Proseminar Wissenschaftliches Arbeiten: Ökonometrie und Statistik

Woche 4: Themenvorstellung – Statistisches Lernen und Regressionsprobleme

Elias Wolf

23. April 2024

Einführende Literatur (kostenlos per VPN)

- Angrist, J. D., & Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
  Link: `https://www.degruyter.com/view/title/563369`
- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2018). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung (15. Auflage.)*. Springer.
  Link: `https://www.springer.com/de/book/9783662566541`
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. Springer.
  Link: `https://www.springer.com/de/book/9780387848570`
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.
  Link: `https://www.springer.com/de/book/9781461471370`
- Neusser, K. (2016). *Time Series Econometrics.* Springer.
  Link: `https://www.springer.com/de/book/9783642334351`
- Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics (Fourth edition)*. Pearson.
  Link: `https://ebookcentral.proquest.com/lib/ulb-bonn/detail.action?docID=5834470`

Statistisches Lernen

- Y: Abhängige/Endogene/Output Variable
- $(X_1, \ldots, X_k)$: Unabhängige/Exogene/Input Variablen
- Wir möchten den Zusammenhang von $X_1, \ldots, X_k$ und Y verstehen.
- Allgemeines Modell:
$$Y = f(X_1, \ldots, X_k) + u$$
- Die Regressionsfunktion $f(X_1, \ldots, X_k)$ beschreibt die in $(X_1, \ldots, X_k)$ enthaltenen systematischen Informationen über Y.
- u ist ein von $(X_1, \ldots, X_k)$ unabhängiger Fehlerterm mit $E[u] = 0$.

- $f(X_1, \ldots, X_k)$ ist unbekannt.
- Wie können wir $f(X_1, \ldots, X_k)$ bestimmen bzw. schätzen?
- Beispielsweise mithilfe einer linearen parametrischen Struktur (lineares Regressionsmodell):
$$f(X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Beispiel: Einfaches lineares Regressionsmodell

- Lineares Regressionsmodell mit einer unabhängigen Variable X:

$$Y = \beta_0 + \beta_1 X + u$$

- Beobachtungen (Y_t, X_t) für $t = 1, \ldots, T$
- Die Parameter β_0 und β_1 können wir mit der Kleinste Quadrate Methode (KQ) bzw. Ordinary Least Squares (OLS) schätzen:

$$\widehat{\beta}_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}, \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X},$$ =argmin Sum ui^2

wobei

$$\overline{X} = \frac{1}{T} \sum_{t=1}^{T} X_t, \quad \overline{X^2} = \frac{1}{T} \sum_{t=1}^{T} X_t^2, \quad \overline{XY} = \frac{1}{T} \sum_{t=1}^{T} X_t Y_t$$

- Geschätztes Modell:

$$\widehat{f}_{OLS}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

Beispiel: Multivariates lineares Regressionsmodell

- Multivariates Regressionsmodell: $f(X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$

$$Y = \boldsymbol{\beta}' \boldsymbol{x} + u, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)', \quad \boldsymbol{x} = (1, X_1, \ldots, X_k)'$$

- Beobachtungen $(Y_t, X_{1,t}, \ldots, X_{k,t})$ für $t = 1, \ldots, T$
- Den Parametervektor $\boldsymbol{\beta}$ schätzen wir mit OLS:

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_k)' = \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t' \right)^{-1} \sum_{t=1}^{T} \boldsymbol{x}_t y_t,$$

wobei $\boldsymbol{x}_t = (1, X_{1,t}, \ldots, X_{k,t})'$.

betahut=u prime u

- Alternativ in Matrixschreibweise:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{y},$$

wobei $\boldsymbol{y} = (Y_1, \ldots, Y_T)'$ und $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]'$.

- Geschätztes Modell:

$$\widehat{f}_{OLS}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_k X_k$$

Mögliche Themen

2 K Nearest Neighbor Regression (KNN)
3 Regressionsbäume (TREE)

Nicht-Parametrische regression

5 Ridge Regression und Lasso Regression (LASSO)
1 Hauptkomponentenregression (PCR)

Big Data

7 Logistische Regression (LR)
4 Lineare Diskriminanzanalyse (LDA)

Binäre Kategorische variablen

6 Endogenität und Instrumentalvariablen (IV) — Kausalität

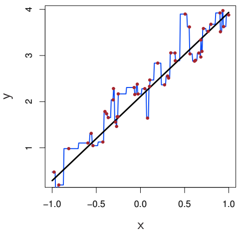
9 Autoregressive Distributed Lag Modell (ADL)
8 Spurious Regression Problem (SR)

