2\, r^2 \\
&= (1 - r^2)\, \tilde{s}_y^2
\end{aligned}$$

Determinationskoeffizient

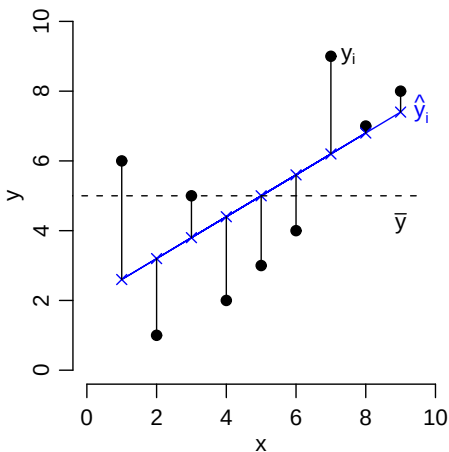
Für die (empirische) Varianz $\tilde{s}_e^2$ der Residuen ergibt sich also

$$\tilde{s}_e^2 = (1 - r^2)\, \tilde{s}_y^2$$

Der Faktor $(1 - r^2)$ kennzeichnet also den Anteil an der Varianz von Y, der durch die Regression auf X *nicht* aufgeklärt wird

Umgekehrt gibt die Größe r^2, die als *Determinationskoeffizient* (Bestimmtheitsmaß) bezeichnet wird, den Anteil der durch die Regression auf X aufgeklärten Varianz von Y an (diese Behauptung wird nachfolgend belegt)

Varianzzerlegung



$$\tilde{s}_y^2 = \tilde{s}_{\hat{y}}^2 + \tilde{s}_e^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Varianzzerlegung

Die Berechnung des Determinationskoeffizienten legt eine Zerlegung der Varianz des Kriteriums Y nahe, die im Folgenden näher untersucht wird

Wir betrachten die Varianz

$$\tilde{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

und erweitern diesen Ausdruck unter Einbeziehung der durch die Regression vorhergesagten Werte $\hat{y}_i$ zu

$$\tilde{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

Varianzzerlegung

Es gilt dann

$$
\begin{aligned}
\tilde{s}_y^2 &= \frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{y}_i)^2 + 2\,(y_i - \hat{y}_i)\,(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\,\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,(\hat{y}_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&= \tilde{s}_e^2 + 2\,\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,(\hat{y}_i - \bar{y}) + \tilde{s}_{\hat{y}}^2
\end{aligned}
$$

Varianzzerlegung

Dabei bezeichnet

$\tilde{s}_e^2$ die Varianz der Residuen

$\tilde{s}_{\hat{y}}^2$ die Varianz der durch die Regression vorhergesagten Werte

Für die behauptete Varianzzerlegung ist also zu zeigen

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,(\hat{y}_i - \bar{y}) = 0$$

Varianzzerlegung

Es ist

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i\,\hat{y}_i - y_i\,\bar{y} - \hat{y}_i^2 + \hat{y}_i\,\bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i - \bar{y}\,\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 + \bar{y}\,\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i - \bar{y}^2 - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 + \bar{y}\,\bar{\hat{y}}\end{aligned}$$

Wir zeigen zunächst die Gleichung $\bar{y} = \bar{\hat{y}}$

Varianzzerlegung

Aus der Regressionsgleichung $\hat{y}_i = \bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}(x_i - \bar{x})$ folgt durch Bildung des arithmetischen Mittels und dessen Nulleigenschaft

$$\begin{aligned}
\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i &= \frac{1}{n}\sum_{i=1}^{n}[\bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}(x_i - \bar{x})] \\
&= \frac{1}{n}\sum_{i=1}^{n}\bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \frac{1}{n} \cdot n \cdot \bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} \cdot 0 \\
&= \bar{y}
\end{aligned}$$

Varianzzerlegung

Mit der Gleichung $\bar{y} = \bar{\hat{y}}$ folgt dann

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,(\hat{y}_i - \bar{y}) &= \frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i - \bar{y}^2 - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 + \bar{y}\,\bar{\hat{y}} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i - \bar{y}^2 - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 + \bar{y}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2
\end{aligned}$$

Wir setzen nun wiederum die Regressionsgleichung $\hat{y}_i = \bar{y} + \dfrac{\tilde{s}_{xy}}{\tilde{s}_x^2}\,(x_i - \bar{x})$ ein

Varianzzerlegung

Nebenrechnung 1

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i &= \frac{1}{n}\sum_{i=1}^{n} y_i\,[\bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\,(x_i - \bar{x})] \\
&= \bar{y}\cdot\frac{1}{n}\sum_{i=1}^{n} y_i + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\cdot\left[\frac{1}{n}\sum_{i=1}^{n} y_i\,x_i - \bar{x}\cdot\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \bar{y}^2 + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\cdot\left[\frac{1}{n}\sum_{i=1}^{n} y_i\,x_i - \bar{x}\cdot\bar{y}\right] \\
&= \bar{y}^2 + \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}
\end{aligned}$$

Varianzzerlegung

Nebenrechnung 2

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 &= \frac{1}{n}\sum_{i=1}^{n}[\bar{y}+\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}(x_i-\bar{x})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[\bar{y}^2+2\cdot\bar{y}\cdot\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}(x_i-\bar{x})+\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^4}(x_i-\bar{x})^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\bar{y}^2+2\cdot\bar{y}\cdot\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})+\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^4}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n}\cdot n\cdot\bar{y}^2+2\cdot\bar{y}\cdot\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\cdot\frac{1}{n}\cdot 0+\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^4}\cdot\tilde{s}_x^2
\end{aligned}$$

