# Crashkurs in Regressionsanalyse

Prof. Dr. Johannes Binswanger

*Herbst 2016*

Universität St.Gallen

# KENNZAHLEN ZUM ZUSAMMENHANG VON VARIABLEN

- Oft werden Daten angeschaut, um eine Schlussfolgerung zu ziehen
  - Etwa, wie stark ist die Exportbranche durch Wechselkursschwankungen bedroht)
- Solche Schlüsse sind oft eine Basis für rationales Planen.
- Visuelle Darstellungen helfen, Einsicht in die Sachlage zu bringen…
- … aber oft bedarf es der Hilfe von rechnerischer Statistik, um zuverlässige Schlüsse zu ziehen.

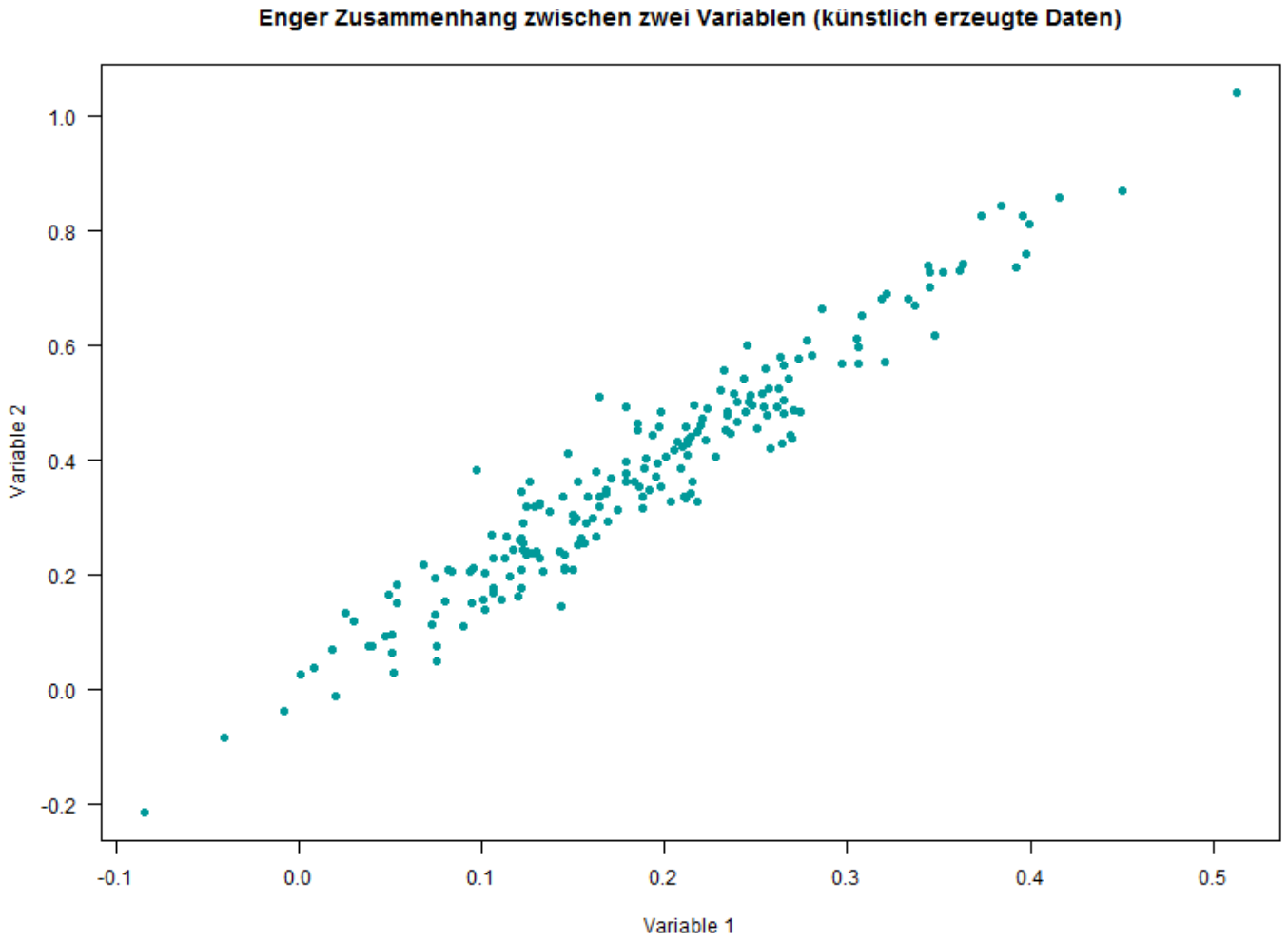
# KENNZAHLEN ZUM ZUSAMMENHANG VON VARIABLEN

- Konkret möchten wir von der rechnerischen Statistik folgende Informationen haben über den Zusammenhang zweier Variablen:
  - Wenn die eine Variable um x Prozentpunkte steigt, um wie viele Prozentpunkte können wir erwarten, dass die andere Variable steigt?
  - Wie „eng“ oder lose ist der Zusammenhang zwischen den beiden Variablen, ausgedrückt durch eine Zahl zwischen 0 und 1?
  - Wie zuversichtlich können wir sein, dass zwischen den beiden Variablen *überhaupt* ein Zusammenhang besteht (ausgedrückt in einer Zahl zwischen 0 und 1)?

