Das lineare Regressionsmodell

Termin 5

Jakob Kapeller

University of Duisburg-Essen
Institute for Socio-Economics &

Johannes Kepler University Linz
Institute for Comprehensive Analysis of the Economy (ICAE)

Editor: *Heterodox Economics Newsletter*

www.jakob-kapeller.org | www.uni-due.de | www.heterodoxnews.com

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

ifso institute for socio-economics

Agenda

- Regressionsanalyse: Motivation und Illustration
  - Motivation und Einführungsbeispiel: Typische Schritte der Regressionsanalyse
  - Erste Kritik und Erweiterung des Regressionsmodells
- Grundlegende Aspekte der Regressionsanalyse
  - Grundbegriffe im Kontext der Regressionsanalyse
  - Formale Aspekte der Regressionsanalyse
- Regressionsanalyse: Erste Erweiterungen
  - Dummy Variablen
  - Logarithmus
  - Funktionale Formen
  - Interaktionsterme

Hinweise

- Der Teil zur quantitativen Sozialforschung wird sie sehr unterschiedlich fordern
  - Insgesamt gibt es viel Stoff: Teamwork ist das A und O
  - Helfen Sie sich gegenseitig - bei der Nachbereitung *und* den Aufgabenblättern
- Der heutige Termin soll vor allem einen Ausblick geben
  - Richtet sich vor allem an jene, die kaum oder keine quantitativen Methodenkurse hatten
  - Falls Sie in bestimmten Bereichen Schwierigkeiten haben, nutzen Sie die Zeit bis zu Termin 10 zur Wiederholung

- Nutzen Sie das R Tutorium und das Forum für Fragen zur praktischen Implementierung und lesen Sie die passenden Kapitel aus dem Skript.

Motivation und Illustration

Motivation

- Die Regressionsanalyse ist die Standardmethode wenn wir am Zusammenhang zwischen quantitativen Größen interessiert sind
  - Durch den *Empirical Turn* (?) in den WiWi noch wichtiger geworden
  - Positiv formuliert: sie bietet eine Methode mit großem Anwendungsfeld
  - Negativ formuliert: die Sozialwissenschaften, insbesondere die Ökonomik, haben eine Obsession für Regressionsanalysen

- In jedem Fall: Grundkenntnisse sind wichtig nicht nur fürs Selbermachen, sondern auch um die aktuelle Forschungslandschaft zu verstehen

Regression zwischen Theorie und Empirie

- Empirische Analyse kann theoriegeleitet sein, muss aber nicht.
  - Meistens haben wir aber eine theoretische Intuition / Vermutung: quantitativ prüfen!
  - Unsere Vermutung: Gewerkschaften behindern Innovationstätigkeit (Union↑ → Inno↓)
  - Erste Intuition: Was ist die „Korrelation" zwischen beiden Variablen?

Get some data (OECD)

R&D in % GDP

```
> head(unioninno, 3)
   Country Year UnionDensity     Tech
1:     AUS 1990         41.3 1.259877
2:     AUS 1991         42.0       NA
3:     AUS 1992         39.2 1.462098
```

10 Länder 30 Jahre % of labor force

Find some way to estimate the relationship!

Kovarianz: $Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sigma_{X,Y}$

Korrelationskoeffizient: $Korr(X,Y) = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{(Var(Y)}}$

$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

gentle reminder…

```
> unions <- dplyr::select(oecd_data, UnionDensity)
> tech <- dplyr::select(oecd_data, Tech)
> cor(unions,tech, use="complete.obs")
                Tech
UnionDensity 0.2442741
```

Korrelationskoeffizent > 0

Erster Test weist eher auf positiven Konnex hin.

- Regressionsverfahren beruht auf derselben Intuition: Korrelation messen
  - Zusammenhang wird durch einen bestimmten Parameter (meist: β) ausgedrückt.
  - Vorteile: Größere Flexibilität, mehr Variablen, explizites mathematisches Modell

Von der Intuition zur Regression

Unsere Vermutung: Union↑ → Inno↓

- **Theoretisches Modell:** Wir formulieren Hypothese als lineares Modell, z.B.

$$Inno = b + m \cdot UnionDensity$$ → Einfach eine lineare Funktion: $y = mx + b$

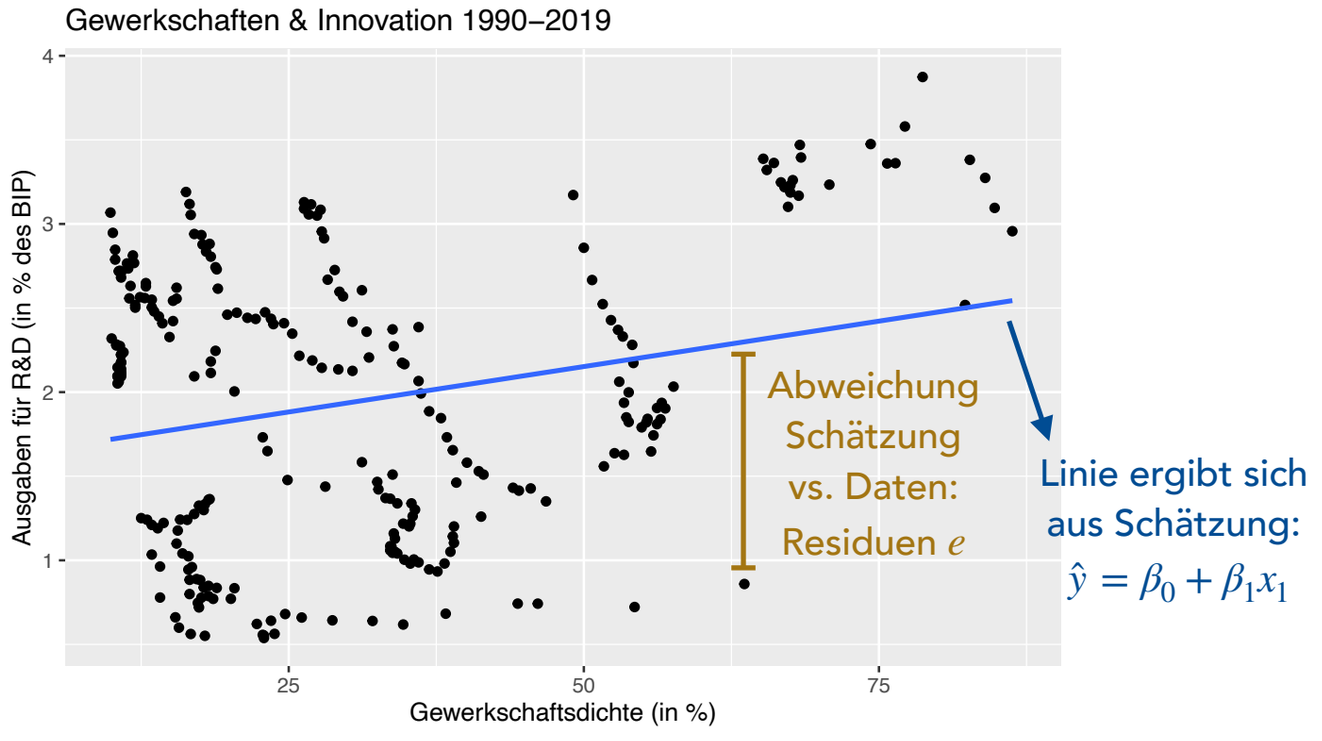
- **Statistische Gleichung:** Andere Schreibweise und Fehlerterm ϵ

$$Inno_i = \beta_0 + \beta_1 UnionDensity_i + \epsilon_i$$ → ϵ_i als Abweichung der Daten von der unterstellten Funktion.

abhängige Variable / „erklärte" Variable / Regressand

unabhängige Variable / „erklärende" Variable / Regressor

- Das Modell „schätzen" →
  - **Rechnerische Interpretation:** Werte für β_0 und β_1 finden.
  - **Graphische Interpretation:** Linie mit dem kleinsten (quadrierten) Abstand zu Daten finden



Regression: Typische Arbeitsschritte

- (1) Was wird gemessen? Ausgangspunkt Theorie: (X → Y, hier als: Union↑ → Inno↓)
  Statistisches Modell: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- (2) In Statistikprogramm schätzen und Output interpretieren:

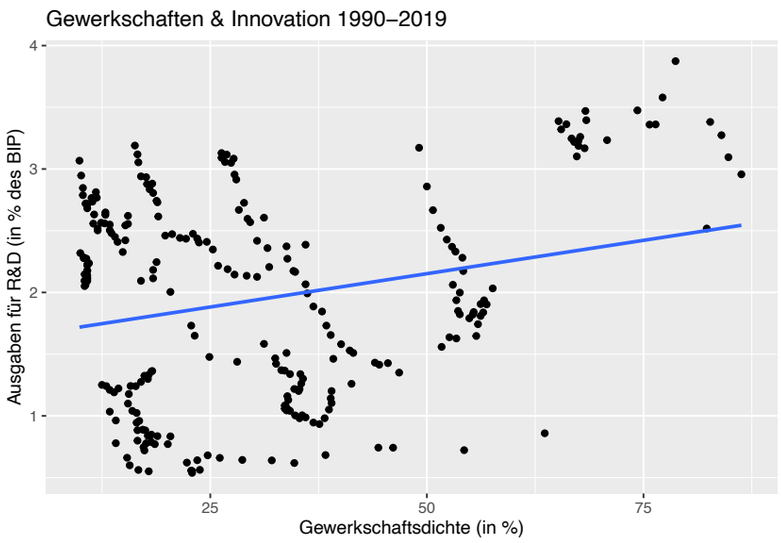
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.61192    0.09779   16.48  < 2e-16 ***
UnionDensity  0.01080    0.00268    4.03 7.35e-05 ***
```

„statistische Signifikanz"

„Anteil der erklärten Variation in den Daten"

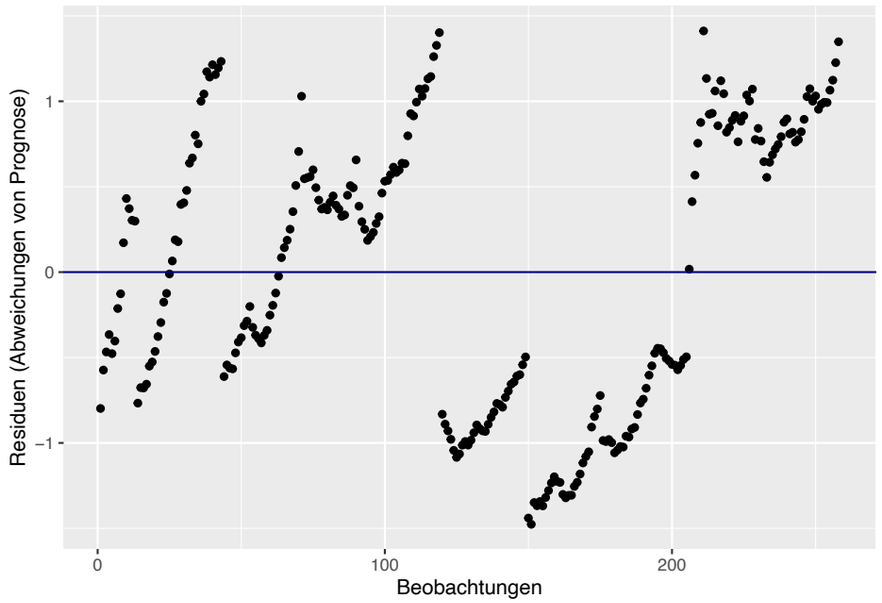
$$Inno_i = 1.6 + 0.01 \cdot UnionDensity_i$$

- (3) Ergebnisse plotten



- (4) Statistische Kennzahlen prüfen (z.B. p-Wert, R^2)
- (5) Residuen e analysieren (gleichmäßig?)
- Intuition: Wenn Modell stimmt müssen Residuen „zufällig" sein (regelmäßiges „weißes Rauschen")…

Residuen e als Differenzen zwischen Regressionsgerade & Datenpunkten



Das Einführungsmodell in R

