CAS Datenanalyse HS16 - DeskStat

Lineare Regression

Lineare Regression

- Das einfache lineare Regressionsmodell beschreibt eine abhängige Variable als lineare Funktion einer unabhängigen Variablen.

$$y = \beta_1 \cdot x + \beta_2 + \epsilon$$

Lineare Regression

- Das einfache lineare Regressionsmodell beschreibt eine abhängige Variable als lineare Funktion einer unabhängigen Variablen.

$$y = \beta_1 \cdot x + \beta_2 + \epsilon$$

- Die beiden Parameter β_1 und β_2 sind unbekannt und sollen durch b_1 und b_2 geschätzt werden.

Lineare Regression

- Das einfache lineare Regressionsmodell beschreibt eine abhängige Variable als lineare Funktion einer unabhängigen Variablen.

$$y = \beta_1 \cdot x + \beta_2 + \epsilon$$

- Die beiden Parameter β_1 und β_2 sind unbekannt und sollen durch b_1 und b_2 geschätzt werden.
- Zum Beispiel:

$$\texttt{eruptions} = \beta_1 \cdot \texttt{waiting} + \beta_2 + \epsilon$$

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.
- Der Erwartungswert der Residuen ist 0.

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.
- Der Erwartungswert der Residuen ist 0.
- Die Streuung der Residuen bleibt konstant.

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.
- Der Erwartungswert der Residuen ist 0.
- Die Streuung der Residuen bleibt konstant.
- Die Residuen sind normalverteilt.

Lineare Regression: Schätzen eines y-Wertes

Problem: Wir modelieren den Zusammenhang zwischen den Eruptionsdauern und den Wartezeiten aus `faithful` mit einem lineare Modell. Wie lange dauert die nächste Eruptions im Schnitt, wenn die Wartezeit 80 Minuten beträgt?

Lineare Regression: Schätzen eines y-Wertes

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
coeffs <- coefficients(eruption.lm)
coeffs

## (Intercept)     waiting
## -1.87401599  0.07562795

waiting <- 80
duration <- coeffs[1] + coeffs[2]*waiting
duration

## (Intercept)
##     4.17622
```

Lineare Regression: Schätzen eines y-Wertes

Erweiterte Antwort:

```
newdata <- data.frame(waiting=80)
predict(eruption.lm, newdata)

##        1
## 4.17622
```

Wir erwarten eine Eruptionsdauer von ungefähr 4 Minuten.

Lineare Regression: Bestimmtheitsmass r^2

- Das Bestimmtheitsmass r^2 gibt an, welcher Anteil der Streuung, die in den Daten `eruptions` steckt, durch das Model erklärt werden kann.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Lineare Regression: Bestimmtheitsmass r^2

- Das Bestimmtheitsmass r^2 gibt an, welcher Anteil der Streuung, die in den Daten `eruptions` steckt, durch das Model erklärt werden kann.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Bei der linearen Regression entspricht das Bestimmtheitsmass dem Quadrat des Korrelationskoeffizienten.

Lineare Regression: Bestimmtheitsmass r^2

Problem: Bestimmen Sie das Bestimmtheitsmass r^2 des linearen Modells zu `faithful`.

Lineare Regression: Bestimmtheitsmass r^2

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
summary(eruption.lm)$r.squared


## [1] 0.8114608
```

Lineare Regression: Signifikanztests

- Ist der Zusammenhang zwischen der abhängigen Variablen und der unabhängigen Variablen überhaupt signifikant oder kommt der Wert von b_1 bloss durch Zufall zustande?