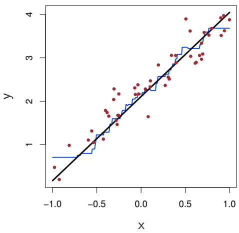
Time series

K Nearest Neighbor Regression (KNN)

- Nichtparametrische Alternative zur linearen Regression.
- Beobachtungen $Y_1, \ldots, Y_T$ und $X_1, \ldots, X_T$
- $\mathcal{N}(x)$ sei die K-Nachbarschaft für den Punkt $x \in \mathbb{R}$. Das sind die K Beobachtungen aus $\{X_1, \ldots, X_T\}$, die am dichtesten an x liegen.
- Die Regressionsfunktion $f(x)$ wird für jeden Punkt $x \in \mathbb{R}$ als Mittel seiner K-Nachbarschaft geschätzt:

$$\widehat{f}_{KNN}(x) = \frac{1}{K} \sum_{t:X_t \in \mathcal{N}(x)} Y_t$$





K Nearest Neighbor Regression (KNN)

Aufgabe:

- Stellen Sie die K Nearest Neighbor Methode zur Schätzung einer Regression dar.
- Gehen Sie auf einfache und auf multiple Regressionsprobleme ein
- Vergleichen Sie diese Methode mit der OLS Methode anhand einer Simulationsstudie
- Diskutieren Sie die Rolle des Parameters K und gehen Sie auf das Problem von Overfitting und den „Fluch der Dimensionalität“ ein

Literatur:

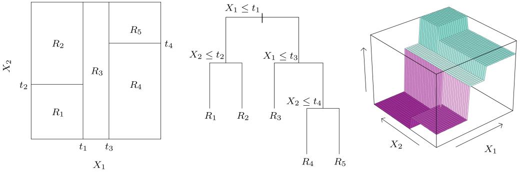
- James et al. (2013). *An Introduction to Statistical Learning*. Chapter 3.
- Hastie et al. (2009). *The Elements of Statistical Learning*. Chapter 2.

Regressionsbäume (TREE)

- Entscheidungsbäume/Binärbäume können auch zum schätzen von Regressionsproblemen verwendet werden
- Der Wertebereich von $X_1, \ldots, X_k$ wird in J Regionen $R_1, \ldots, R_J$ aufgeteilt.
- Sei $N_j = \#\{\boldsymbol{x}_t \in R_j\}$ die Anzahl der Beobachtungen in Region R_j. Für $x \in R_i$ wird $f(x)$ als Mittel über alle Beobachteten Y in der Region R_j geschtzt:

$$\widehat{f}_{TR}(x) = \frac{1}{N_j} \sum_{t:\boldsymbol{x}_t \in R_j} Y_t, \quad x \in R_j$$

- Erweiterungen: Bagging, Random Forests, Boosting



Regressionsbäume (TREE)

Aufgabe:

- Stellen Sie die Idee der Binärbäume dar und erläutern Sie wie diese zum schätzen von Regressionsproblemen genutzt werden können.
- Erklären Sie, wie die Regionen $R_1, \ldots, R_J$ bestimmt werden können.
- Führen Sie eine Simulationsstudie durch, in der Sie die Methode illustrieren und auch das Problem von Overfitting darstellen.
- Gehen Sie auf eine der Erweiterungen wie Bagging, Random Forests oder Boosting ein.

Literatur:

- James et al. (2013). *An Introduction to Statistical Learning*. Chapter 8.
- Hastie et al. (2009). *The Elements of Statistical Learning*. Chapter 10.

Ridge Regression und Lasso Regression (LASSO)

- Typische Regressionssituation: Viele Beobachtungen T und wenige Regressoren k
- Big Data Situation: $k \approx T$ oder sogar $k >> T$ (hochdimensional)
- Problem: Varianz des OLS-Schätzers steigt in k
- Shrinkage-Methoden mit zusätzlicher Strafkonstante $\lambda \geq 0$:

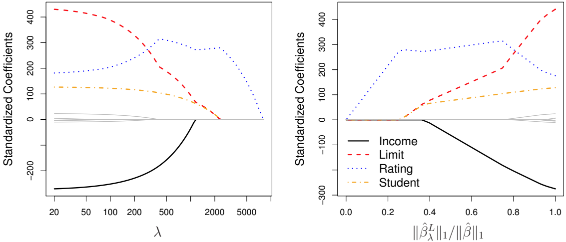
$$\text{OLS:} \quad \min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_k)'} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