Varianzzerlegung

Mit den beiden Nebenrechnungen folgt nun

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)\,(\hat{y}_i-\bar{y}) &= \frac{1}{n}\sum_{i=1}^{n} y_i\,\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 \\ &= \bar{y}^2 + \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2} - \bar{y}^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2} \\ &= 0\end{aligned}$$

Insgesamt haben wir daher gezeigt

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)\,(\hat{y}_i-\bar{y}) = 0$$

Varianzzerlegung

Als Endergebnis der obigen Berechnungen erhalten wir somit die Varianzzerlegung

$$\tilde{s}_y^2 = \tilde{s}_e^2 + \tilde{s}_{\hat{y}}^2$$

Die Varianz des Kriteriums Y lässt sich also additiv zerlegen in einen Anteil, der durch die Regression auf X aufgeklärt wird (erklärte Varianz), und die Varianz der Regressionsresiduen

Die jeweiligen Anteile an der Gesamtvarianz von Y werden durch den Determinationskoeffizienten bestimmt

$$\tilde{s}_e^2 = (1 - r^2) \cdot \tilde{s}_y^2$$
$$\tilde{s}_{\hat{y}}^2 = r^2 \cdot \tilde{s}_y^2$$

Beispiel

Mit der Regressionsgleichung

$$\hat{y}_i = 1.54 + 0.69 \cdot x_i$$

berechnet man

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$y_i - \hat{y}_i$	$\hat{y}_i - \bar{y}$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})^2$
1	2	3	2.92	-0.2	0.08	-0.28	0.04	0.01	0.08
2	1	2	2.23	-1.2	-0.23	-0.97	1.44	0.05	0.94
3	3	4	3.61	0.8	0.39	0.41	0.64	0.15	0.17
4	2	3	2.92	-0.2	0.08	-0.28	0.04	0.01	0.09
5	4	4	4.30	0.8	-0.30	1.10	0.64	0.09	1.21
$\sum$	12	16	$\approx$16.00	0.00	0.00	0.00	2.80	0.31	2.49
$\frac{1}{n}\sum$	2.4	3.2	$\approx$3.2	0.00	0.00	0.00	0.56	0.06	0.50

Beispiel

Aus der Tabelle entnimmt man

$$\tilde{s}_y^2 = \tilde{s}_e^2 + \tilde{s}_{\hat{y}}^2 = 0.06 + 0.50 = 0.56$$

und

$$\frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}} = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2} = \frac{0.50}{0.56} \approx 0.89$$

Für den Determinationskoeffizienten berechnet man

$$r^2 = 0.943^2 = 0.89$$

Berechnung in R mit Funktion `lm()`

Die Funktion `lm()` erlaubt die Schätzung linearer Modelle, wie sie durch lineare Regression, aber etwa auch durch die in Statistik D noch zu besprechende Varianzanalyse, gegeben sind

Wird die Variable `x` als Prädiktor für die Kriteriumsvariable `y` verwendet, wird das durch die Angabe des Arguments `y ~ x` ausgedrückt (siehe `?formula`)

Beispiel: Maximalpuls und Alter

```
dat <- data.frame(
  x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37),
  y = c(202,186,187,180,156,169,174,172,153,199,193,
        174,198,183,178)
)
lm(y ~ x, dat)
```

Berechnung in R mit Funktion `lm()`

Als Ausgabe erhält man dann

```
Call:
lm(formula = y ~ x, data = dat)

Coefficients:
(Intercept)            x
   210.0485      -0.7977
```

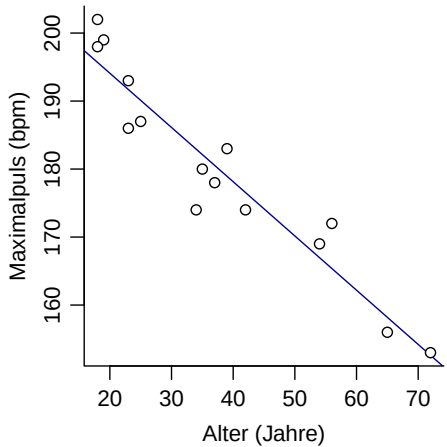
Mit der Eingabe

```
lm.result <- lm(y ~ x, dat)
```

werden sämtliche Ausgaben der Funktion `lm()` in dem Objekt `lm.result` gespeichert und stehen für weitere Abfragen zur Verfügung

Berechnung in R mit Funktion `lm()`

Streudiagramm mit geschätzter Regressionsgeraden




```
plot(y ~ x, dat,
     xlab = "Alter (Jahre)",
     ylab = "Maximalpuls (bpm)"
)
abline(lm.result)
```

Berechnung in R mit Funktion `lm()`

Ausgabe der Koeffizienten

```
coef(lm.result)
# (Intercept)           x
#     210.048    -0.7977
```

Ausgabe der vorhergesagten Werte

```
predict(lm.result)
#      1      2      3      4      5      6          15
# 195.69 191.70 190.11 182.13 158.20 166.97 ... 180.53
```

Ausgabe der Residuen

```
resid(lm.result)
#     1      2      3      4      5     6          15
# 6.311 -5.701 -3.105 -2.128 -2.196 2.029 ... -2.533
```

Berechnung in R mit Funktion `lm()`

Zusammenfassende Ausgabe des lm-Objekts

```
summary(lm.result)
# Call:
# lm(formula = y ~ x, data = dat)
#
# Residuals:
#    Min     1Q Median     3Q    Max
# -8.926 -2.538  0.388  3.187  6.624
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  210.048      2.867    73.3  < 2e-16
# x             -0.798      0.070   -11.4  3.8e-08
#
# Residual standard error: 4.58 on 13 degrees of freedom
# Multiple R-squared:  0.909,   Adjusted R-squared:  0.902
# F-statistic:  130 on 1 and 13 DF,  p-value: 3.85e-08
```