```
> techmodel <- lm(Tech ~ UnionDensity, data=oecd_data)
> summary(techmodel)
```

```
Call:
lm(formula = Tech ~ UnionDensity, data = oecd_data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4763 -0.7305  0.1139  0.7593  1.4117
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.61192    0.09779   16.48  < 2e-16 ***
UnionDensity  0.01080    0.00268    4.03 7.35e-05 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.8111 on 256 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.05967,   Adjusted R-squared:  0.056
F-statistic: 16.24 on 1 and 256 DF,  p-value: 7.346e-05
```

Die F-Statistik testet die gemeinsame Signifikanz aller erklärenden Variablen: Nullhypothese ist, dass für alle Parameter die Nullhypothese hält.

Wir lassen das Modell mit dem Befehl `lm` laufen, speichern es als `techmodel` und rufen mit `summary` einen ausführlichen Bericht auf.

R schreibt hier nochmal unsere Eingabe ab.

Hier werden Angaben zur Verteilung der Residuen gemacht.
Bezug zu Schritt (5) auf vorangegangener Folie!

Das ist der zentrale Output mit Angaben zur Größe der Parameter (1. Spalte) und zur statistischen Signifikanz (letzte Spalte).

Legende für Symbole in letzter Spalte.

Fehlende Datenpunkte und Freiheitsgrade (= Zahl der freien Variablen: $(n - k)$)
(Standardfehler der Residuen: dazu später mehr)

Das R^2 gibt an welcher Anteil der Variation in der abhängigen Variable durch das Modell erklärt wird.

Regression: Zentrale Intuitionen

- Wir beschreiben den interessierenden Zusammenhang als lineare Funktion
  - Rechnerische und graphische Interpretation!
  - Wir können mit der errechneten Funktion auch Prognosen machen:

$$Inno_i = 1.6 + 0.01 \cdot UnionDensity_i$$

z.B.: Wenn $UD = 50\,\%$ dann ist $Inno = 2.15\,\%$

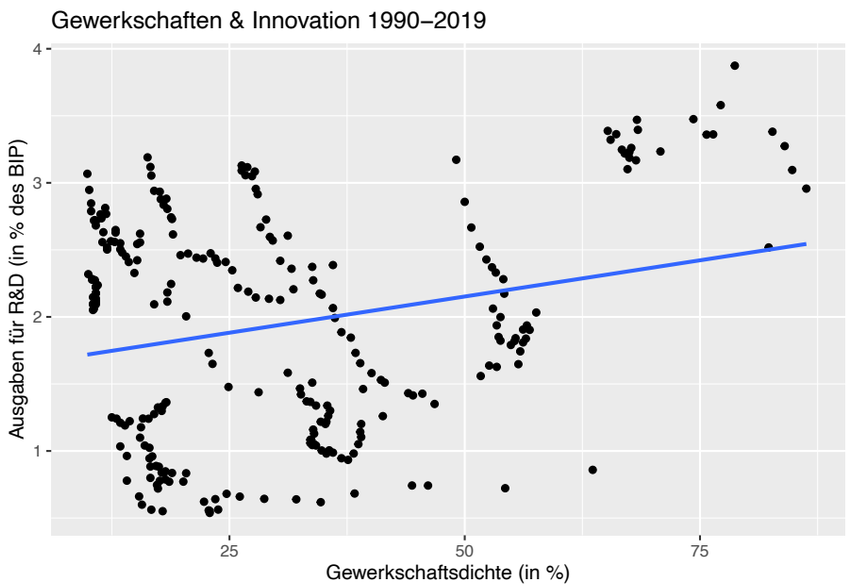
z.B.: Wenn $\Delta UD = 10\,\%$ dann ist $\Delta Inno = 0.11\,\%$