=u prime u =

$$\text{Ridge:} \quad \min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_k)'} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}$$

=sum beta^2

$$\text{Lasso:} \quad \min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_k)'} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{k} |\beta_j|$$



Ridge Regression und Lasso Regression (LASSO)

Aufgabe:

- Stellen Sie die Ridge und Lasso Methode zur Schätzung einer Regression dar
- Erläutern Sie wie der Ridge und der Lasso Schätzer von der unbekannten Strafkonstante λ beeinflusst wird.
- Illustrieren Sie anhand von simulierten Beispielen, wie der Ridge und Lasso Schätzer von λ abhängen.

Literatur:

- James et al. (2013). *An Introduction to Statistical Learning*. Chapter 6.
- Hastie et al. (2009). *The Elements of Statistical Learning*. Chapter 3.
- Stock and Watson (2020). *Introduction to Econometrics*. Chapter 14.

Hauptkomponentenregression (PCR)

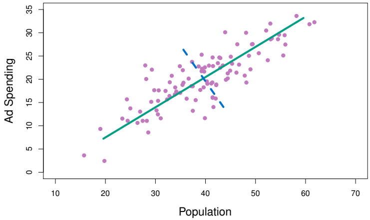
- Principal Component Regression (PCR)
- Big Data Situation: Viele Regressoren k
- Das hochdimensionale Problem wird in ein Problem mit niedriger Dimension überführt.

Etwas schwerer aber lohnt sich

- Modell mit $p << k$ Hauptkomponenten:

$$Y_t = \mu + \gamma_1 Z_1 + \ldots + \gamma_p Z_p + \epsilon_t$$

- Die Hauptkomponenten $Z_1, \ldots, Z_p$ sind Linearkombinationen der Variablen $X_1, \ldots, X_k$.
- Die Linearkombinationen werden dabei geschickt gewählt, sodass möglichst viel Variation in den Variablen erklärt wird.



Hauptkomponentenregression (PCR)

Aufgabe:

- Stellen Sie die Methode der Hauptkomponentenregression dar.
- Zeigen Sie wie die Hauptkomponenten bestimmt/geschätzt werden.
- Gehen Sie auf Methoden zur Bestimmung der Anzahl der Hauptkomponenten p ein.
- Illustrieren Sie die Methode anhand von simulierten Beispielen.

Literatur:

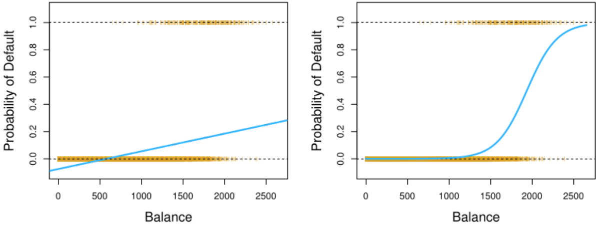
- James et al. (2013). *An Introduction to Statistical Learning*. Chapter 6.
- Hastie et al. (2009). *The Elements of Statistical Learning*. Chapter 3, 14.
- Backhaus et al. (2018). *Multivariate Analysemethoden*. Kapitel 7.
- Stock and Watson (2020). *Introduction to Econometrics*. Chapter 14, 17.

Logistische Regression (LR)

- Logistic Regression
- Beobachtungen $Y_1, \ldots, Y_T$ und $X_1, \ldots, X_T$, wobei Y binär skaliert ist: $Y_t \in \{0, 1\}$
- Klassifikationsproblem: Was ist $p(x) = P(Y = 1|X = x)$?
- Logistische Regression:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- Parameterschätzung mit der Maximum Likelihood Methode



Logistische Regression (LR)

Aufgabe:

- Stellen Sie die Methode der Logistischen Regression dar.
- Gehen Sie auch auf die multivariate logistische Regression ein.
- Zeigen Sie wie das Modell geschätzt wird.
- Illustrieren Sie die Methode mithilfe einer Simulationsstudie und bewerten Sie die Schätzgüte.
- Gehen Sie auf den Kompromiss zwischen der positiven und der negativen Fehlklassifikationsrate und auf die ROC Kurve ein.