# REGRESSIONSANALYSE

- Um zu den Kennzahlen zu gelangen, bedienen wir uns einer Technik, die man *Regressionsanalyse* nennt.
  - Sie kennen das vielleicht auch von «Trendlinie hinzufügen» in Excel.
  - Wir betrachten die Regressionsanalyse erst für – zu didaktischen Zwecken – künstlich erzeugte Daten.

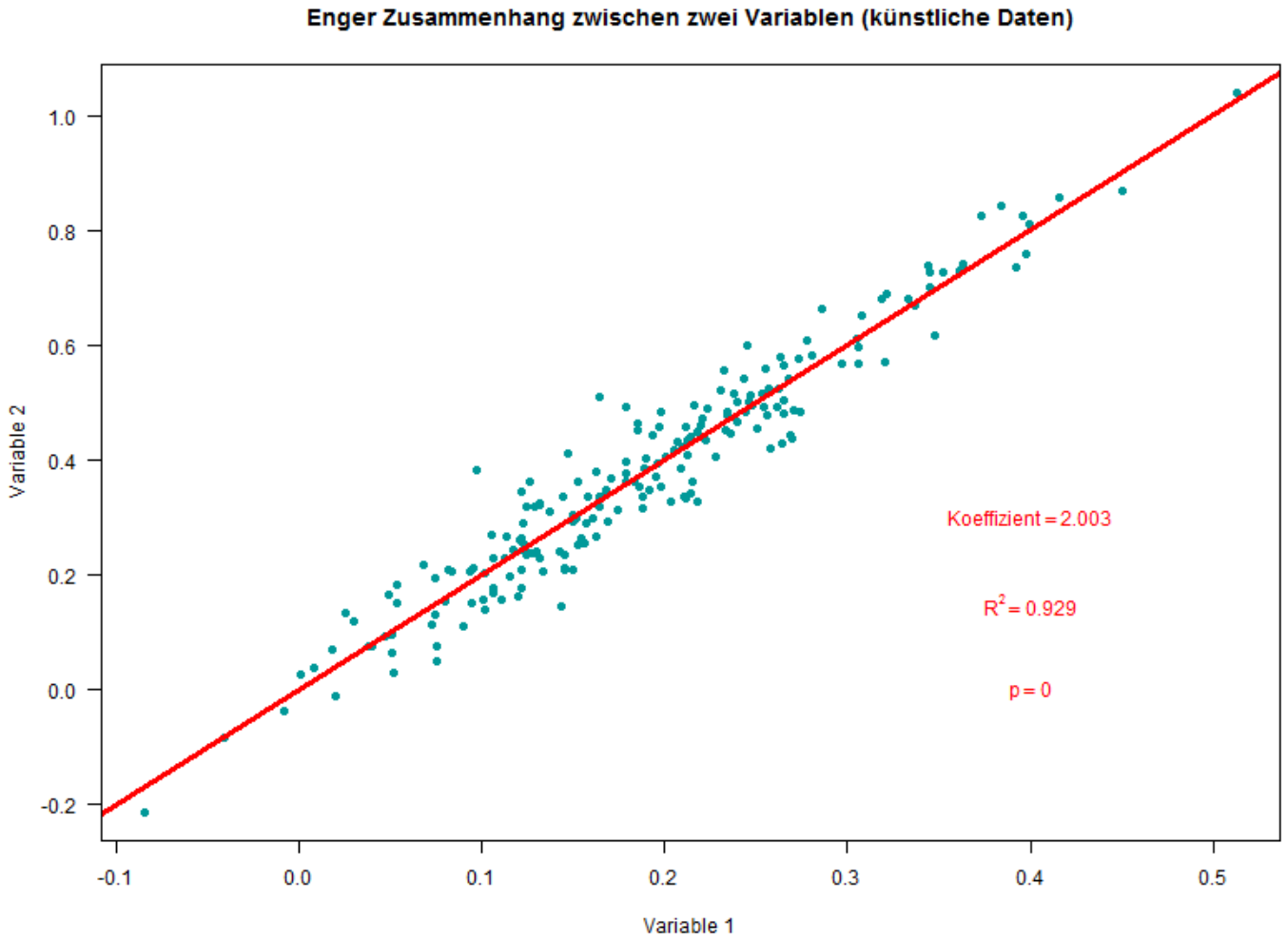
# REGRESSIONSANALYSE



# REGRESSIONSANALYSE

- Rein visuell besteht in der vorangehenden Graphik offensichtlich ein enger Zusammenhang zwischen den beiden (künstlichen didaktischen) Variablen.
- Frage: Wie kann man diesen Zusammenhang *messen*?