```
> beta_0 <- techmodel[["coefficients"]][1]
> beta_1 <- techmodel[["coefficients"]][2]
> unname(beta_0+beta_1*50)
[1] 2.152046
> unname(beta_1*10)
[1] 0.1080254
```

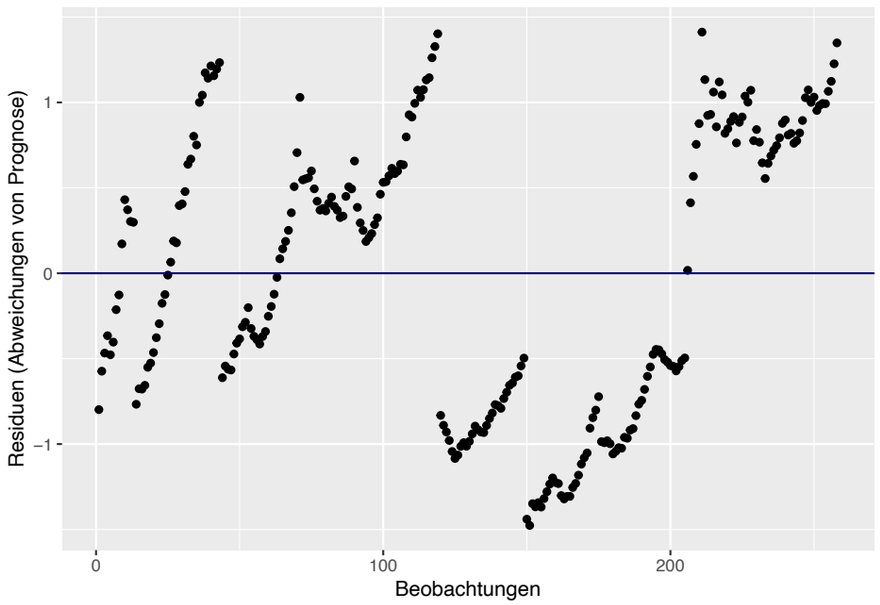
- Falsifikationsdenken ist auch hier hilfreich
  - Eine Korrelation ist kein Beweis eines kausalen Zusammenhangs, aber eine fehlende/ falsch gerichtete bivariate Korrelation taugt als (aller)erstes skeptisches Argument.
  - Ein statistisches Modell kann an statistischen Kriterien (z.B. Verteilung der Residuen, R^2) scheitern, aber es kann durch statistische Kriterien nicht theoretisch verifiziert werden.
- Rolle von Fehlertermen bzw. Residuen wichtig für besseres Verständnis
  - **Erste Lektion:** Fehlerterm (ϵ) ≠ Residuum (e). Erstere beziehen sich auch Population, zweitere auf Stichprobe (dazu später mehr).

Graphische Intuition I: Scatterplot und Residuenplot

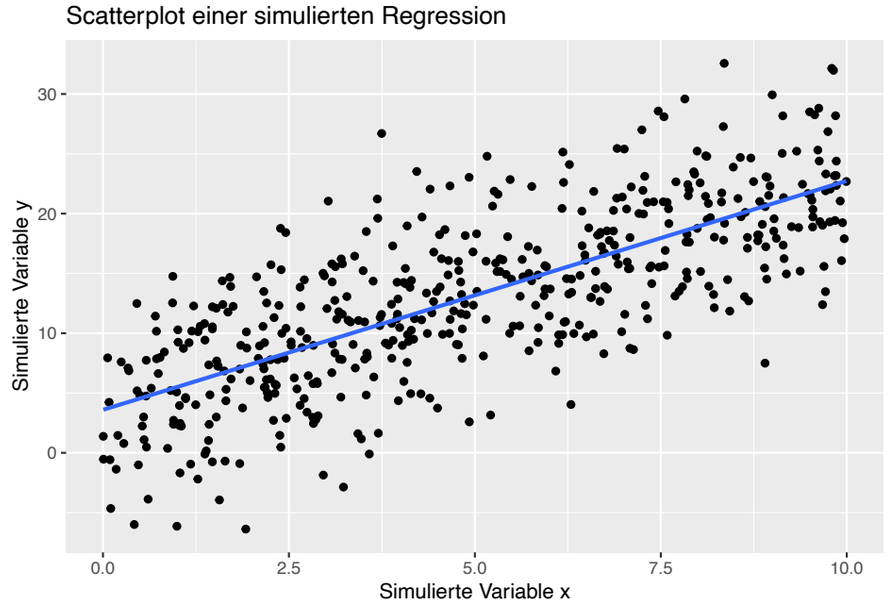
„Struktur in den Residuen“ vs. „Weißes Rauschen“



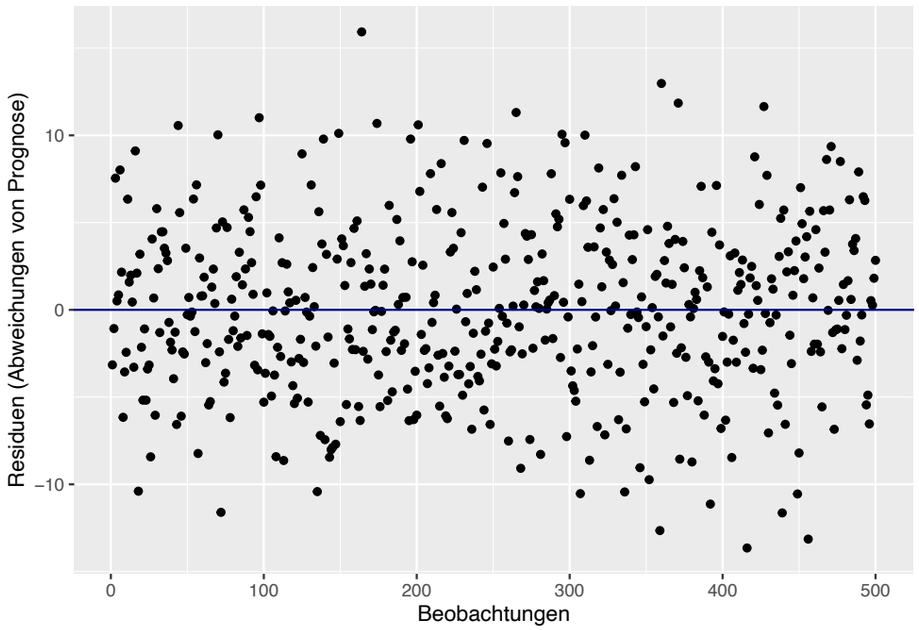
Scatterplot: Unser Einführungsmodell



Residuenplot: Unser Einführungsmodell



Scatterplot: "white noise“

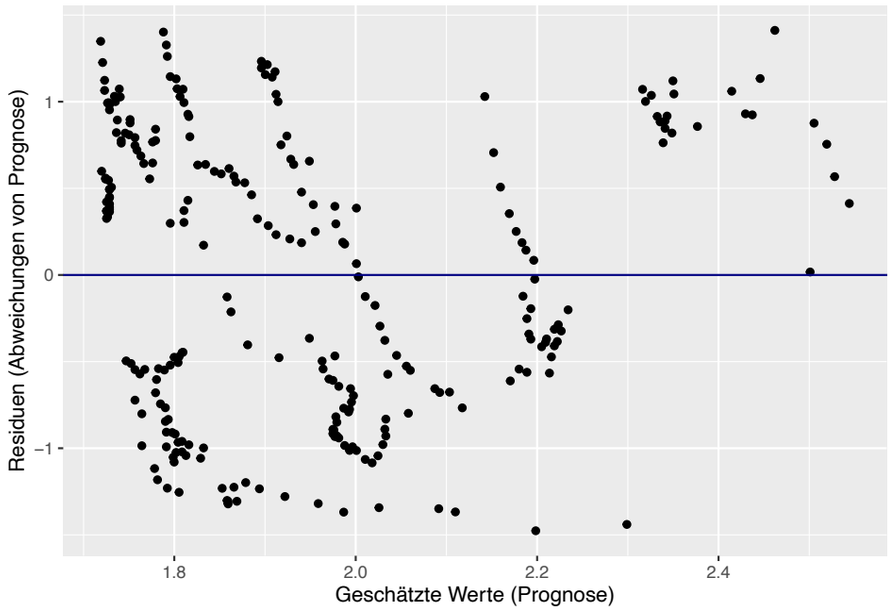


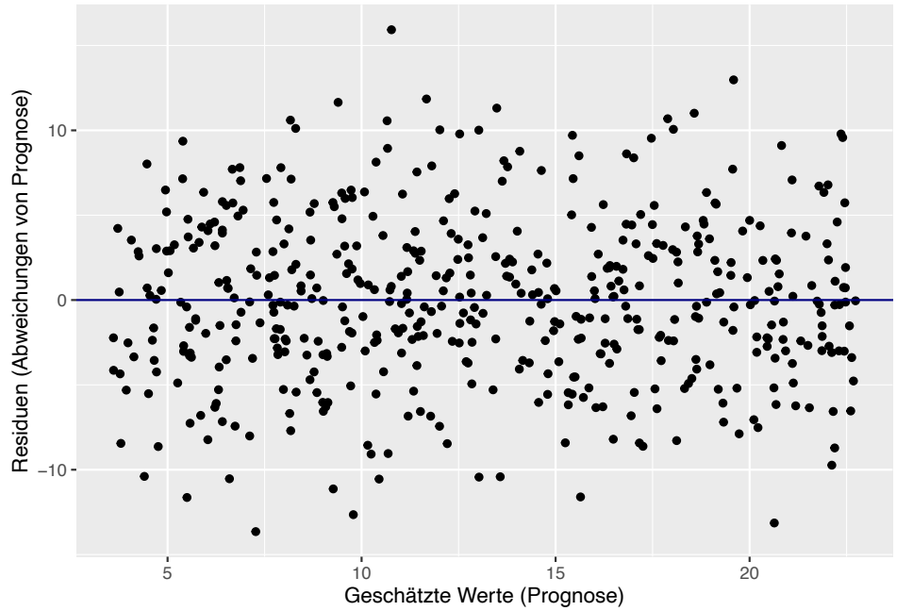
Residuenplot: "white noise“

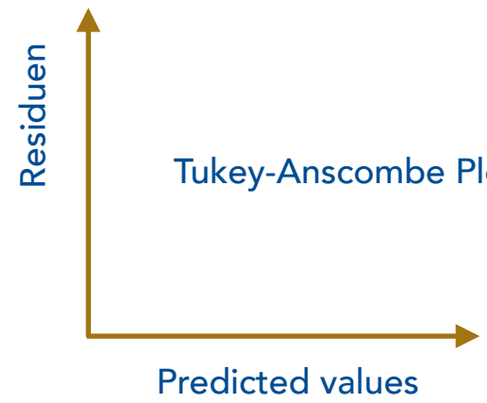
Graphische Intuition II: Tiefergehende Plots

Tukey-Ascombe-Plot und Plots mit größerer Informationstiefe

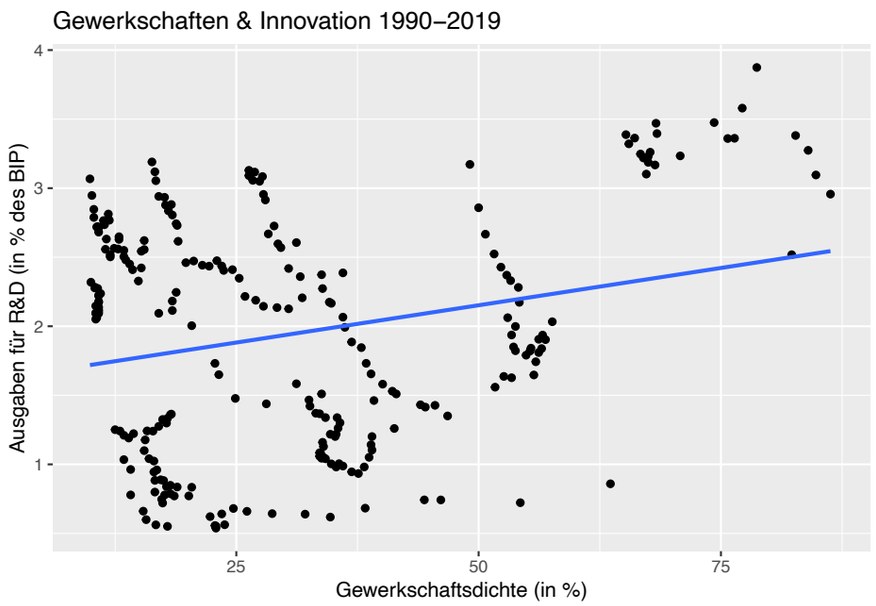
```
TAdata <- data.table("resids"=techmodel1[["residuals"]], "fittedvalues"=predict(techmodel1))
```





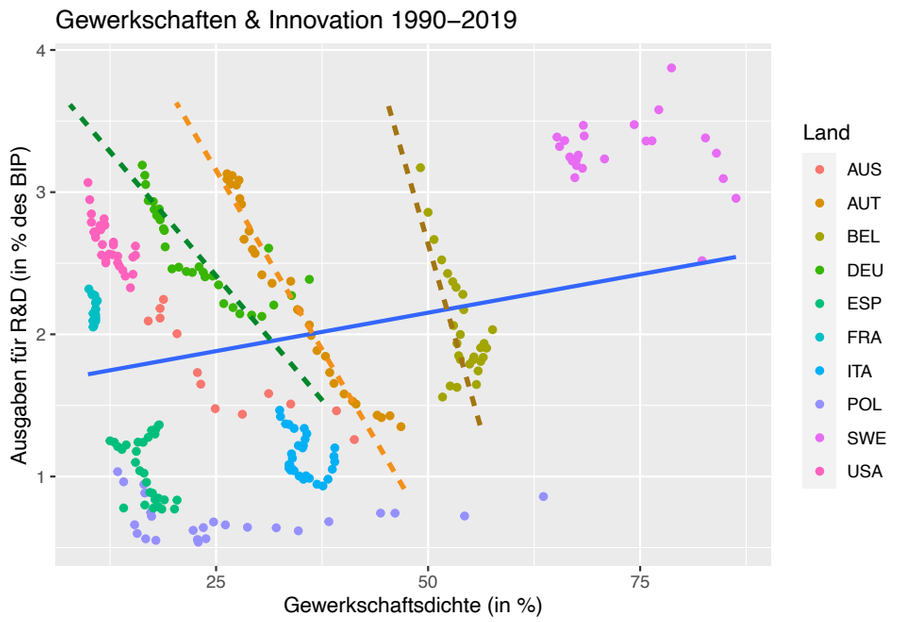


Tukey-Anscombe Plot: Unser Einführungsmodell



Scatterplot: Unser Einführungsmodell

Tukey-Anscombe Plot: „white noise“



Scatterplot mit größerer Informationsdichte: Länder

Typischer Fall:
Zwischen Ländern und Innerhalb von Ländern beobachten wir unterschiedliche Dynamiken

Mögliche Kritikpunkte an unserem Modell

$$Inno_i = \beta_0 + \beta_1 UnionDensity_i + \epsilon_i$$

• Positive Korrelation würde auch entstehen, wenn *Inno* positiv auf *UD* wirkt oder es eine Wechselwirkung gibt.	• „Simultanität“ Termin 10/11
• Wir ignorieren, dass es in unserem Datensatz eine räumliche und eine zeitliche Dimension gibt.	• Paneldaten: Zweite Veranstaltung im Modul
• Fehlende Daten als Problem? (auch: Daten valide, reliabel…)	• „Selection Bias“ Termin 10/11
• Positive Korrelation würde auch entstehen, wenn beide von einem dritten Faktor beeinflusst werden.	• Drittvariablenproblem Termin 10/11
• *Inno* könnte auch vielen anderen Faktoren beeinflusst werden und *UD* ist damit gar nicht so wichtig bzw. muss in einem größeren Kontext verstanden werden?	• Können wir gleich hier mal angehen…

