1. Lösungen zu Kapitel 8

Übungsaufgabe 8.1

a) *Falsch!*
Die Nichtberücksichtigung von unwichtigen Variablen für die Identifikation kausaler Effekte stellt kein Problem dar, sofern diese Variablen keinen Beitrag zur Erklärung der abhängigen Variablen liefern und nicht mit den im Modell berücksichtigten Faktoren korreliert sind.
Betrachtet man z.B. das folgende Modell

$$Y = \beta_0 + \beta_1 X_1 - \beta_2 X_2 + \varepsilon,$$

so führt Auslassen der Variable X_2 zur Schätzung von

$$Y = \beta_0 + \beta_1 X_1 + \nu.$$

Die Variable X_2 würde zum Teil des Fehlerterms ($\nu = \beta_2 X_2 + \varepsilon$) und der Erwartungswert des Fehlerterms hätte die Form

$$E(\nu) = E(\beta_2 X_2 + \varepsilon) = E(\beta_2 X_2) + E(\varepsilon) = E(\beta_2 X_2) = \beta_2 X_2.$$

Solange X_2 nicht mit Y korreliert ist (somit im statistischen Sinne unwichtig ist ($\beta_2 = 0$)), ist die Annahme (3) des linearen Regressionsmodells ($E(\nu) = 0$) nicht verletzt. In diesem Fall gilt

$$Cov(X_1, \nu) = Cov(X_1, \beta_2 X_2 + \varepsilon) = \beta_2 Cov(X_1, X_2) + Cov(X_1, \varepsilon) = 0.$$

b) *Richtig!*
Während sowohl für Proxy-Variablen als auch für Instrumente gilt, dass sie mit der endogenen erklärenden Variablen korreliert und für die strukturelle Gleichung redundant sein müssen, ist der entscheidende Unterschied, dass das Instrument nicht mit unbeobachteten Faktoren im Fehlerterm der strukturellen Gleichung korreliert sein darf.
Proxy-Variablen zeichnen sich im Unterschied zu einem Instrument gerade durch die Eigenschaft aus, dass sie mit der unbeobachteten Variable im Fehlerterm korreliert sind.

Übungsaufgabe 8.2

Annahme (8.18) ist die Identifikationsannahme, die erfüllt sein muss, damit es sich um eine *gute* Proxy-Variable handelt. Sie lautet:

$$E(X_2|X_1, \widetilde{X}_2) = E(X_2|\widetilde{X}_2) = \alpha_0 + \alpha_1 \widetilde{X}_2.$$

Ist diese Annahme verletzt, so gilt

$$E(X_2|X_1, \widetilde{X}_2) = \alpha_0 + \delta_1 X_1 + \alpha_1 \widetilde{X}_2.$$

Dies führt dazu, dass

$$\begin{aligned} E(Y|X_1, \widetilde{X}_2) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ &= \beta_0 + \beta_1 X_1 + \beta_2(\alpha_0 + \delta_1 X_1 + \alpha_1 \widetilde{X}_2) \\ &= (\beta_0 + \alpha_0) + (\beta_1 + \beta_2 \delta_1) X_1 + \beta_2 \alpha_1 \widetilde{X}_2. \end{aligned}$$

Der Erwartungswert von $\hat{\beta}_1$ ist nicht mehr β_1 sondern $E(\hat{\beta}_1) = \beta_1 + \beta_2 \delta_1$.