# REGRESSIONSANALYSE



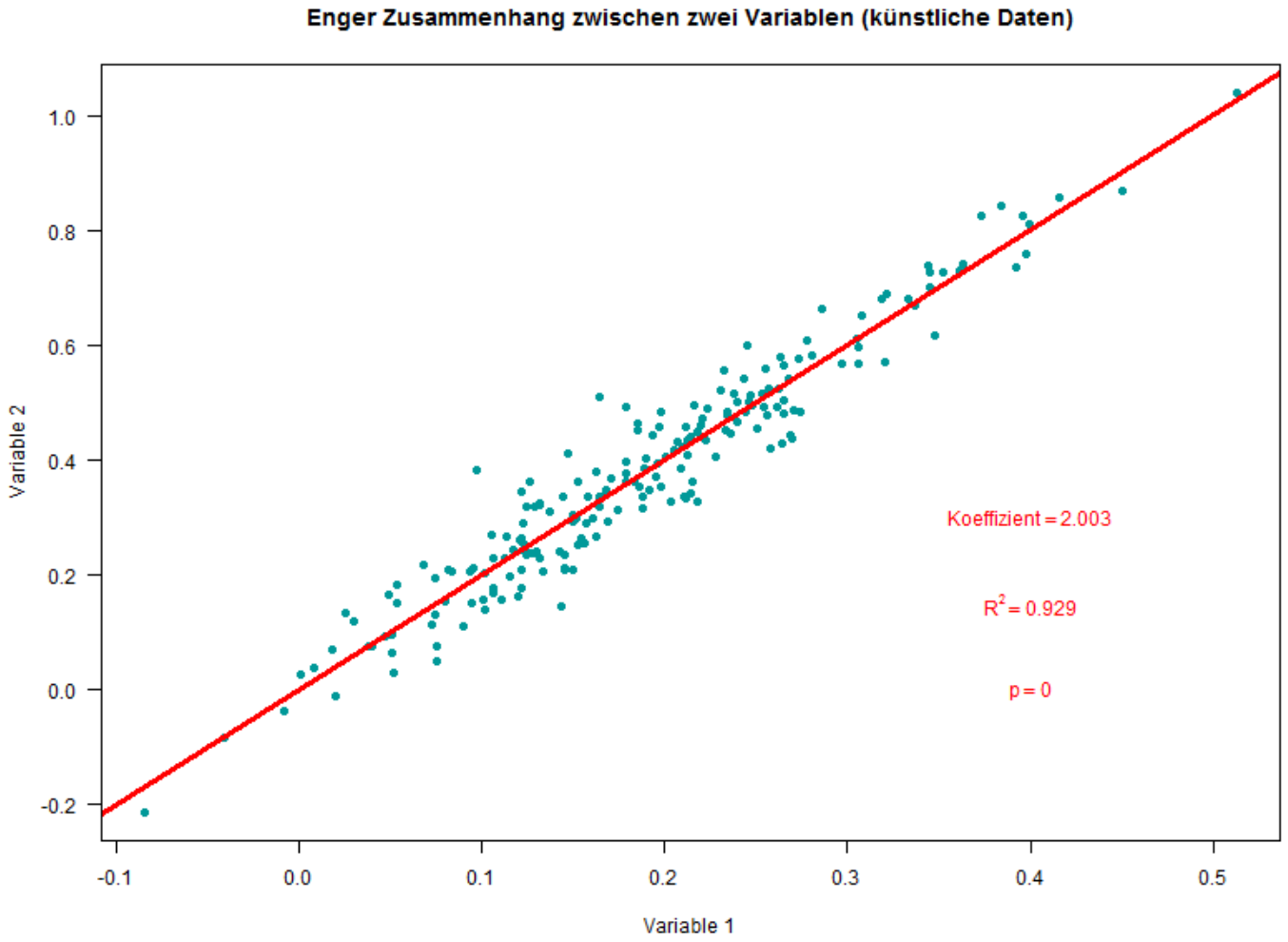
# REGRESSIONSANALYSE

- Lösung: Wir legen eine Gerade durch die Punkte, so dass die Punkte im Mittel «am nächsten» bei der Gerade liegen («Trendlinie»).
  - Details brauchen uns hier nicht zu kümmern.
- Diese Gerade heisst Regressionsgerade.
  - Vermutlich das am meisten verwendete statistische Konzept überhaupt, abgesehen vom Mittelwert.
- Falls die Abstände zur Gerade in Summe sehr klein sind, besteht ein «enger» Zusammenhang zwischen zwei Variablen.

# REGRESSIONSANALYSE

- Die Regressionstechnik liefert uns eine Reihe von Kennzahlen, wovon uns drei besonders interessieren:
  - Kennzahl für Wert der Steigung der Regressionsgerade: *(Steigungs-)Koeffizient*
  - Kennzahl, wie gut die Gerade die Streuungstendenz in der Punktwolke «nachbildet»: $R^2$ (auch *Bestimmtheitsmass* oder *Determinationskoeffizient*)
  - Kennzahl dafür, wie wahrscheinlich es ist, die Daten zu beobachten, so wie sie sind, wenn man davon ausgeht, dass **kein** Zusammenhang/Trend besteht: *p-Wert* (oder *Wahrscheinlichkeitswert*)

# REGRESSIONSANALYSE: STEIGUNGSKOEFFIZIENT



# REGRESSIONSANALYSE: STEIGUNGSKOEFFIZIENT

- Der Steigungskoeffizient (oder, kurz, Koeffizient) im vorangegangenen Beispiel ist 2.003.
- Bedeutung: Wenn die Variable auf der horizontalen Achse um eine Einheit zunimmt, dann nimmt die Variable auf der vertikalen Achse *trendmässig* um 2.003 Einheiten zu
  - Anstatt *trendmässig* sagt man auch *erwartungsgemäss* oder *im Durchschnitt*