Literatur:

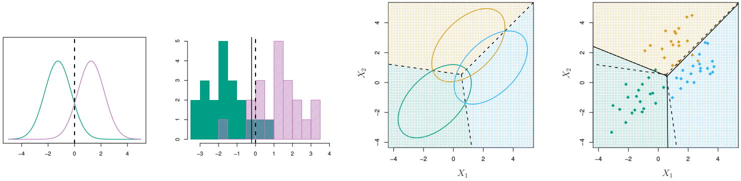
- James et al. (2013). *An Introduction to Statistical Learning*. Chapter 4.
- Hastie et al. (2009). *The Elements of Statistical Learning*. Chapter 4.
- Backhaus et al. (2018). *Multivariate Analysemethoden*. Kapitel 4.
- Stock and Watson (2020). *Introduction to Econometrics*. Chapter 11.

Lineare Diskriminanzanalyse (LDA)

- (Linear) discriminant analysis
- Beobachtungen $Y_1, \dots, Y_T$ und $X_1, \dots, X_T$
- Klassifizierungsproblem: Y ist nominalskaliert mit K Klassen: $Y_t \in \{1, 2, \dots, K\}$, X ist metrisch skaliert
- X wird als normalverteilt angenommen und für jede Klasse $k = 1, \dots, K$ werden die Parameter der Normalverteilung separat geschätzt, mit Dichte $f_k(x)$
- $P(Y|X = x)$ erhalten wir mithilfe des Satz von Bayes:

$$P(Y = k|X = x) = \frac{P(Y = k)f_k(x)}{\sum_{l=1}^{K} P(Y = l)f_l(x)}$$

ein bisschen ähnlich wie das vorherige Thema



Lineare Diskriminanzanalyse (LDA)

Aufgabe:

- Stellen Sie die Lineare Diskriminanzanalyse dar.
- Betrachten Sie auch die multivariate LDA sowie die quadratische Diskriminanzanalyse.
- Zeigen Sie wie das Modell geschätzt wird.
- Illustrieren Sie die Methode mithilfe einer Simulationsstudie und bewerten Sie die Schätzgüte.
- Gehen Sie auf den Kompromiss zwischen der positiven und der negativen Fehlklassifikationsrate und auf die ROC Kurve ein.

Literatur:

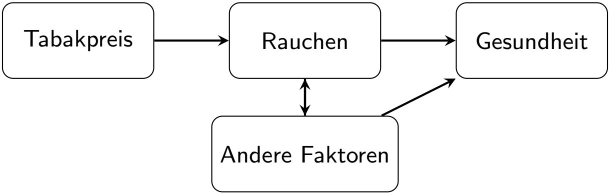
- James et al. (2013). *An Introduction to Statistical Learning*. Chapter 4.
- Hastie et al. (2009). *The Elements of Statistical Learning*. Chapter 4.
- Backhaus et al. (2018). *Multivariate Analysemethoden*. Kapitel 4.

Endogenität und Instrumentalvariablen (IV)

- Wir betrachten das lineare Regressionsmodell

$$Y_t = \beta_0 + \beta_1 X_t + u_t, \quad t = 1, \ldots, T.$$

  Um β_1 zu schätzen, nimmt man normalerweise an, dass X_t mit dem Fehlerterm u_t unkorreliert ist, d.h. $Cov(X_t, u_t) = 0$ (X_t ist exogen).
- Falls diese Bedingung verletzt ist, d.h. falls $Cov(X_t, u_t) \neq 0$, spricht man von **Endogenität**.
- In diesem Fall liefert die herkömmliche Methode der Kleinsten Quadrate einen verzerrten und sogar inkonsistenten Schätzer.
- Der **Instrumentalvariablen- ansatz** ist eine Methode, um den Parameter β zu schätzen, wenn Endogenität vorliegt.



Endogenität und Instrumentalvariablen (IV)

Aufgabe:

- Stellen Sie den Ansatz der Instrumentalvariablenschätzung im linearen Regressionsmodell dar und erläutern Sie den zweistufigen OLS-Schätzer (two stage least squares, kurz 2SLS).
- Erklären Sie, warum der herkömmliche OLS-Schätzer nicht gut funktioniert, wenn Endogenität vorliegt.
- Führen Sie Simulationen durch, die den herkömmlichen OLS-Schätzer mit dem 2SLS-Schätzer vergleichen.