Übungsaufgabe 8.3

a) Spalte (1) der Tabelle enthält die Ergebnisse einer OLS-Schätzung der Lohnfunktion basierend auf 3.010 Beobachtungen. Der logarithmierte Lohn ist die abhängige Variable und die Anzahl der Schuljahre die erklärende Variable.
Der Koeffizient der Schuljahre beträgt 0,075 und ist statistisch signifikant von Null verschieden ($|t| = 25$). Im Schnitt bedeutet ein weiteres Schuljahr 7,5% höhere Löhne. Das Bestimmtheitsmaß der Schätzung ist 0,300. 30% der Variation der Löhne kann folglich mit der Variation der absolvierten Schuljahren erklärt werden.

b) Es ist wahrscheinlich, dass der $School_i$-Koeffizient aufgrund unbeobachtbarer Eigenschaften nach oben verzerrt ist. Zu diesen Eigenschaften gehören z.B. angeborene Fähigkeiten, Motivationen, etc. Personen, die motivierter sind, haben wahrscheinlich auch mehr Schuljahre absolviert. Der Koeffizient spiegelt folglich nicht nur den Effekt der Schuljahre auf die Löhne wieder, sondern greift auch einen Teil des Effekts der Motivation mit auf.
Dies kann auch formal gezeigt werden. Das ursprünglich geschätzte Modell hat die Form

$$\ln(w) = \beta_0 + \beta_1 School_i + \nu.$$

Aufgrund unbeobachtbarer Eigenschaften X, die nicht in dem geschätzten Modell enthalten sind, jedoch mit den Schuljahren korreliert sind, ist die Annahme $E(School_i \nu) = 0$ verletzt. Dies führt zu einer Verzerrung des geschätzten Koeffizienten $\hat{\beta}_1$:

$$
\begin{aligned}
E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^{N}(School_i - \overline{School})(Y_i - \overline{Y})}{\sum_{i=1}^{N}(School_i - \overline{School})^2}\right) \\
&= E\left(\frac{\sum_{i=1}^{N}(School_i - \overline{School})(\beta_1(School_i - \overline{School}) + \nu)}{\sum_{i=1}^{N}(School_i - \overline{School})^2}\right) \\
&= \beta_1 + E\left(\frac{\sum_{i=1}^{N}(School_i - \overline{School})\nu}{\sum_{i=1}^{N}(School_i - \overline{School})^2}\right).
\end{aligned}
$$

c) Die Ergebnisse der Instrumentvariablen-Schätzung stimmen nicht mit der Erwartung überein, dass der Effekt eines weiteren Schuljahres aufgrund unbeobachtbarer Heterogenität überschätzt ist, d.h. der Koeffizient nach oben verzerrt ist. Bei der IV-Schätzung steigt der Koeffizient von $School_i$ sogar im Vergleich zum OLS-Schätzer.
Die Erklärung könnten heterogene Maßnahmeneffekte sein, d.h. das Maßnahmeeffekt nicht für alle Personen gleich sind. Dies bedeutet, dass man theoretisch nicht β_1 sondern β_{1i} schätzen müsste. Es kann gezeigt werden, dass der OLS-Schätzer β_1 ein erwartungstreuen Schätzer des durchschnittlichen Maßnahmeneffekts darstellt.
Dies gilt jedoch nicht für den Instrumentvariablenschätzer. Hier liefert der Schätzer nur den Effekt für die Bevölkerungsgruppe, deren Verhalten durch das Instrument beeinflusst wurde (*local average treatment effect*).
Im vorliegenden Beispiel erfolgt die Identifikation des kausalen Effekts nur für diejenigen Personen, deren Anzahl an Schuljahren durch die Nähe des Colleges beeinflusst wurde. Dies sind insbesondere Personen, die prinzipiell eher eine geringere Anzahl von Schuljahren anstreben, sich aber durch die Nähe zum College (und durch die damit verbundenen geringeren Kosten der Bildung) doch für ein weiteres Schuljahr entscheiden. Diese Personen scheinen eine überdurchschnittliche Rendite aus einem weiteren Schuljahr zu ziehen. Ein weiteres Schuljahr erhöht die Löhne im Schnitt um 13,2%.