# REGRESSIONSANALYSE: STEIGUNGSKOEFFIZIENT

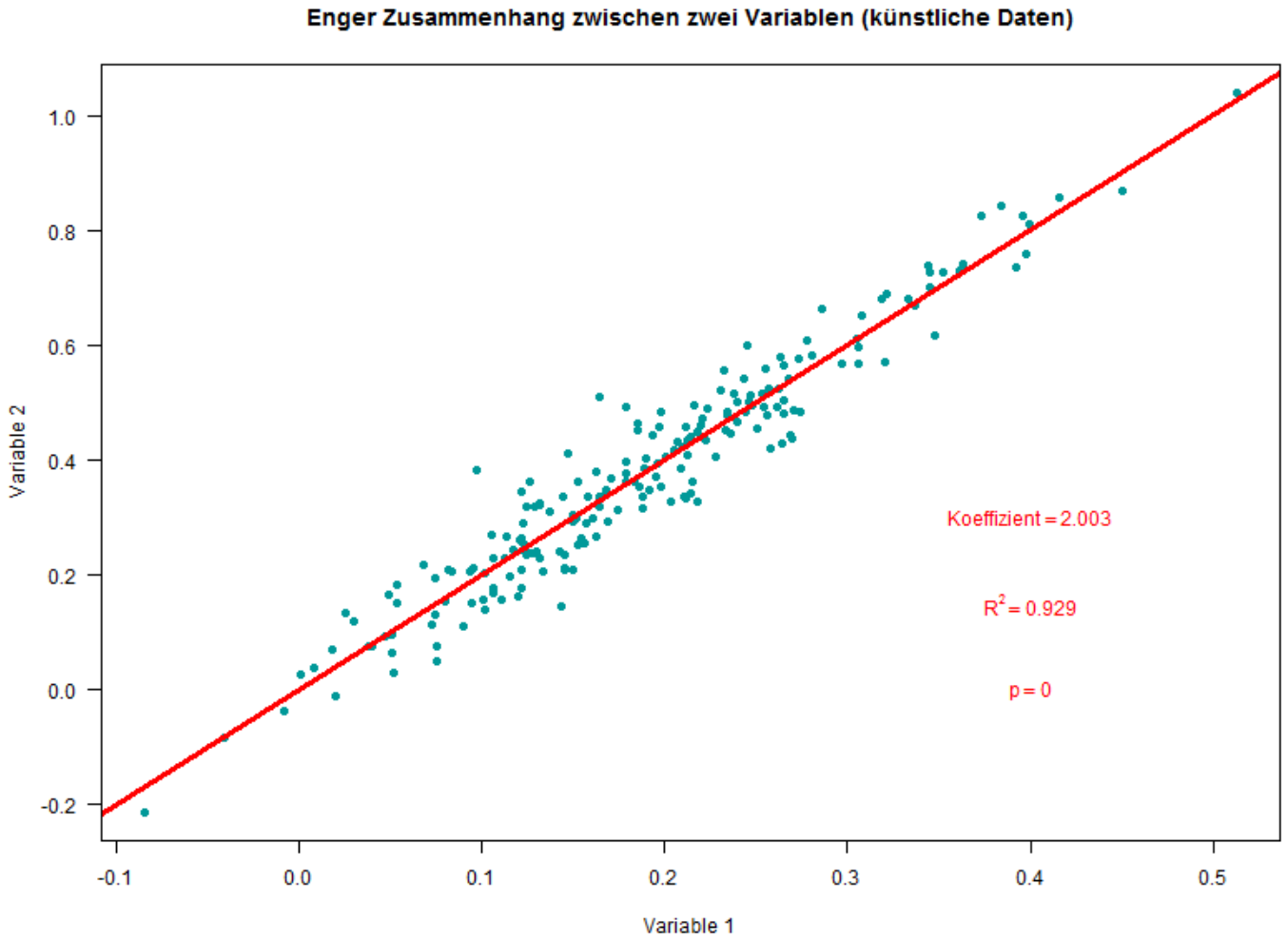
- Trendmässig heisst natürlich nicht, dass man sich sicher sein kann
  - Die einzelnen Punkte liegen ja nicht *auf* der Geraden
  - Es gibt «Störeinflüsse» von anderen Einflussfaktoren

- Wie gut der Trend zu einer zuverlässigen Planung geeignet ist, wird durch die andern beiden Kennzahlen ausgedrückt.