```
> head(unioninno, 3)
   Country Year UnionDensity     Tech
1:     AUS 1990         41.3 1.259877
2:     AUS 1991         42.0       NA
3:     AUS 1992         39.2 1.462098
```

Bivariate und multivariate Regression

- Regressionen können mehrere erklärende Variablen enthalten, z.B.

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$$

- Vermutung: das durchschnittliche Einkommen (*GDPpc*) und die Größe des öffentlichen Sektors (*TAX*) haben Einfluss auf die Innovationsausgaben *Inno*.
  - Dann könnten wir beispielsweise das folgende Modell schätzen.

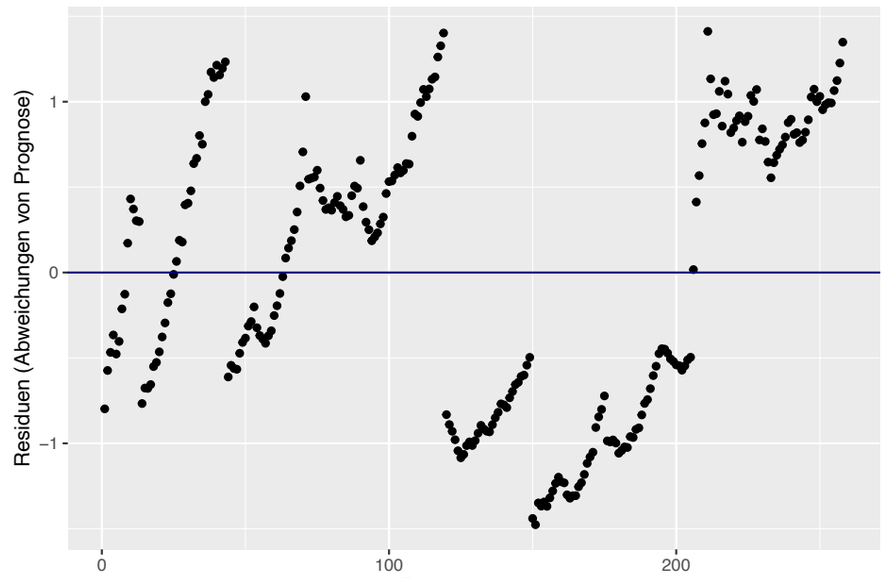
```
> techmodel2 <- lm(Tech ~ UnionDensity + GDPpc + TAX, data=oecd_data)
> summary(techmodel2)

Call:
lm(formula = Tech ~ UnionDensity + GDPpc + TAX, data = oecd_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81954 -0.34235 -0.03506  0.29677  1.16416