Lineare Regression: Signifikanztests

- Ist der Zusammenhang zwischen der abhängigen Variablen und der unabhängigen Variablen überhaupt signifikant oder kommt der Wert von b_1 bloss durch Zufall zustande?
- Wir testen die Hypothesen

$$H_0 : \beta_1 = 0 \text{ und } H_1 : \beta_1 \neq 0$$

Lineare Regression: Signifikanztests

- Ist der Zusammenhang zwischen der abhängigen Variablen und der unabhängigen Variablen überhaupt signifikant oder kommt der Wert von b_1 bloss durch Zufall zustande?
- Wir testen die Hypothesen

$$H_0 : \beta_1 = 0 \text{ und } H_1 : \beta_1 \neq 0$$

- Ist $\beta_1 = 0$, dann ist auch der Korrelationskoeffizient $\rho = 0$. In diesem Fall besteht kein linearer Zusammenhang zwischen den beiden Grössen x und y.

Lineare Regression: Signifikanztest für β_1

Problem: Untersuchen Sie, ob zwischen den Grössen `eruptions` und `waiting` aus `faithful` ein signifikanter Zusammenhang besteht.

Lineare Regression: Signifikanztest für β_1

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
summary(eruption.lm)
```

```
##
## Call:
## lm(formula = eruptions ~ waiting, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Lineare Regression: Signifikanztest für β_1

Antwort: Der p-Wert ist nahezu gleich 0. Die Nullhypothese $\beta_1 = 0$ wird verworfen. Offenbar besteht ein signifikanter Zusammenhang zwischen der Wartezeit und den Eruptiondauer.

Lineare Regression: Konfidenzintervalle für y

- Gemäss dem errechneten Modell führt eine Wartezeit von $x = 80$ Minuten zu einer durchschnittlichen Eruptionsdauer von $y = 4$ Minuten.

Lineare Regression: Konfidenzintervalle für y

- Gemäss dem errechneten Modell führt eine Wartezeit von $x = 80$ Minuten zu einer durchschnittlichen Eruptionsdauer von $y = 4$ Minuten.
- Dieser Wert wurde aufgrund einer Stichprobe ermittelt. Der wahre Durchschnittswert wird von diesem Wert abweichen.

Lineare Regression: Konfidenzintervalle für y

- Gemäss dem errechneten Modell führt eine Wartezeit von $x = 80$ Minuten zu einer durchschnittlichen Eruptionsdauer von $y = 4$ Minuten.
- Dieser Wert wurde aufgrund einer Stichprobe ermittelt. Der wahre Durchschnittswert wird von diesem Wert abweichen.
- Wir schätzen den wahren Wert mit einem Konfidenzintvervall ab.

Lineare Regression: Konfidenzintervalle für y

Problem: Bestimmen Sie ein 95%-Konfidenzintervall für die durchschnittliche Eruptionsdauer bei einer Wartezeit von 80 Minuten.

Lineare Regression: Konfidenzintervalle für y

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
newdata <- data.frame(waiting=80)
predict(eruption.lm, newdata, interval="confidence")

##        fit       lwr      upr
## 1 4.17622 4.104848 4.247592
```

Die durchschnittliche Eruptionszeit beträgt bei einer Wartezeit von 80 Minuten zwischen 4.10 und 4.24 Minuten, bei einem Signifikanzniveau von 95%.

Lineare Regression: Prognoseintervalle für y

- Das Prognoseintervall liefert einen Wertebereich für die zu erwartenden Lage eines einzelnen vorhergesagten Wertes der abhängigen Variablen.

Lineare Regression: Prognoseintervalle für y

- Das Prognoseintervall liefert einen Wertebereich für die zu erwartenden Lage eines einzelnen vorhergesagten Wertes der abhängigen Variablen.
- Dieser Wertebereich ist wiederum abhängig von einem Konfidenzniveau α.

Lineare Regression: Prognoseintervalle für y

- Das Prognoseintervall liefert einen Wertebereich für die zu erwartenden Lage eines einzelnen vorhergesagten Wertes der abhängigen Variablen.
- Dieser Wertebereich ist wiederum abhängig von einem Konfidenzniveau α.
- Das Prognoseintervall ist wird einen grösseren Wertebereich als das Konfidenzintervall liefern.

Lineare Regression: Prognoseintervalle für y

Problem: Bestimmen Sie ein 95%-Prognoseintervall für die Eruptionsdauer bei einer Wartezeit von 80 Minuten.

Lineare Regression: Prognoseintervalle für y

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
newdata <- data.frame(waiting=80)
predict(eruption.lm, newdata, interval="predict")

##        fit      lwr      upr
## 1 4.17622 3.196089 5.156351
```

Die Eruptionszeit beträgt bei einer Wartezeit von 80 Minuten zwischen 3.20 und 5.16 Minuten, bei einem Signifikanzniveau von 95%.