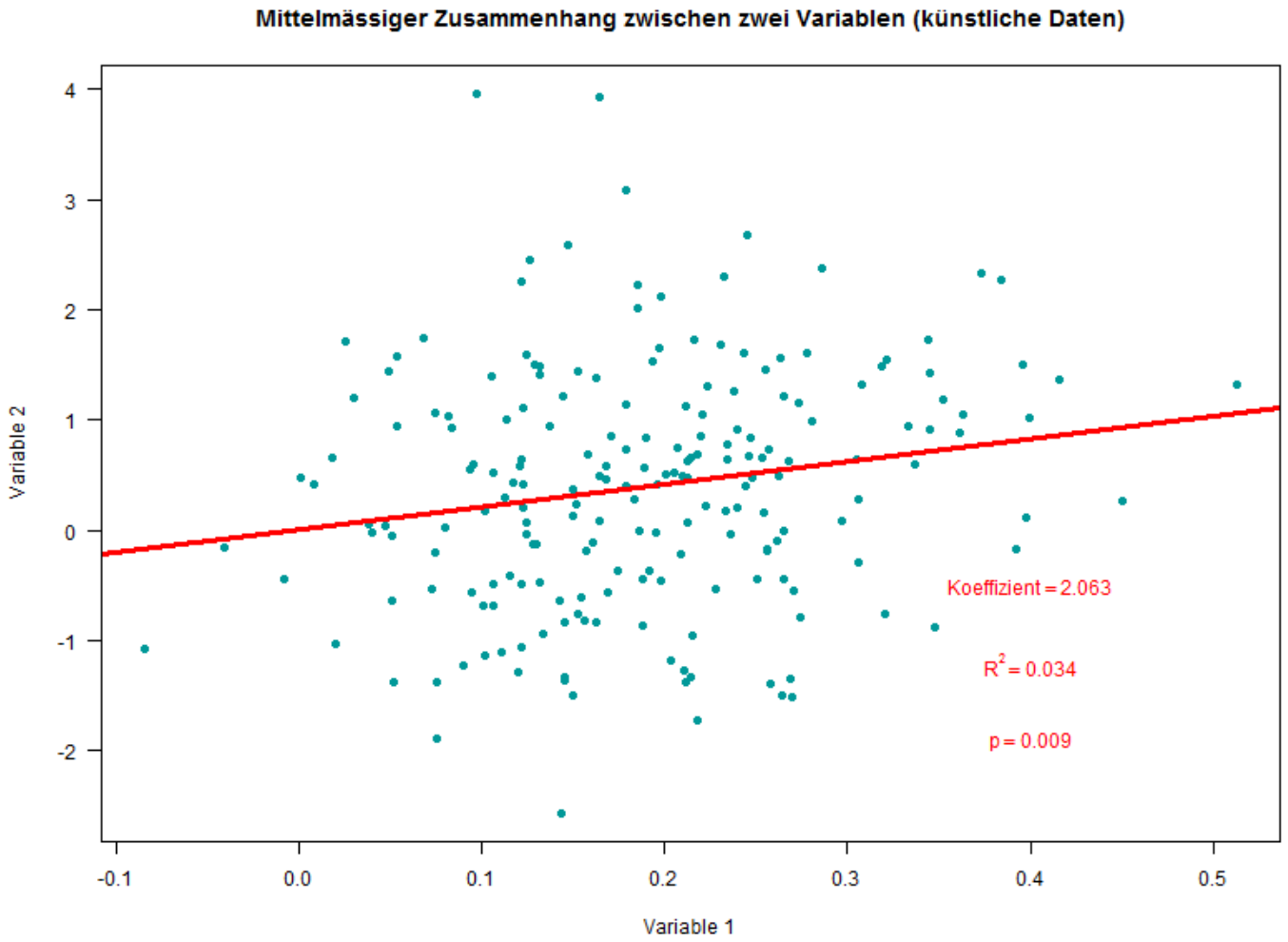
# REGRESSIONSANALYSE: $R^2$

- «Eichung» der Kennzahl zur «Repräsentativität» des Trends:
  - Alle Punkte liegen AUF der Gerade: Wert 100% (oder 1)
  - Mindestens ein paar Punkte liegen nicht auf der Gerade: Wert <100% (<1)
  - Minimaler Wert: 0
- Die Resultierende Grösse heisst $R^2$ («Bestimmtheits-mass» oder «Determinationskoeffizient»).
- Wir vergleichen erst einmal drei Situationen.