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.240e-01  1.186e-01   2.732  0.00674 **
UnionDensity 1.394e-02  1.870e-03   7.458 1.40e-12 ***
GDPpc        5.384e-05  2.227e-06  24.180  < 2e-16 ***
TAX         -3.505e-02  5.766e-03  -6.078 4.45e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4282 on 253 degrees of freedom
  (43 observations deleted due to missingness)
Multiple R-squared:  0.7392,    Adjusted R-squared:  0.7362
F-statistic: 239.1 on 3 and 253 DF,  p-value: < 2.2e-16
```

$$Tech_i = \beta_0 + \beta_1 UnionDensity_i + \beta_2 GDPpc_i + \beta_3 TAX_i + \epsilon_i$$

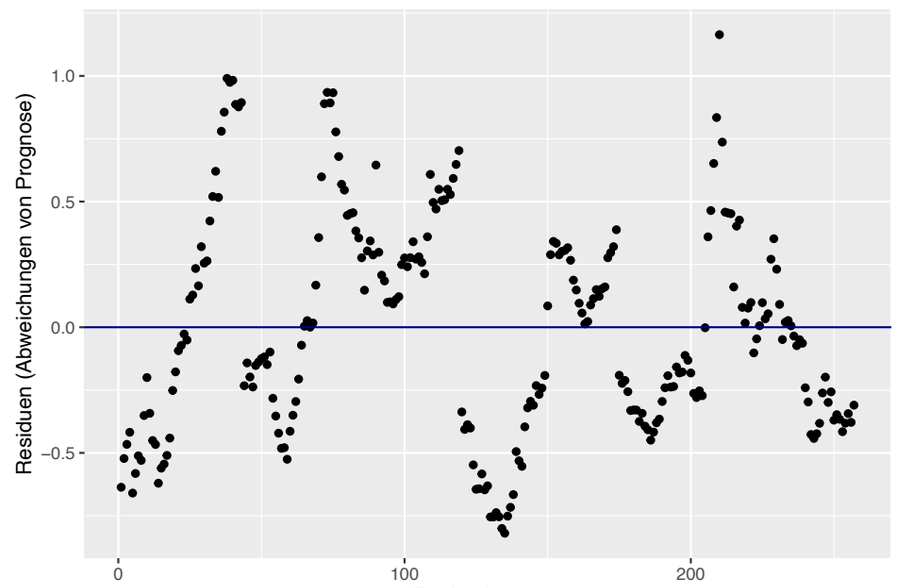
- Koeffizienten des Modells: „Konditionale Effekte“
  - Hier gibt β_1 den Effekt von *UnionDensity* an, wenn man die anderen beiden Variablen konstant hält.
  - „Wahrscheinlichkeitstheoriesprache“: konditionale Effekte
  - „Wissenschaftstheoriesprache“: ceteris paribus Effekte
  - „Ökonometrie-Sprech“: Wir kontrollieren für …

Wiederum: Residuenanalyse

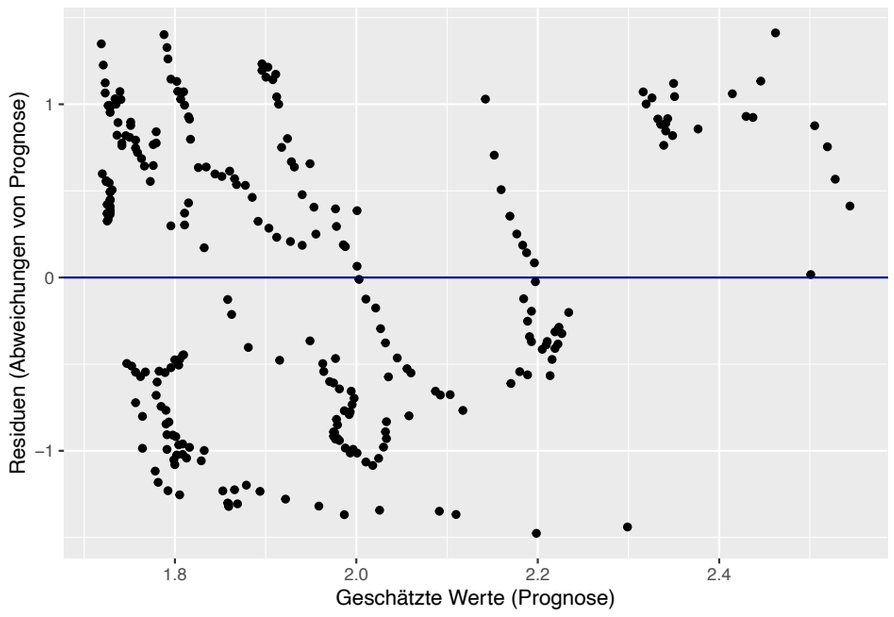
„Struktur in den Residuen“ vs. „Weißes Rauschen“



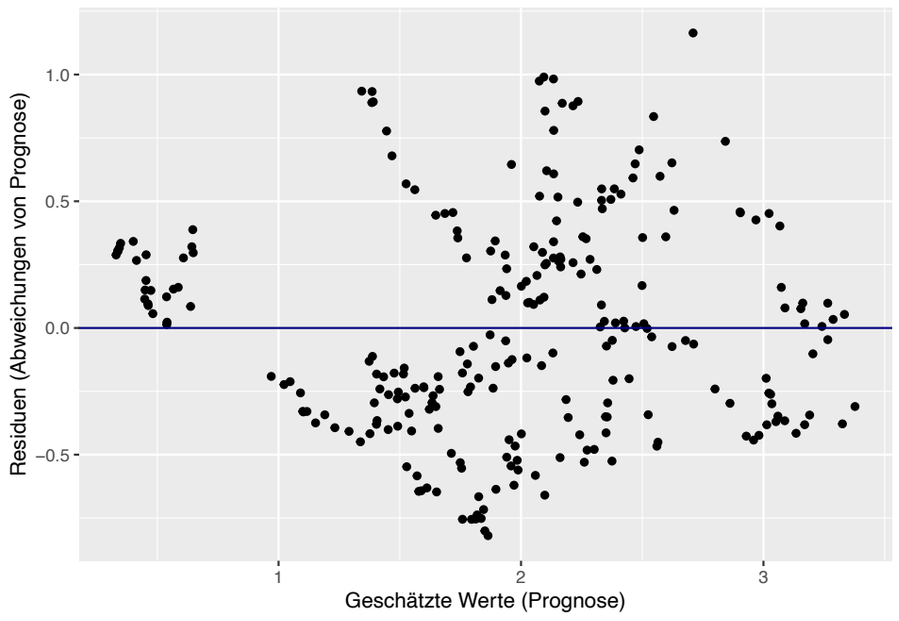
Residuenplot: Unser Einführungsmodell



Residuenplot: Erweitertes Modell



Tukey-Anscombe Plot: Unser Einführungsmodell



Tukey-Anscombe Plot: Erweitertes Modell

Regression:
Grundbegriffe und formale Aspekte

Grundlegende Begriffe

- Ökonometrie ~ auf die Ökonomik angewandte Stochastik
  - Für die Ökonometrie ist sowohl Wahrscheinlichkeitstheorie als auch Statistik wichtig.

- Die Wahrscheinlichkeitstheorie befasst sich mit der mathematischen Beschreibung von Zufallsprozessen
- Wir formulieren wahrscheinlichkeitstheoretische Modelle und untersuchen welche Realisationen aus den angenommenen Mechanismen resultieren
- **Beispiel:** Wir definieren ein Modell eines Würfelwurfes indem wir einen Ereignisraum $\Omega = \{1,2,3,4,5,6\}$ festlegen, sowie entsprechende Wahrscheinlichkeiten definieren, nämlich $\mathbb{P}(\omega = 1) = \mathbb{P}(\omega = 2) = \ldots = \mathbb{P}(\omega = 6) = \frac{1}{6}$. Aus diesem Modell können wir dann die Verteilung der gewürfelten Augen ableiten.

Grundlegende Begriffe

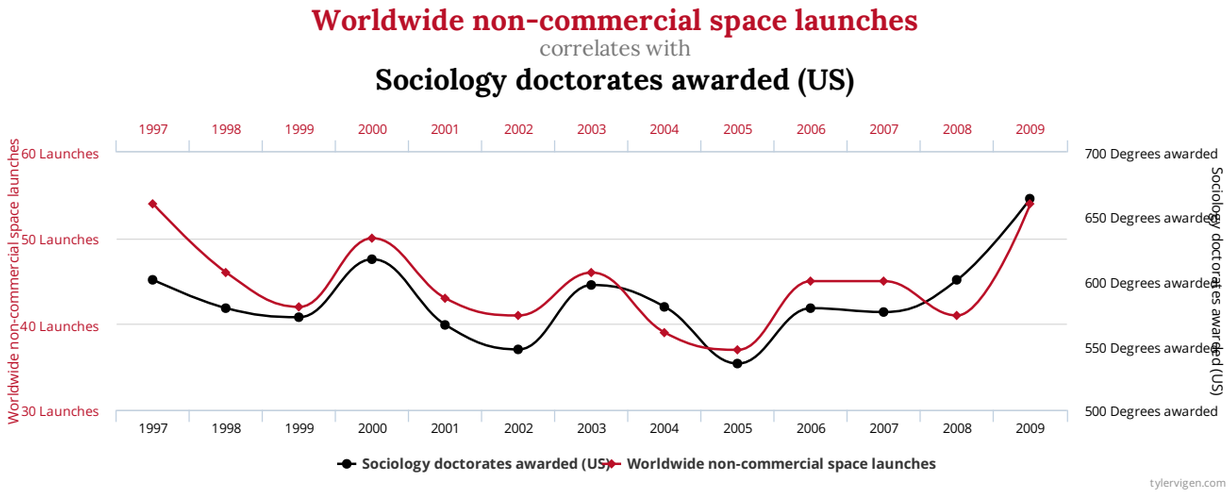
- Ökonometrie ~ auf die Ökonomik angewandte Stochastik
  - Für die Ökonometrie ist sowohl Wahrscheinlichkeitstheorie als auch Statistik wichtig
  - Die Wahrscheinlichkeitstheorie befasst sich mit der mathematischen Beschreibung von Zufallsprozessen.

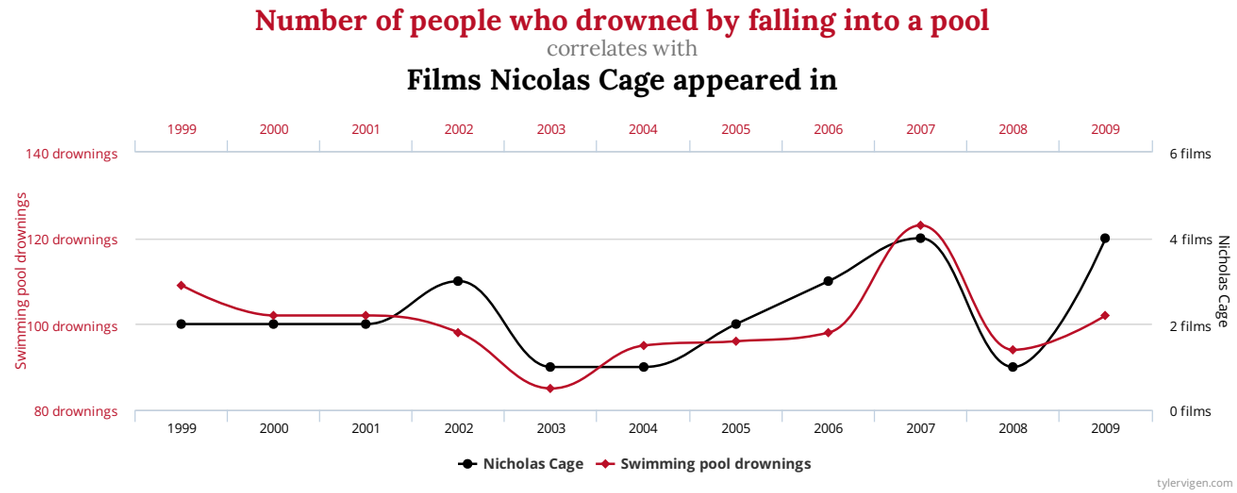
- In der Statistik ist es genau andersherum:
  - Wir beobachten Daten als Realisationen von Zufallsprozessen und versuchen Rückschlüsse auf die zugrundeliegenden Prozesse zu ziehen, bzw. entscheiden ob unsere bisherigen Vorstellungen über den zugrundeliegenden Prozess plausibel sind.
  - **Beispiel:** Wir werfen einen Würfel sehr häufig und schauen uns die Verteilung der gewürfelten Augen an. Auf dieser Basis entscheiden wir, ob ein Modell, in dem alle Seiten mit der Wahrscheinlichkeit $\frac{1}{6}$ auftauchen plausibel ist.

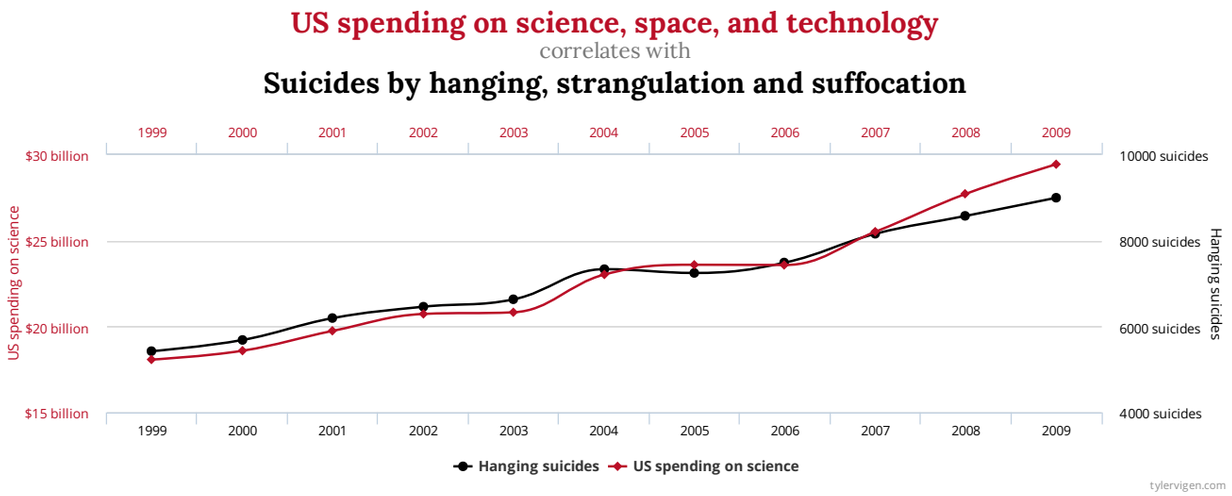
Grundlegende Begriffe

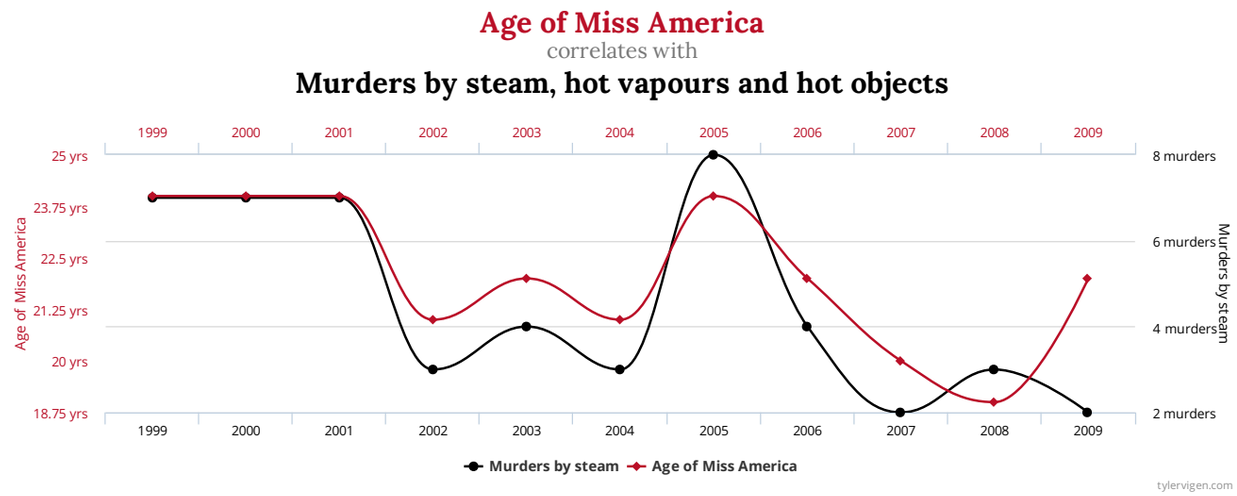
- Regression als parametrische Schätzung
  - Kovarianz und Korrelationskoeffizient als Beispiele für nicht-parametrische Verfahren.
- Statistische Denkschulen: frequentistische und Bayesianische Statistik
  - Hier: Fokus auf frequentistische Statistik.
  - Grundidee: Fokus auf möglichst ‚objektives' Maß von Wahrscheinlichkeit (‚Frequentismus') vs. Fokus auf wie neue Daten subjektive Einschätzungen über Wahrscheinlichkeit rationalerweise ändern sollte (‚Bayesianismus').
- Korrelation ≠ Kausalität
  - Korrelation: Bewegung der Variablen X und Y verläuft teilweise parallel.
  - Kausalität: X hat tatsächlich eine kausale Wirkung auf Y.
  - Scheinkorrelationen als allgegenwärtiges Phänomen…
  - Signifikanz beweist keine Kausalität. Sie ist definiert als Wahrscheinlichkeit der Daten gegeben die Hypothese, dass kein Zusammenhang vorliegt (Nullhypothese): $p(D \mid H_0)$

Unterhaltsame Korrelationen zur Auflockerung









Quelle: https://www.tylervigen.com/spurious-correlations

Noch ein paar Grundbegriffe

Stichprobe vs. Population

- Stichprobe vs. Population als zentrale Unterscheidung
  - Grundidee: Wir entnehmen Stichproben (z.B. Befragung, Gewichtsmessung) aus Populationen (z.B. alle Personen in Deutschland).
  - In der deskriptiven Statistik werden Größen wie die Varianz für eine Population und eine Stichprobe unterschiedlich berechnet.
  - In der schließenden Statistik geht es darum von einer Stichprobe Rückschlüsse auf die Population zu ziehen
- „Parameter“ & „Statistik/Schätzer“, „Fehlerterme“ (ϵ) & „Residuen“ (e)
  - Streng genommen: Die Begriffe „Parameter“ und „Fehlerterme“ beziehen sich auf die Population, die wir aber nie kennen. Unser „Schätzer“ bzw. die „Residuen“ ergeben sich hingegen immer aus einer Stichprobe.
  - Wichtiger Hinweis: Manchmal ist in der Ökonometrie nicht so klar, was durch die Stichprobe repräsentiert werden soll/die Population ausmacht (insbesondere bei makroökonomischen Fragen).

Ein letztes Mal: Population vs. Stichprobe

- „Parameter“ & „Statistik/Schätzer“, „Fehlerterme“ & „Residuen“
  - Streng genommen: Die Begriffe „Parameter“ und „Fehlerterme“ beziehen sich auf die Population, die wir aber nie kennen. Unser „Schätzer“ bzw. die „Residuen“ ergeben sich hingegen immer aus einer Stichprobe.
  - Wir hoffen mit unserem Schätzer dem wahren Parameter nahe zu kommen: Erwartungstreue (Schätzer ist im Mittel richtig, zentraler Grenzwertsatz gilt), Effizienz (Schätzer hat die geringstmögliche Varianz) und Konsistenz (Schätzer wird wachsendem n genauer).
- Für die Fehlerterme treffen wir eine Annahme: $E(\epsilon_i \mid X) = 0$
  - Das bedeutet, dass die erklärenden Variablen keine Information über die Fehlerterme enthalten („Exogenität“) – diese sind im Mittel gleich Null.
  - Für die Fehlerterme müssen wir dies einfach annehmen und können es nicht nachweisen.
  - Für die Residuen gilt aufgrund der Mechanik der Schätzung immer, dass erklärenden Variablen und Residuen unkorreliert sind.

Exkurs: Was ist Erwartungstreue?

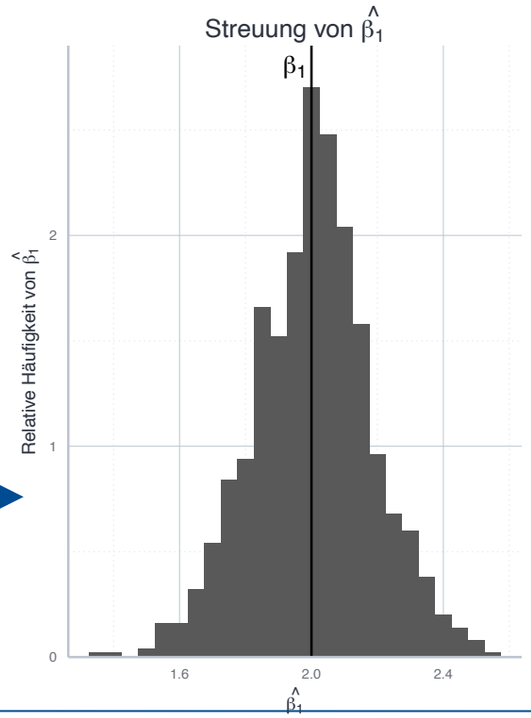
Erläuterung mit Hilfe einer Monte-Carlo Simulation

- **Erwartungstreue:** Schätzer ist im Mittel richtig, zentraler Grenzwertsatz gilt…
  - **Genauer:** Unser Schätzer β ist eine Zufallsvariable (jede Stichprobenziehung ist ein Würfelwurf) und folgt der Verteilung $\hat{\beta}_1 \propto \mathcal{N}\left(\beta_1, SE_{\beta_1}\right)$
  - Informelle Definition **„Monte Carlo Simulation"**: Wir lassen einen Zufallsprozess mehrfach laufen und gucken was passiert…

$$Y_i = 3 + 2x_i + \epsilon_i, \quad \epsilon_i \propto \mathcal{N}(0,5)$$