d) Eine Instrumentvariable Z muss zwei Annahmen erfüllen. Erstens muss sie exogen sein, d.h. sie darf nicht mit den im Fehlerterm enthaltenen unbeobachtbaren Variablen korreliert sein ($Cov(Z, \nu) = 0$). Im vorliegenden Beispiel darf das Instrument folglich nicht mit unbeobachtbaren Eigenschaften wie Motivation oder angeborenen Fähigkeiten korreliert sein.
Zweitens muss sie mit der endogenen Variablen korreliert sein, d.h. sie muss relevant sein ($Cov(Z, X) \neq 0$). Die Nähe zum College muss also einen Einfluss auf die absolvierten Schuljahre haben.
Sind beide Annahmen erfüllt, ist $Near$ ein valides Instrument.

e) Die Annahme der Exogenität kann nicht getestet werden. Die zweite Annahme, sprich die Relevanz des Instruments, kann hingegen durch Regression der endogenen Variablen X auf das Instrument Z überprüft werden kann.
Im vorliegenden Beispiel schätzt man hierfür die Regression:

$$School_i = \gamma_0 + \gamma_1 Near_i + \nu_i.$$

Mit Hilfe des t-Tests kann getestet werden, ob γ_1 statistisch signifikant von Null verschieden ist und somit die Nähe zum College einen signifikanten Einfluss auf die Schuljahre hat.

f) Es ist gut möglich, dass sowohl in der abhängigen als auch in der erklärenden Variable Messfehler vorliegen. Solange der Messfehler der abhängigen Variablen, sprich der Löhne, nicht mit den Schuljahren korreliert ist, liegt kein Problem vor. Dies kann zwar zu einer Verzerrung der Konstanten führen, aber der Steigungsparamter β_1 wäre weiterhin unverzerrt.
Wird jedoch *School* mit einem Fehler gemessen, so ist die Kovarianz dieser Variablen mit dem Fehlerterm nicht mehr Null und die OLS-Schätzung führ zu verzerrten und inkonsistenten Koeffizienten. Häufig führt dies zu einer Verzerrung von $\hat{\beta}_1$ gegen Null (*attenuation bias*).
Die Tatsache, dass die IV-Schätzung einen weitaus größeren Effekt für *School* ergibt, könnte als Hinweis für einen Messfehler von *School* interpretiert werden.

g) Wird angenommen, dass $Near$ ein mögliches Instrument für *School* ist, kann die Dummy-Variable prinzipiell keine geeignete Proxy-Variable sein.
Mit Proxy-Variablen möchte man für Variablen kontrollieren, die nicht beobachtet werden können. Sie müssen folglich mit diesen korreliert sein.
Im vorliegenden Beispiel möchten man den Effekt von *School* auf die Löhne erhalten. Das Problem ist, dass der Lohneffekt der Schulausbildung wahrscheinlich mit unbeobachtbaren Eigenschaften wie Motivation, Talent, etc. korreliert und β_1 somit verzerrt.
Möchte man nun den Proxy-Variablen-Ansatz verwenden, so wäre es das Ziel, eine Variable zu finden, die mit diesen unbeobachtbaren Eigenschaften und folglich dem Fehlerterm korreliert ist. Hat man eine geeignete Proxy-Variable, so erhält man einen unverzerrten Schätzer für β_1.
Beim IV-Ansatz sucht man hingegen nach einer Variablen, die gerade *nicht* mit dem Fehlerterm korreliert ist, um den Effekt der exogenen Variation von *School* zu erhalten.

Übungsaufgabe 8.4

a) Die vorliegende Schätzung basiert auf Daten des Sozioökonomischen Panels (SOEP) und die Stichprobe umfasst 3.159 Beobachtungen. Es wird folgende Lohngleichung geschätzt

$$\ln(w_i) = \beta_0 + \beta_1 S_i + \beta_2 B_i + \beta_3 B_i^2 + \beta_4 F_i + \beta_5 F_i \cdot S_i + \beta_6 R_i + \varepsilon_i,$$

wobei S_i die Jahre der Schulausbildung von Individuum i sind, B_i ist die Arbeitsmarkterfahrung, F_i ist eine Dummy-Variable, die den Wert 1 annimmt, wenn die Person weiblich ist, R_i ist die Anzahl der Zigaretten, die pro Tag konsumiert werden, und ε_i ist der normalverteilte Fehlerterm mit dem Mittelwert 0 und der Varianz σ^2.
Der geschätzte Koeffizient der Schulausbildung beträgt 0,084 und ist statistisch signifikant verschieden von Null. Dies bedeutet, dass ein weiteres