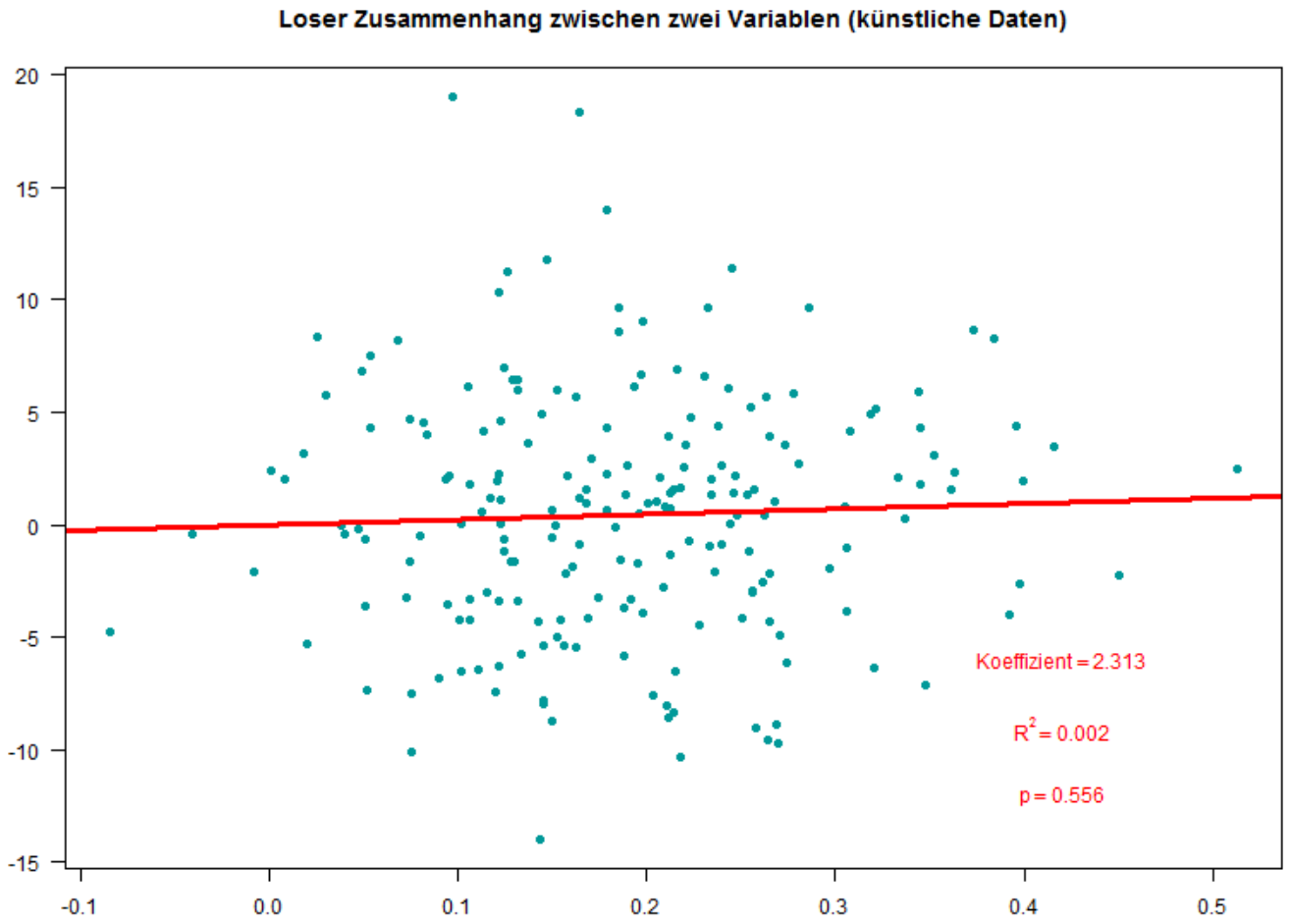
# REGRESSIONSANALYSE: $R^2$



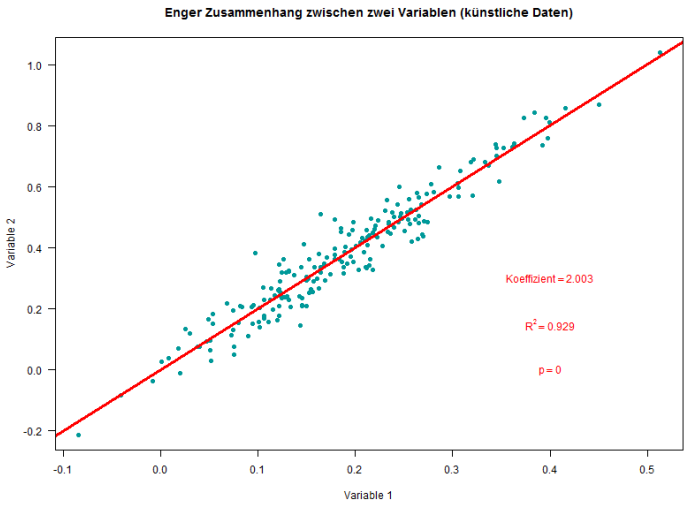
# REGRESSIONSANALYSE: $R^2$

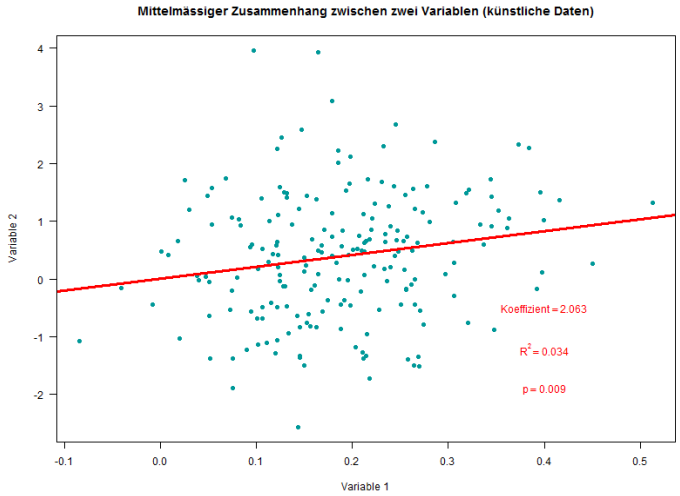


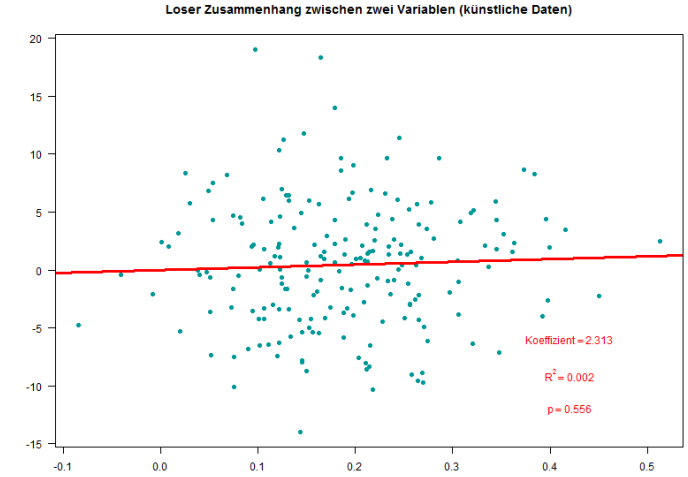
# REGRESSIONSANALYSE: $R^2$



# REGRESSIONSANALYSE: $R^2$



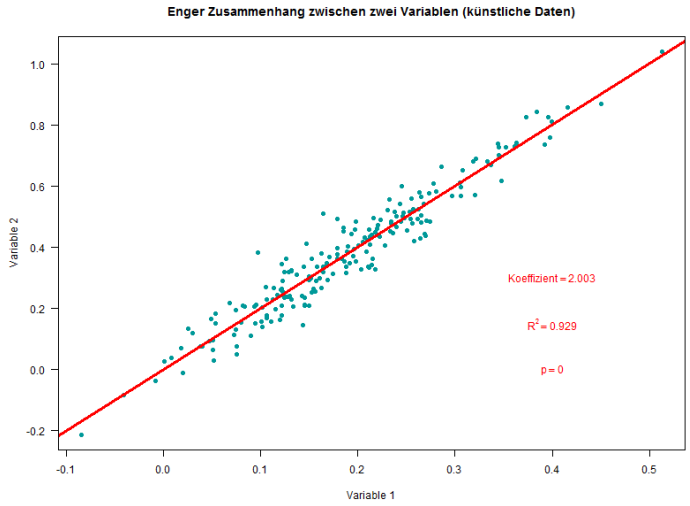




Beachten Sie, dass die Skalen der vertikalen Achsen ganz unterschiedlich sind, die Regressionslinie/ Trendkurve ist de facto immer gleich steil!

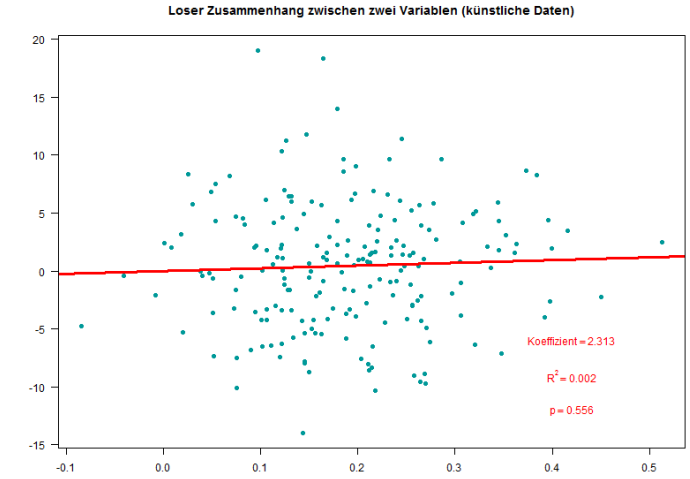
# REGRESSIONSANALYSE: $R^2$

- Es gibt zwei Arten von Streuung der Punkte:
  - Streuung, welche durch die Regressionslinie/Trend «erklärt» wird
  - Streuung welche *nicht* «erklärt» wird

# REGRESSIONSANALYSE: $R^2$




- Die Regressionsgerade «erklärt» sehr viel der Streuung: Hohes $R^2$.




- Die Regressionsgerade «erklärt» praktisch nichts von der Streuung: $R^2$ nahe bei 0.

# REGRESSIONSANALYSE: $R^2$

- Formal gibt das $R^2$ das Verhältnis an der Streuung, die durch die Regressionsgerade erklärt wird, zur gesamten Streuung:
  - Alles wird erklärt: 100%/100% → Alle Punkte liegen *auf* der Geraden, $R^2$ = 100% (oder 1)
  - Nichts wird erklärt: 0%/100% → Kein Trend, Gerade ist flach, $R^2$ = 0
- Je höher das $R^2$, desto nützlicher ist eine Variable zum Planen!