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:
  - Der Erwartungswert der Residuen ist 0.

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:
  - Der Erwartungswert der Residuen ist 0.
  - Die Residuen haben eine gleichbleibende Streuung.

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:
  - Der Erwartungswert der Residuen ist 0.
  - Die Residuen haben eine gleichbleibende Streuung.
  - Die Residuen sind normalverteilt und unabhängig.

Lineare Regression: Residuen-Plot

Problem: Stellen Sie die Residuen des linearen Modells zwischen der Eruptionsdauer und der Wartezeit aus `faithful` grafisch dar.

Lineare Regression: Residuen-Plot

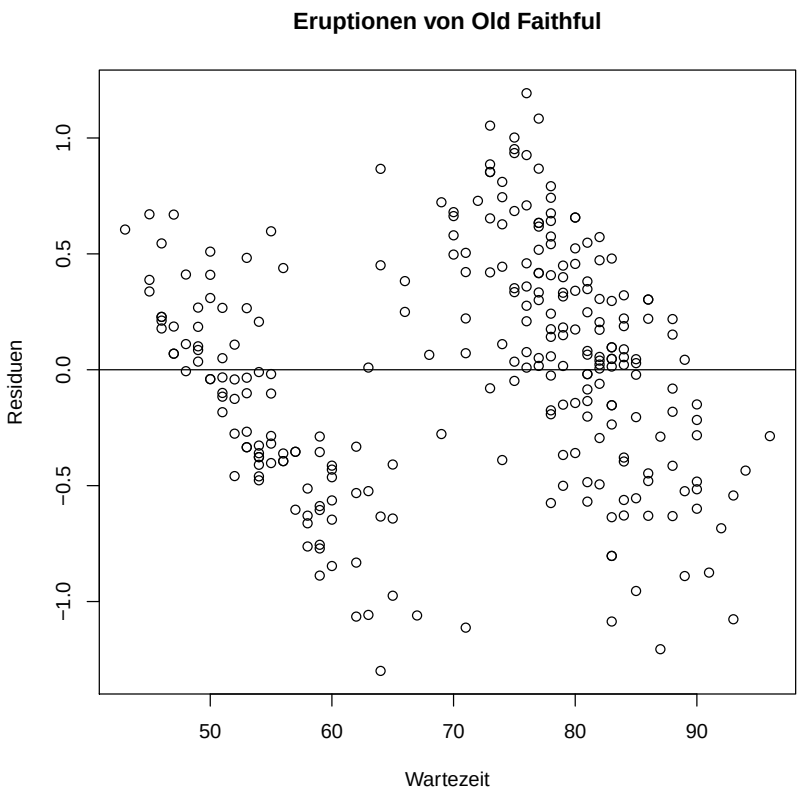
Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
eruption.res <- resid(eruption.lm)
```

Lineare Regression: Residuen-Plot

Antwort:

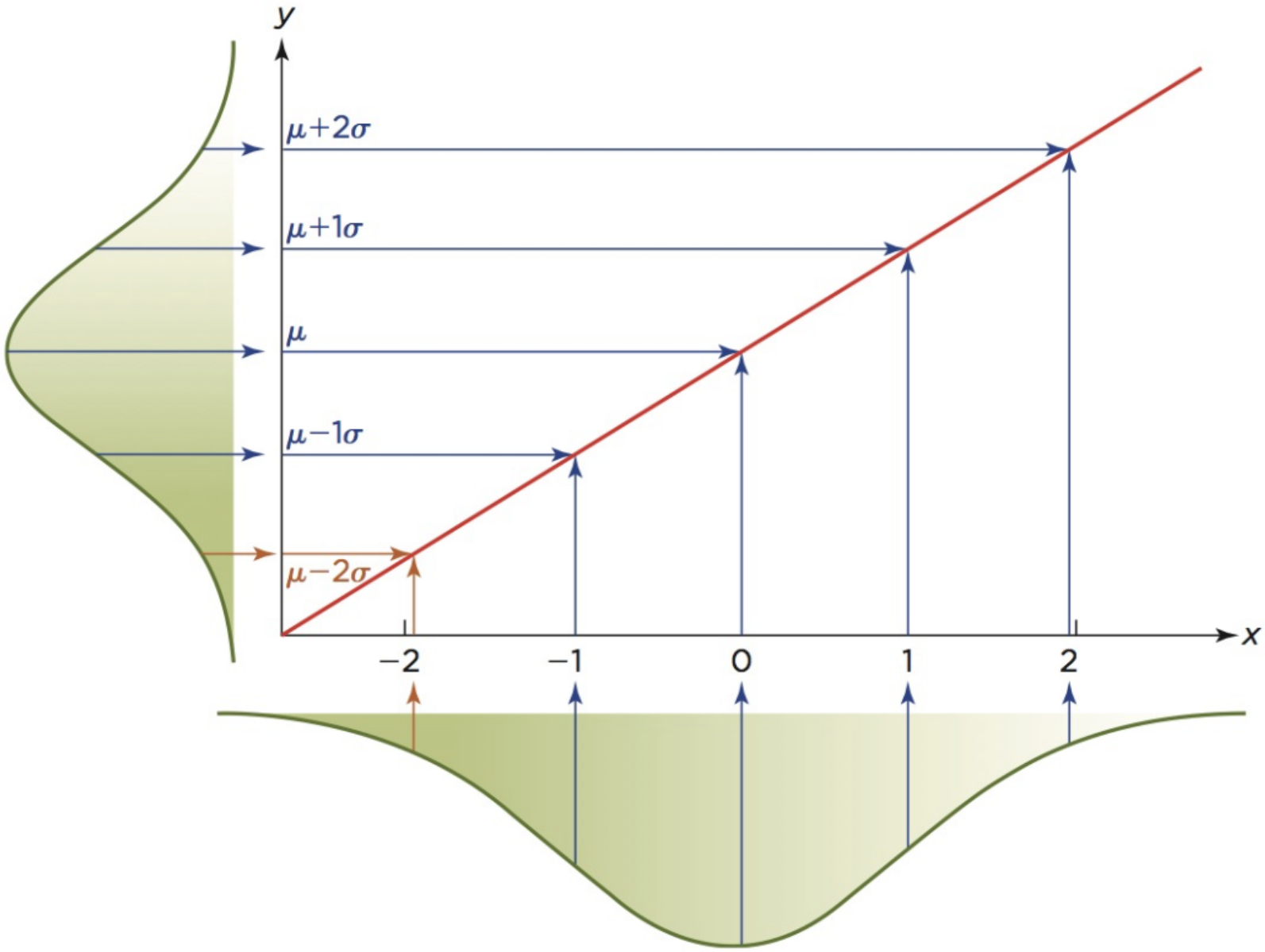
```
plot(faithful$waiting, eruption.res, ylab="Residuen",
 xlab="Wartezeit", main="Eruptionen von Old Faithful")