Jahr Schulausbildung einen marginalen Effekt von 8,4% auf die Löhne von Männern hat.
Arbeitsmarkterfahrung hat einen positiven und abnehmenden Effekt auf die Löhne. Der Effekt gleicht einem umgedrehten U. Der marginale Effekt hängt vom aktuellen Niveau ab, so dass er nur für konkrete Anzahl von Jahren bestimmt werden kann. Ab acht Jahren Berufserfahrung ist der Effekt von Arbeitsmarkterfahrung auf die Löhne negativ.
Frauen haben signifikant niedrigere Löhne als Männer. Im Schnitt liegt diese Differenz bei 20,1%. Darüber hinaus haben Frauen im Vergleich zu Männern eine geringere Rendite der Schulausbildung. Der Koeffizient des Interaktionsterms zwischen den Jahren der Schulausbildung und der Dummy-Variablen für Frauen ist statistisch signifikant verschieden von Null und beträgt -0,015. Während die Rendite der Schulausbildung für Männer 8,4% beträgt, ist sie für Frauen nur 6,9%.
Ebenfalls einen negativen Effekt auf die Löhne hat der Zigarettenkonsum. Pro Zigarette mehr pro Tag sinken die Löhne um 2,4%.
Das Bestimmtheitsmaß dieser Schätzung beträgt 0,301. Dementsprechend können 30,1% der Varianz in den Löhnen mit dem vorliegenden Modell erklärt werden.

b) Um zu testen, ob die Variablen B_i und B_i^2 gemeinsam einen statistisch signifikanten Einfluss auf die Löhne haben, verwendet man den F-Test oder den LM-Test ($H_0 : \beta_2 = \beta_3 = 0$).
Für den F-Test wird das Modell einmal mit den zu testenden Variablen B_i und B_i^2 (*ur=unrestricted*) und einmal ohne diese geschätzt (*r=restricted*). Anschließen kann aus den Bestimmtheitsmaßen dieser Schätzungen die Teststatistik errechnet werden.

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(N - k)}$$

Überschreitet die F-Statistik den kritischen Wert der F-Verteilung bei gegebener Irrtumswahrscheinlichkeit, kann die Nullhypothese abgelehnt werden.

c) Es ist gut möglich, dass die Variable R_i, Zigaretten pro Tag, endogen ist. Da z.B. in der Lohnregression nicht für den Gesundheitszustand kontrolliert wird, bedeutet dies, dass dieser im Fehlerterm enthalten ist. Da der Gesundheitszustand wiederum mit dem Rauchen korreliert ist und Personen mit einem schlechten Gesundheitszustand aufgrund ihrer geringeren Produktivität im Durchschnitt einen geringeren Lohn erhalten könnten, ist $\hat{\beta}_6$ potentiell verzerrt.
Die Verzerrung kann formal folgendermaßen dargestellt werden:

$$\begin{aligned} E(\hat{\beta}_6) &= \beta_6 + \beta_7 \cdot \frac{Cov(R_i Gesundheit)}{Var(R_i)} \\ &= \beta_6 + \beta_7 \cdot \alpha, \end{aligned}$$