# REGRESSIONSANALYSE: P-WERT

- Die dritte Kennzahl ist der p-Wert.
- Der p-Wert ist ein Indikator dafür, inwiefern *überhaupt* ein Zusammenhang/Trend besteht zwischen zwei Variablen.
- Aufgrund der Systematik der zugrundeliegenden statistischen Konzepte ist das Konzept des p-Wertes etwas «auf den Kopf gestellt»:

  *Der p-Wert beantwortet die Frage: Wie wahrscheinlich ist es, dass ich die Daten beobachte, so wie sie sind, wenn ich davon ausgehe, dass **kein** Zusammenhang/ Trend besteht. → Wert zwischen 0 (Zusammenhang extrem wahrscheinlich) und 100 (Zusammenhang extrem unwahrscheinlich).*

# REGRESSIONSANALYSE: P-WERT

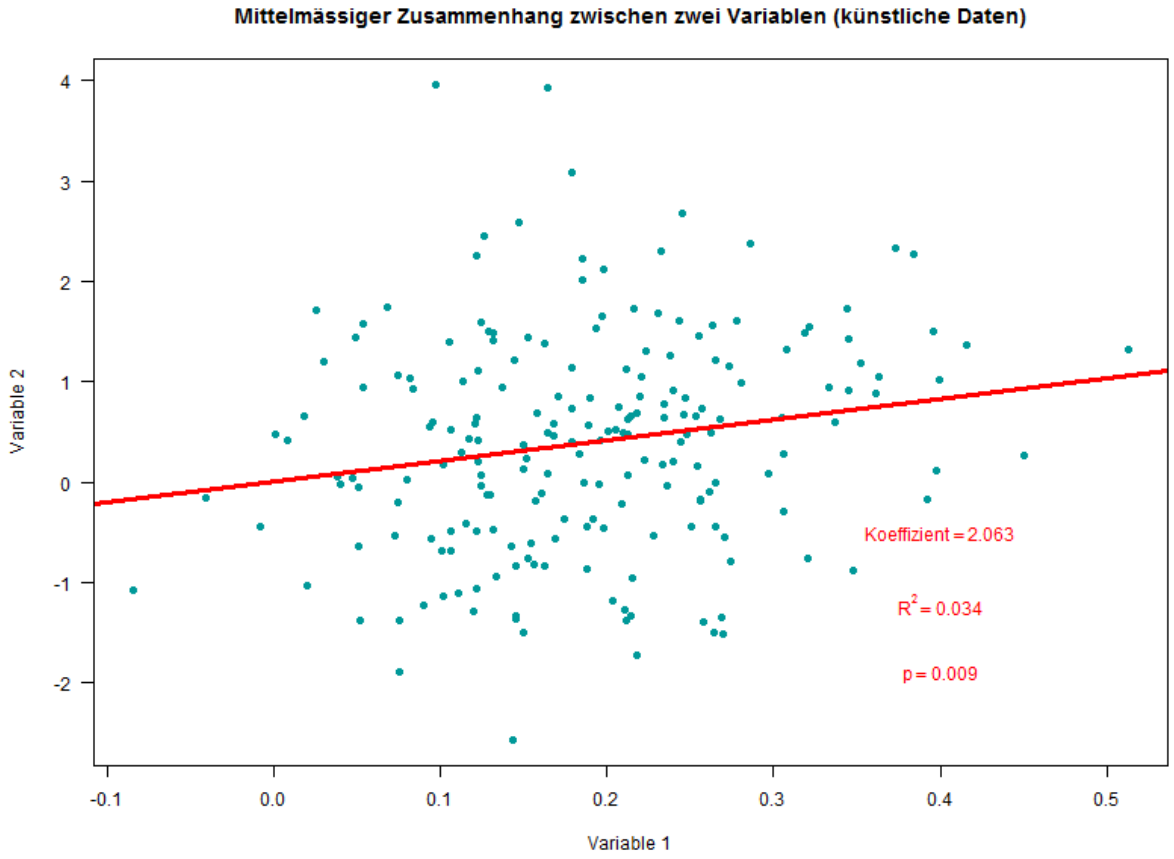
- Im Fachjargon nennt man einen systematischen Einfluss, gemessen durch einen niedrigen p-Wert, *signifikant*.
  - Es ist üblich, zu sagen, dass «Variable 1» signifikant ist (bezüglich ihres Einflusses auf «Variable 2») wenn der p-Wert unter 0.05 (gleichbedeutend mit 5%) zu liegen kommt.
  - Man sagt auch, dass Variable 1 signifikant sei auf dem Niveau von p.
    - Allerdings redet man hiervon nur, wenn p relativ nahe bei 0 liegt, sicher mindestens unter 15%.

# REGRESSIONSANALYSE: P-WERT

- Achtung, der p-Wert und das $R^2$ messen zwei unterschiedliche Dinge!
- Auch bei einem niedrigen $R^2$ kann der p-Wert hoch sein!
  - Eine Variable kann auf das Gesamtergebnis nur einen bescheidenen Einfluss haben im Vergleich zu anderen Faktoren, aber dieser kann dennoch sehr robust/systematisch sein.
  - Z.B. hat Meereshöhe systematischen Einfluss auf Temperatur, auch wenn viele andere Faktoren einen vielleicht grösseren Einfluss haben.
  - «Nur einen geringen Teil der Streuung erklären» ist nicht dasselbe wie «kein Einfluss haben»!

# REGRESSIONSANALYSE: P-WERT

Hier sehen wir ein niedriges $R^2$ und gleichzeitig einen sehr niedrigen p-Wert.



# SCHWEIZER EXPORTE UND WECHSELKURSE

- Nun sind wir gerüstet, Regressionen mit echten Daten zu verstehen…