abline(0,0)
```



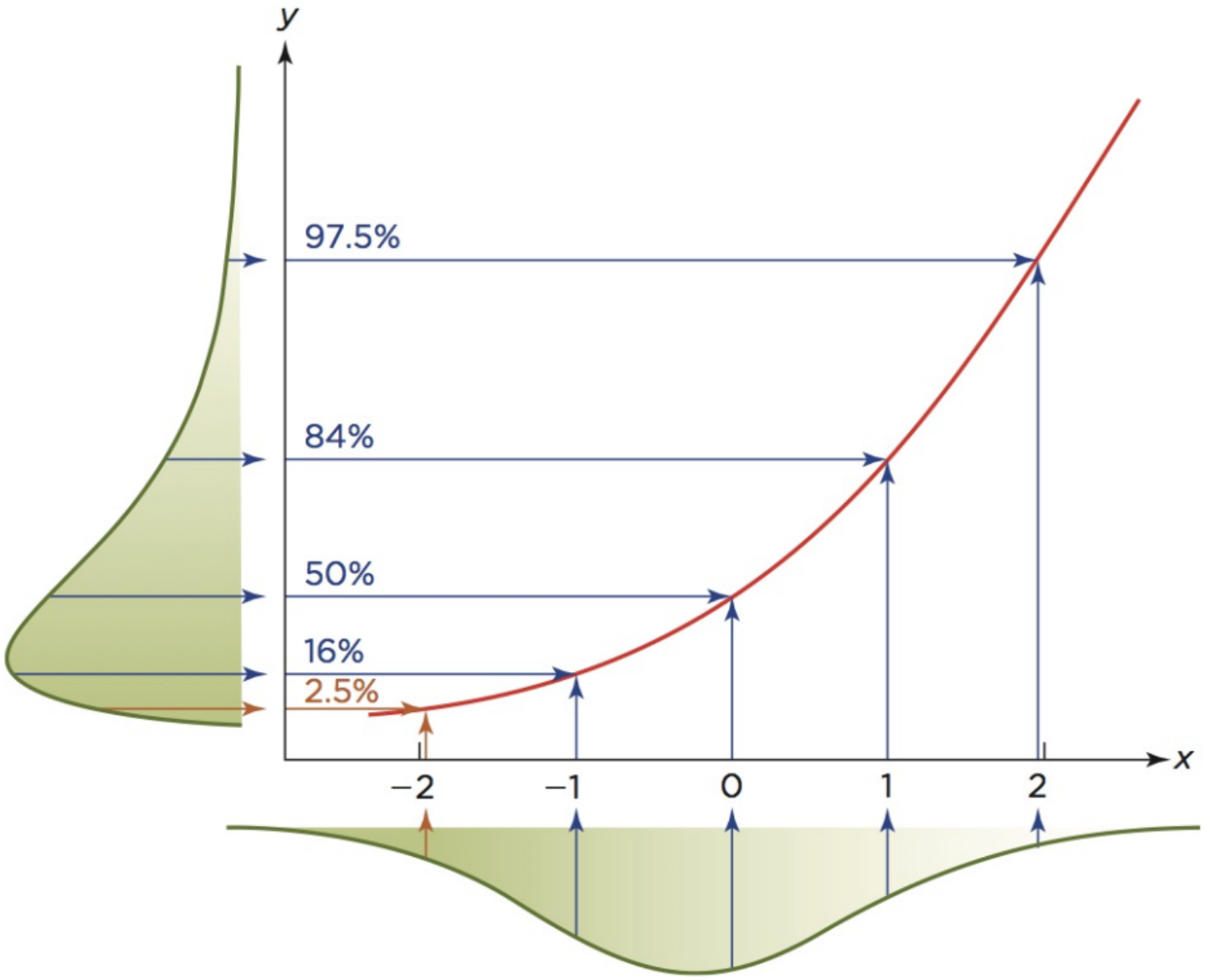
Lineare Regression: QQ-Plot

- Mit dem Normal-Wahrscheinlichkeits-Diagramm (auch Quantile-Quantile-Plot) der Residuen vergleichen wir die Residuen mit der Normalverteilung.

Lineare Regression: QQ-Plot



Lineare Regression: QQ-Plot



Lineare Regression: QQ-Plot

Problem: Erstellen Sie das Normal-Wahrscheinlichkeits-Diagramm der Residuen aus dem Datensatz `faithful`.

Lineare Regression: QQ-Plot

Antwort:

```
plot(eruption.lm, which=2)
```