```r
set.seed(123)
true_DGP <- function(x, b0, b1){
  y <- b0 + b1*x + rnorm(length(x), 0, 5)
  return(y)
}
beta_0_wahr <- 3
beta_1_wahr <- 2
sample_size <- 100
x <- runif(sample_size, 0, 10)
set.seed(123)
n_datensaetze <- 1000
beta_0_estimates <- rep(NA, n_datensaetze)
beta_1_estimates <- rep(NA, n_datensaetze)

for (i in 1:n_datensaetze){
  daten_satz <- data.frame(
    x = x,
    y = true_DGP(x, beta_0_wahr, beta_1_wahr)
  )
  schaetzung_2 <- lm(y~x, data = daten_satz)
  beta_0_estimates[i] <- schaetzung_2[["coefficients"]][1]
  beta_1_estimates[i] <- schaetzung_2[["coefficients"]][2]
}
```

- Wir erstellen 1000 Ziehungen dieses Prozesses
  - Da die Y über die Fehler zufällig sind, sieht der Datensatz jedes Mal anders aus
  - Wir sehen im Mittel trifft der Schätzer den wahren Parameter: Er ist **erwartungstreu**.



Zentrale Elemente des Regressionsoutputs

Formale Aspekte der „Methode der kleinsten Quadrate“

- In der bivariaten Regression haben hat der zentrale Koeffizient des Regressionoutputs, β_1, eine intuitive Interpretation.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad Inno_i = \beta_0 + \beta_1 UnionDensity_i + \epsilon_i$$

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sum (Y - \bar{y})(X - \bar{x})}{\sum (X - \bar{x})^2} \qquad tstat_1 = \frac{\hat{\beta}_1}{SE_{\beta_1}}$$

(2) Geschätzte Werte

(4) t-Statistik und Signifikanzwert

```
             Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
β̂0 (Intercept)                1.61192    0.09779   16.48  < 2e-16 ***
β̂1 UnionDensity               0.01080    0.00268    4.03 7.35e-05 ***
```

(1) Bezeichnung der „Schätzer“

(3) Standardfehler (von β_1)

$$SE_{\beta_1} = \sqrt[2]{\frac{Var(\epsilon)}{Var(X) \cdot n}} = \sqrt[2]{\frac{\frac{1}{n-2}\sum_{i=1}^{n} e^2}{\sum (X - \bar{x})^2}}$$

Varianz (Standardfehler) der Residuen

Mechanik des Regressionsverfahrens

Formale Aspekte der „Methode der kleinsten Quadrate“

- Methode der kleinsten Quadrate = *ordinary least squares (OLS)*
  - **Grundidee:** Wir suchen jene Linie bei der die quadrierten Abweichungen der Datenpunkte minimal sind.
  - Was heißt das mathematisch?

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$Y_i - \beta_0 - \beta_1 X_i = \epsilon_i$$

$$(Y_i - \beta_0 - \beta_1 X_i)^2 = \epsilon_i^2$$

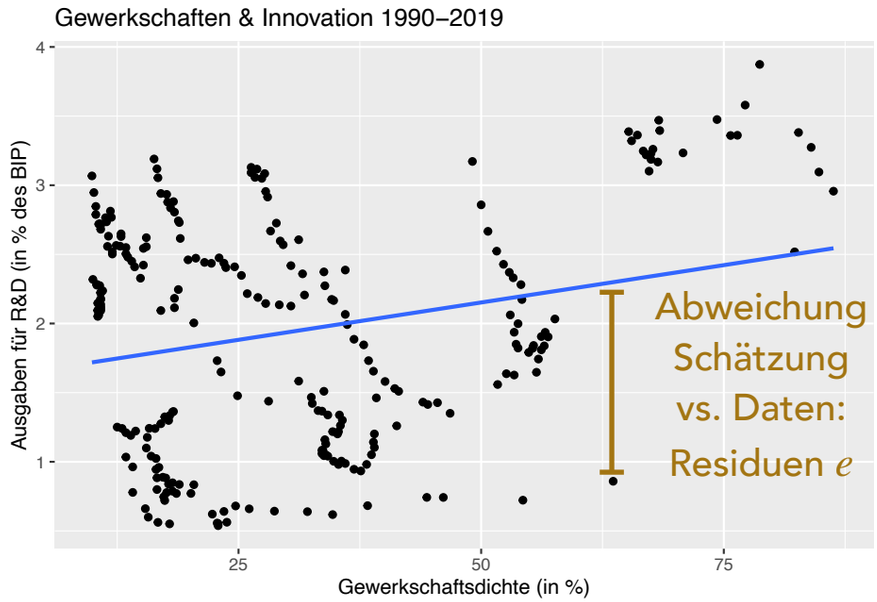
$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

Die Fehler sollen minimal sein.

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

führt zu…

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$



$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Das kleine Dach („*hat*“) drückt aus, dass es sich um eine Schätzung und nicht um den wahren Parameter handelt.

- Wir suchen jene Werte für β_0 und β_1 für die Fehler minimal klein sind.

Das Ausmaß erklärter Variation

Formale Aspekte der „Methode der kleinsten Quadrate“

- R^2 als Maß für Anteil der „erklärten“ Variation $R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$
  - Intuitiv: Welchen Anteil der Schwankung in Y erklärt X?
  - Formal: $TSS = ESS + RSS$ („SS“ steht für *sum of squares: total, explained and residual).*
  - TSS und RSS kann man einfach als Summe der quadratischen Abweichung ausdrücken: $TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ und $ESS = \sum_{i=1}^{n} (e_i)^2$. Der zweite Ausdruck ist einfacher, weil $\bar{e} = 0$.

- R^2 als sehr „grobes“ Gütekriterium
  - Tauglichkeit hängt ab von Fragestellung
  - Simples R^2 steigt immer mit zusätzlichen Variablen – „angepasstes“ R^2 als Alternative: $\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e^2/(n-k-1)}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2/(n-1)}$

Freiheitsgrade! (Grob: Zahl der Beobachtungen – Zahl der errechneten Variablen)
k = Zahl der erklärenden Variablen im Modell, (-1) aufgrund des Mittelwerts

Zur statistischen Signifikanz

Formale Aspekte der „Methode der kleinsten Quadrate“

- Signifikanz beweist keine Kausalität.
  - Sie ist definiert als Wahrscheinlichkeit der Daten gegeben die Hypothese, dass kein Zusammenhang vorliegt (Nullhypothese): $p(D \mid H_0)$
- Schrittweise Betrachtung
  - Die Nullhypothese H_0 besagt, dass zwischen Y und X kein Zusammenhang besteht ($\beta_1 = 0$).
  - Wenn wir H_0 verwerfen können, dann ist der Effekt *signifikant.*
  - Die Wahrscheinlichkeit, unser $\hat{\beta}_1$ unter der H_0 zu beobachten wir durch den p-Wert, also $p(D \mid H_0)$, angegeben. Reagiert stark auf Größe der Stichprobe (n).
  - Wenn $p < \alpha$ ist der Schätzer signifikant auf dem α-Niveau (=gewünschtes Signifikanzniveau).
  - Wieder gilt: Hohe p-Werte sind geeignet, skeptische Fragen zu stellen, niedrige p-Werte verifizieren aber keine Hypothese (Falsifikation…)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.240e-01  1.186e-01   2.732  0.00674 **
UnionDensity 1.394e-02  1.870e-03   7.458 1.40e-12 ***
GDPpc        5.384e-05  2.227e-06  24.180  < 2e-16 ***
TAX         -3.505e-02  5.766e-03  -6.078 4.45e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regressionsanalyse: Erste Erweiterungen

Erweiterungen des Regressionsmodells

Erweiterung #1: Dummy-Variablen

- Wir überlegen was wir mit dem Regressionsmodell alles anstellen können…

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$$

- Dummy-Variable
  - Wir können nominale (ordinale) Daten mit einem Vektor darstellen, dessen Elemente aus 0 und 1 bestehen („ist nicht der Fall“ vs. „ist der Fall“), z.B. Mann-Frau, Teil von Europa,…
  - In unserem Beispiel: „Liberale Marktwirtschaften“ eher innovativ.
  - USA und AUS als „liberale Marktwirtschaften“ in unserem Datensatz.
  - Kreation Dummy-Var:

```
oecd_data <- dplyr::mutate(oecd_data, "LME"=dplyr::if_else(Country %in% c("AUS","USA"),1,0))
```

$$Inno_i = \beta_0 + \beta_1 UnionDensity_{1,i} + \beta_2 GDPpc_{2,i} + \beta_3 TAX_{3,i} + \beta_4 LME + \epsilon_i$$

Typischer Praxisfall:
Wir müssen uns die Variable, die uns interessiert selbst konstruieren und unserem Modell hinzufügen.

```
> techmodel3 <- lm(Tech ~ UnionDensity+GDPpc+TAX+LME, data=oecd_data)
> summary(techmodel3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.201e-01  1.101e-01   2.907  0.00398 **
UnionDensity 1.046e-02  1.819e-03   5.748 2.60e-08 ***
GDPpc        6.312e-05  2.521e-06  25.040  < 2e-16 ***
TAX         -4.102e-02  5.434e-03  -7.549 8.03e-13 ***
LME         -5.668e-01  8.807e-02  -6.435 6.17e-10 ***
```

Erweiterungen des Regressionsmodells

Erweiterung #2: Funktionale Formen

- Wir überlegen was wir mit dem Regressionsmodell alles anstellen können…

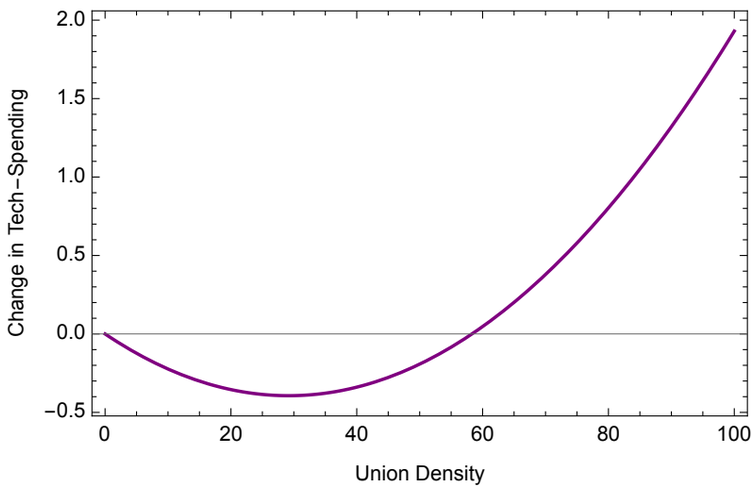
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$$

- Alternative funktionale Formen
  - Nur die Parameter sind notwendig linear: die Variablen können transformiert werden und auch mehrfach verwendet werden.

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{1,i}^2 + \beta_3 X_{2,i} + \epsilon_i$$

  - Blaue Gleichung zeigt das Beispiel einer quadratischen Funktion.
  - Was wäre, wenn es ein optimales Gewerkschaftsniveau gibt, unter- und oberhalb dessen die Innovationsausgaben abnehmen?

$$Inno_i = \beta_0 + \beta_1 UnionDensity_{1,i} + \beta_2 GDPpc_{2,i} + \beta_3 TAX_{3,i} + \beta_4 UnionDensity^2 + \epsilon_i$$



Wie so oft, findet man das Gegenteil des Erwarteten:
Full Unionization is optimal ;-)

```
> techmodel4 <- lm(Tech ~ UnionDensity+I(UnionDensity^2)+GDPpc+TAX, data=oecd_data)
> summary(techmodel4)

                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.661e-01  1.310e-01   5.848 1.54e-08 ***
UnionDensity      -2.659e-02  6.686e-03  -3.977 9.12e-05 ***
I(UnionDensity^2)  4.633e-04  7.378e-05   6.280 1.48e-09 ***
GDPpc              5.079e-05  2.131e-06  23.830  < 2e-16 ***
TAX               -1.862e-02  5.976e-03  -3.115  0.00205 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Erweiterungen des Regressionsmodells

Erweiterung #3: Verwendung des Logarithmus

- Wir überlegen was wir mit dem Regressionsmodell alles anstellen können…

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$$

- Logarithmus als hilfreiche Transformation
  - Kann geeignet sein um Daten, die nicht linear sind, zu „linearisieren“ → T7 / T10-11
  - Kann Interpretation erleichtern: z.B. *UnionDensity* und *Inno* beides in Prozent.
  - Angenommen wir möchten den Wert fürs GDPpc auch so interpretieren. Dann können wir schreiben: $Inno_i = \beta_0 + \beta_1 UnionDensity_{1,i} + \beta_2 Ln(GDPpc_{2,i}) + \beta_3 TAX_{3,i} + \epsilon_i$