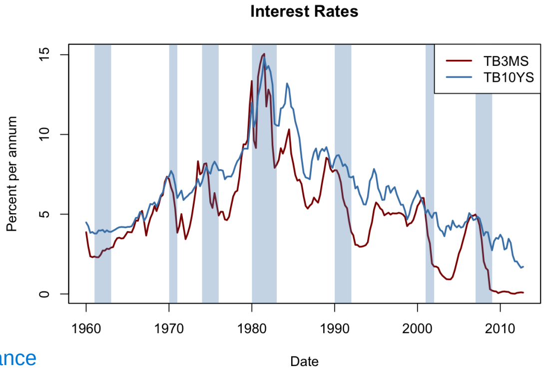
Literatur:

- Angrist and Pischke (2009). *Mostly Harmless Econometrics*. Chapter 4.
- Stock and Watson (2020). *Introduction to Econometrics*. Chapter 12.

Autoregressive Distributed Lag Modell (ADL)

- Makroökonomische Zeitreihen sind in der Regel autokorreliert.
- Daher werden oft dynamische Regressionsmodelle betrachtet:

$Y_t = \alpha + \rho_1 Y_{t-1} + u_t$	Autoregressives Modell (AR)
$Y_t = \alpha + \beta_1 X_t + u_t$	Einfache Regression
$Y_t = \alpha + \beta_1 X_t + \beta_2 X_{t-1} + u_t$	Distributed Lag Modell (DL)
$Y_t = \alpha + \rho_1 Y_{t-1} + \beta_1 X_t + \beta_2 X_{t-1} + u_t$	ADL(1,1) Modell



Makro oder Finance

Autoregressive Distributed Lag Modell (ADL)

Aufgabe:

- Stellen Sie dar, wie das ADL Modell geschätzt wird und wie es zur Prognose von Y dienen kann.
- Erläutern Sie wie die Koeffizienten als kurzfristige und langfristige Multiplikatoren interpretiert werden können.
- Zeigen Sie anhand von Simulationen, wie die Laglängen p und q im allgemeinen ADL(p, q) Modell mithilfe von Informationskriterien wie AIC und BIC geschätzt werden können.

Literatur:

- Neusser (2016). *Time Series Econometrics*, Chapter 3, 5
- Stock and Watson (2020). *Introduction to Econometrics*, Chapter 15

Spurious Regression Problem (SR)

- Wir betrachten das lineare Regressionsmodell

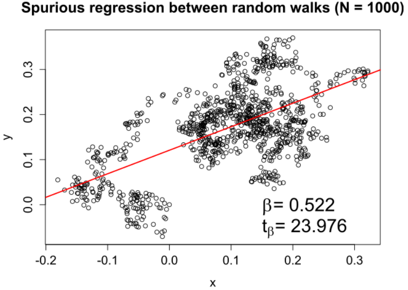
$$Y_t = \beta_0 + \beta_1 X_t + u_t, \quad t = 1, \ldots, T.$$

- Im Fall $\beta_1 = 0$ besteht kein Zusammenhang zwischen den Variablen, und die t-Statistik sowie das R^2 sollten nahe bei 0 liegen.
- Dennoch können die t-Statistik und das R^2 irreführend hoch sein, wenn die Zeitreihen X_t und Y_t hoch autokorreliert sind.
- Zum Beispiel im Fall von Random Walks:

$$X_t = X_{t-1} + v_t, \quad Y_t = Y_{t-1} + w_t,$$

wobei v_t und w_t unabhängig und i.i.d. sind.

Spannend wenn man Zeitreihen mag, sonst ein bisschen spezifiesch



Spurious Regression Problem (SR)

Aufgabe:

- Erläutern Sie das Problem einer Spurious Regression im einfachen Regressionsmodell.
- Illustrieren Sie den Effekt der irreführend hohen t-Statistiken anhand von Simulationen.
- Diskutieren Sie Lösungsmöglichkeiten, wie die Verwendung alternativer Standardfehler, das Transformieren der Zeitreihen in Differenzen oder die Aufnahme zusätzlicher Lags.

Literatur:

- Neusser (2016). *Time Series Econometrics*, Chapter 7
- Stock and Watson (2020). *Introduction to Econometrics*, Chapter 15

Einführende Literatur (kostenlos per VPN)

- Angrist, J. D., & Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
  Link: `https://www.degruyter.com/view/title/563369`
- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2018). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung (15. Auflage.)*. Springer.
  Link: `https://www.springer.com/de/book/9783662566541`
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. Springer.
  Link: `https://www.springer.com/de/book/9780387848570`
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.
  Link: `https://www.springer.com/de/book/9781461471370`
- Neusser, K. (2016). *Time Series Econometrics.* Springer.
  Link: `https://www.springer.com/de/book/9783642334351`
- Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics (Fourth edition)*. Pearson.
  Link: `https://ebookcentral.proquest.com/lib/ulb-bonn/detail.action?docID=5834470`