Interpretation:
Wenn das GDP um 1% steigt,
dann steigen die Tech-Ausgaben um 0.015%-Punkte

```
> techmodel5 <- lm(Tech ~ UnionDensity+log(GDPpc)+TAX, data=oecd_data)
> summary(techmodel5)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.722301   0.647821 -15.008  < 2e-16 ***
UnionDensity  0.015702   0.002198   7.144 9.65e-12 ***
log(GDPpc)    1.154740   0.061864  18.666  < 2e-16 ***
TAX          -0.040321   0.006788  -5.940 9.39e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Erweiterungen des Regressionsmodells

Erweiterung #3: Verwendung des Logarithmus

- Wir überlegen was wir mit dem Regressionsmodell alles anstellen können…

- Logarithmus als hilfreiche Transformation
  - Kann geeignet sein um Daten, die nicht linear sind, zu „linearisieren“ → T7 / T10-11
  - Interpretationshilfe für Logarithmen in Regressionsgleichungen

Modellart	**Schätzgleichung**	**Interpretation der Koeffizienten**
Level-Level	$y = \beta_0 + \beta_1 x_1 + \epsilon$	Ändert sich x_1 um 1 ändert sich y um β_1
Log-Level	$\ln(y) = \beta_0 + \beta_1 x_1 + \epsilon$	Ändert sich x_1 um 1 ändert sich y c.p. um ca. $100 \cdot \beta_1 \%$
Level-Log	$y = \beta_0 + \beta_1 \ln(x_1) + \epsilon$	Ändert sich x_1 um ca. 1% ändert sich y c.p. um ca. $\beta_1/100$
Log-Log	$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \epsilon$	Ändert sich x_1 um ca. 1% ändert sich y c.p. um ca. $\beta_1 \%$

Erweiterungen des Regressionsmodells

Erweiterung #4: Interaktionsterme

- Wir überlegen was wir mit dem Regressionsmodell alles anstellen können…

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$$

- **Wechselwirkungen zwischen erklärenden Variablen: Interaktionsterme**
  - Wir können eine Regressionsgleichung auch so schreiben: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} \cdot X_{2,i} + \epsilon_i$
  - Dabei werden zwei Variablen multipliziert. Wenn dieser Termin signifikant ist, hängt die Wirkung einer Variable von der Größe der jeweils anderen ab (z.B. Wirkung eines Medikaments hängt vom Geschlecht ab).
  - **Ein einfaches Beispiel zu US-Präsidentschaftswahlen mit Dummy-Variablen:**

$$Donation = \beta_0 + \beta_1 GenderCandidate + \beta_2 GenderDonor + \beta_3 GenderCandidate * GenderDonor$$

  - Man nehme an, dass (a) Männer mehr spenden, (b) Männer mehr Spenden bekommen, (c) Frauen anderen Frauen mehr spenden als Männer anderen Männern.
  - Also würden wir folgendes erwarten: $\beta_1 < 0,\ \beta_2 < 0,\ \beta_3 > (\beta_1 + \beta_2)(-1)$

    (wenn Frau=1, also Frauen durchgehen mit 1 und Männer mit 0 codiert werden)

Erweiterungen des Regressionsmodells

Erweiterung #4: Interaktionsterme

- Wir überlegen was wir mit dem Regressionsmodell alles anstellen können…

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$$

- Wechselwirkungen zwischen erklärenden Variablen: Interaktionsterme
  - Wir können eine Regressionsgleichung auch so schreiben: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} \cdot X_{2,i} + \epsilon_i$
  - Angewendet auf unser Beispiel könnte man beispielsweise die Vermutung aufstellen, dass die Wirkung von mehr Gewerkschaften vom allgemeinen Entwicklungsstand der Ökonomie abhängen könnte:

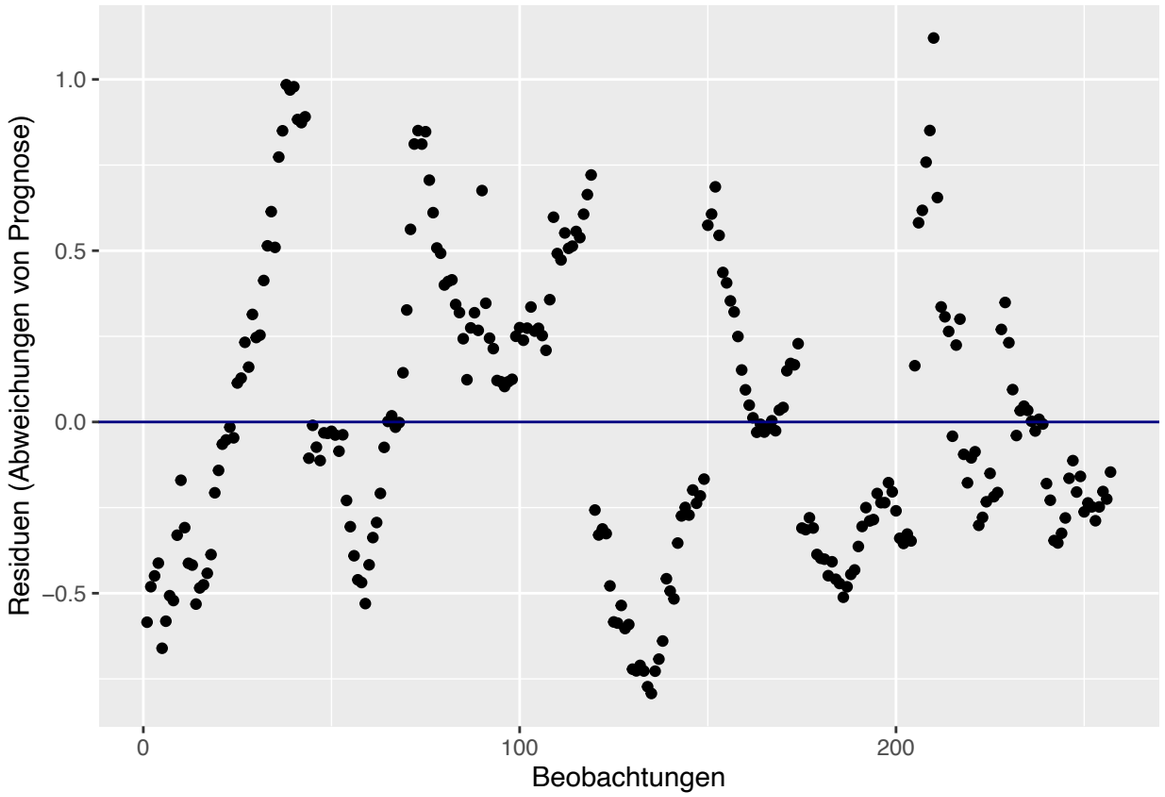
$$Inno_i = \beta_0 + \beta_1 UnionDensity_i + \beta_2 GDPpc_i + \beta_3 TAX_i + \beta_4 UnionDensity_i \cdot GDPpc_i + \epsilon_i$$

```
> techmodel6 <- lm(Tech ~ UnionDensity+GDPpc+TAX+UnionDensity*GDPpc, data=oecd_data)
> summary(techmodel6)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.218e-01  1.833e-01   4.483 1.12e-05 ***
UnionDensity        -5.549e-03  5.850e-03  -0.948 0.343806
GDPpc                4.125e-05  4.198e-06   9.826  < 2e-16 ***
TAX                 -3.480e-02  5.642e-03  -6.167 2.75e-09 ***
UnionDensity:GDPpc   4.876e-07  1.390e-07   3.508 0.000535 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

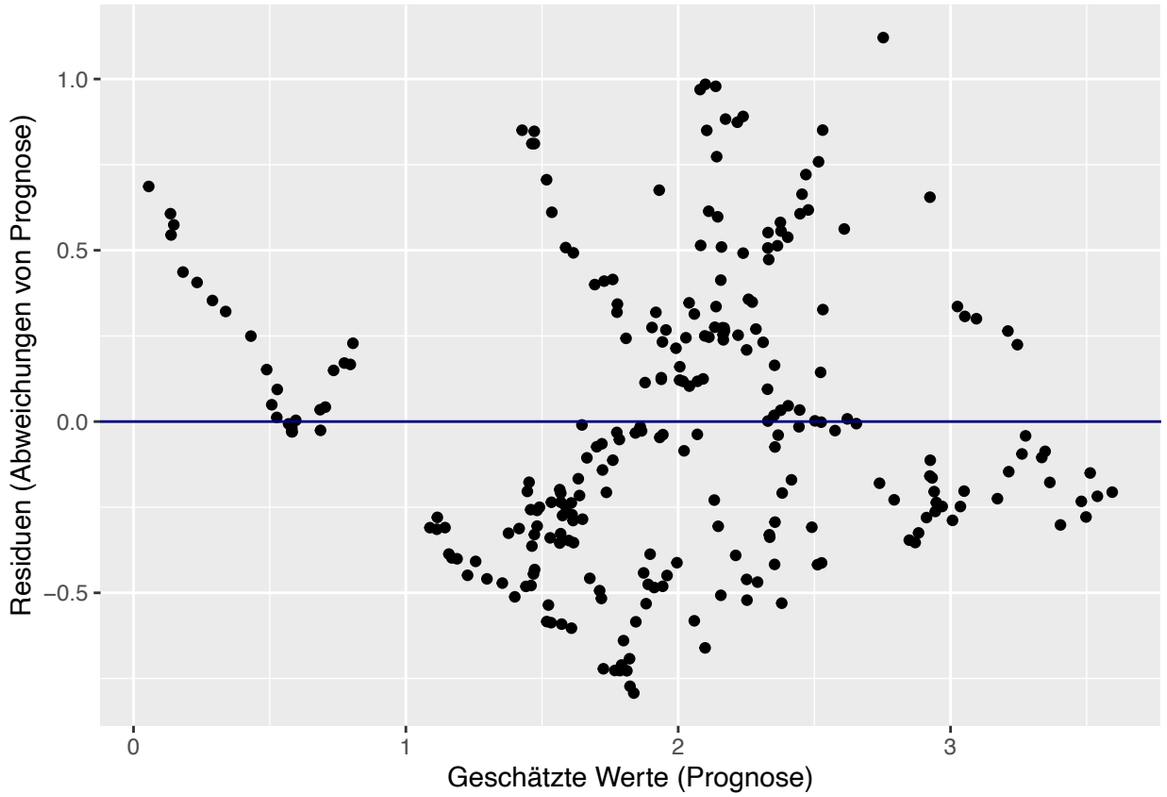
Interpretation:
Die positiven Effekte von Gewerkschaften auf die Innovationsausgaben zeigen sich eher in reichen Ökonomien

Passt zur theoretischen Idee, dass höherer Lohndruck Innovationsanreize setzt (Kaldor-Verdoorn).
Kompatibel mit bestehenden Resultaten (z.B. von Kleinknecht et al.)

Residuenplots zum finalen Modell



Residuenplot: Interaktionsmodell



Tukey-Anscombe Plot: Interaktionsmodell

Anstatt eines Resümees: That was fun!

Aber: alles ruht auf Annahmen…

- Zentrale Einsicht: Das „richtige Modell“ ist nicht leicht zu finden.
  - Muss zum Prozess passen, der die Daten generiert.
  - Unser letztes Modell beschreibt die Struktur der Daten nur unzureichend.
  - Kein statistischer Test gibt definitive Auskunft darüber, ob ein Modell „richtig“ ist.

- Regressionsverfahren ruht auf Annahmen

  – Bezug zur Frage des „richtigen Modells“
  - A1: „Linearität“ – Das Modell ist linear in seinen Parameters.
  - A2: „Exogenität“– Die unabhängigen Variablen sind exogen
  - A3: „Multikollinearität“– Die unabhängigen Variablen sind nicht linear abhängig
  - A4: „Homoskedastie und keine Autokorrelation“– Die Fehler haben konstante Varianz und sind nicht miteinander korreliert

Wiederholungsfragen zum Selbststudium

Wiederholungsfragen

- Was ist der zentrale Unterschied zwischen der Berechnung des Korrelationskoeffizienten um der Bestimmung eines Regressionsparameters?

- Was verstehen wir unter einer abhängigen und einer unabhängigen Variable?

- Fassen Sie kurz den idealtypischen Ablauf ökonometrischer Forschung zusammen.

- Malen Sie sich einen Datensatz mit 5-10 Beobachtungen für eine abhängige und eine unabhängige Variable auf und zeichnen Sie grob eine Regressionsgrade ein und beachten dabei die Intuition des OLS Schätzers.

Wiederholungsfragen

- Was haben Wahrscheinlichkeitstheorie und Statistik mit der Ökonometrie zu tun?
- IGeben Sie jeweils zwei Beispiele für Korrelation und Kausalität und erläutern Sie den Unterschied.
- Grenzen Sie die Konzepte TSS, RSS und ESS voneinander ab.
- Was ist der Unterschied zwischen den Fehlertermen und den Residuen?
- Was hat es mit der *Verteilung eines Schätzers* auf sich?
- Was bedeutet es wenn wir sagen, dass ein geschätzter Wert ‘signifikant’ ist?
- Erläutern Sie den Unterschied zwischen dem adjustierten und nicht adjustierten Bestimmtheitsmaß. Welches der beiden ist besser?

Wiederholungsfragen

- Was ist eine Dummy-Variable wozu kann eine solche in ökonometrischen Modellen hilfreich sein?

- Geben Sie Beispiele für die Flexibilität und Grenzen des Regressionsverfahrens an, indem Sie aufzeigen welche grundsätzlichen funktionalen Formen mit diesem kompatibel sind?