wobei $beta_7$ den Effekt des Gesundheitszustands auf den Lohn angibt und α den Steigungskoeffizienten einer Regression von R_i auf den Gesundheitszustand.
Da die Kovarianz zwischen der Gesundheit und dem Rauchen negativ ist ($\alpha < 0$), ist davon auszugehen, dass der Koeffizient von R_i nach unten verzerrt ist.
Damit die Dummy-Variable RJ_i, die den Wert 1 annimmt, wenn eine Person schon vor Erreichen des 16. Lebensjahr regelmäßig geraucht hat, ein valides Instument für R_i ist, müssen zwei Annahmen erfüllt sein.
Erstens muss $Cov(RJ_i\varepsilon_i) = 0$ gelten, sprich RJ_i darf nicht mit den unbeobachtbaren Faktoren korreliert sein, die einen Effekt auf die Löhne haben.
Zweitens darf die Kovarianz zwischen RJ_i und R_i nicht Null sein, d.h. RJ_i muss einen Einfluss auf R_i haben. Da $\hat{\delta}_6$ statistisch signifikant verschieden von Null ist, ist die zweite Annahme erfüllt.
Die erste Annahme kann nicht empirisch getestet werden, sondern nur durch theoretische Überlegungen gerechtfertigt werden. Auf der einen Seite könnte man argumentieren, dass der Zigarettenkonsum mit 16 zwar den Zigarettenkonsum heutzutage erklärt, aber nicht direkt mit dem Gesundheitszustand von heute verbunden ist. Auf der anderen Seite könnte man sagen, dass gerade der frühe Zigarettenkonsum zu gesundheitlichen Schäden geführt hat, die stark mit dem heutigen Gesundheitszustand korreliert sind.

d) Der Test auf Endogenität testet, ob die OLS-Schätzer der Koeffizienten signifikant verschieden von den TSLS-Schätzern sind. Ist diese Differenz signifikant, geht man davon aus, dass Endogenität vorliegt.
Im ersten Schritt schätzt man eine Hilfsregression, die die Zigaretten pro Tag auf das Instrument und die Kontrollvariablen regressiert.

$$R_i = \beta_0 + \delta_1 S_i + \delta_2 B_i + \delta_3 B_i^2 + \delta_4 F_i + \delta_5 F_i \cdot S_i + \delta_6 RJ_i + \nu_i$$

Die geschätzten Residuen ($\hat{\nu}_i$) werden im zweiten Schritt in das strukturelle Modell als zusätzliche erklärende Variable aufgenommen. Geschätzt wird:

$$\ln(w_i) = \beta_0 + \beta_1 S_i + \beta_2 B_i + \beta_3 B_i^2 + \beta_4 F_i + \beta_5 F_i \cdot S_i + \beta_6 R_i + \alpha\hat{\nu}_i + \varepsilon_i.$$

Nun wird mit Hilfe des t-Tests die Nullhypothese überprüft, dass der Koeffizient der geschätzten Residuen gleich Null ist ($H_0 : \alpha = 0$). Kann H_0 abgelehnt werden, geht man davon aus, dass die Zigaretten pro Tag R_i endogen sind, da ε_i und ν_i miteinander korreliert sind.

Übungsaufgabe 8.5

a) Der zweistufige Kleinstquadratschätzer (TSLS) ist eine Möglichkeit den Instrumentvariablenansatz zu implementieren.

Hierzu wird im ersten Schritt der durchschnittliche tägliche Alkoholkonsum der Mutter während der Schwangerschaft (in Gramm) auf das Instrument, in diesem Fall den Alkoholkonsum des Partners der werdenden Mutter (AP_i), regressiert:

$$A_i = \delta_0 + \delta_1 AP_i + \nu_i. \tag{1.1}$$

Im zweiten Schritt wird das strukturelle Modell mit dem geschätzten Alkoholkonsum der Mutter ($\hat{A}_i$) aus Gleichung (1.1) anstelle des tatsächlichen Alkoholkonsums (A_i) geschätzt:

$$\ln(Geburtsgewicht) = \boldsymbol{X}\alpha + \beta\hat{A}_i + \varepsilon_i.$$

Damit diese Methode zu konsistenten Schätzern führt, müssen zwei wichtige Identifikationsannahmen erfüllt sein. Erstens darf der Alkoholkonsum des Partners nicht mit den unbeobachtbaren Faktoren, die im Zusammenhang mit dem Gesundheitsbewusstsein der Mutter stehen, korreliert sein ($Cov(AP_i\varepsilon_i) = 0$). Das Instrument muss exogen sein.
Zweitens muss ein signifikanter Zusammenhang zwischen dem Alkoholkonsum des Partners und dem Alkoholkonsum der werdenden Mutter bestehen. Dies bedeutet, dass $Cov(A_i AP_i) \neq 0$. Ist dies der Fall, so ist δ_1 signifikant von Null verschieden
Während die zweite Annahme getestet werden kann, kann die erste Annahme nur über Plausibilitätsüberlegungen verteidigt werden. Auf der einen Seite kann man argumentieren, dass der Alkoholkonsum des Partners nicht direkt mit dem Gesundheitsbewusstsein der werdenden Mutter korreliert ist. Auf der anderen Seite könnte man argumentieren, dass sich zwei Partner vielleicht dadurch finden, dass sie sich unter anderem auch in solchen Eigenschaften sehr ähneln.

b) Der Koeffizient der TSLS-Schätzung ist - genau wie der Koeffizient der OLS-Schätzung - signifikant verschieden von Null und negativ. Während 1 Gramm Alkohol das Geburtsgewicht in der OLS-Schätzung um 2,1% reduziert, beträgt dieser Effekt 0,5% bei der TSLS-Schätzung.
Dies stimmt mit der Erwartung überein, dass der Koeffizient der OLS-Schätzung nach unten verzerrt ist, da er den Effekt von anderen Gesundheits- und Verhaltensfaktoren der Mutter auffängt. Ohne das Instrument wird der negative Einfluss des Alkoholkonsums auf das Geburtsgewicht des Kindes überschätzt.

Übungsaufgabe 8.6

Die Rendite aus Schulbildung ist in einer Mincer'schen Lohnfunktion häufig aufgrund unbeobachtbarer Heterogenität verzerrt. Fixed-Effects-Modelle bieten keine Lösung, da sie es nicht ermöglichen, erklärende Variablen zu identifizieren, die nur eine geringe Variation über die Zeit aufweisen, da alle zeitinvarianten Faktoren eliminiert werden.

Die Idee Zwillingsdaten zu verwenden beruht auf der Tatsache, dass eineiige Zwilling sich durch identische genetische Informationen auszeichnen. Durch Bildung von Differenzen kann der Einfluss angeborener kognitiver Fähigkeiten eliminiert werden.

Die einfache Lohnregression von Zwilling i in Familie j sieht folgendermaßen aus:

$$\ln(w_{ij}) = \boldsymbol{X}_{ij}\beta + \delta S_{ij} + \alpha_j + \varepsilon_{ij}.$$

Durch die Bildung von Differenzen entfällt der fixe Familieneffekt α_j.

$$\ln(w_{1j}) - \ln(w_{2j}) = (\boldsymbol{X}_{1j} - \boldsymbol{X}_{2j})\beta + \delta(S_{1j} - S_{2j}) + (\varepsilon_{1j} - \varepsilon_{2j})$$

Eine notwendige Bedingung, damit dieses Vorgehen einen unverzerrten Schätzer für die Bildungsrendite liefert, ist die ausreichende Anzahl von beobachteten Zwillingspaaren, für die die Schulausbildung variiert.

Übungsaufgabe 8.7

a) Die Modell schätzt die Note eines Studenten in einer bestimmten Vorlesung in Abhängigkeit davon, ob es sich um ein Fach der Speziellen VWL oder BWL handelt, der Anwesenheit des Studenten in der Vorlesung, der wöchentlichen Arbeitszeit neben dem Studium sowie der Anzahl der Fachsemester.
   Die Note eines Studenten ist signifikant höher, wenn es sich bei der Vorlesung um eine Vorlesung der Speziellen Volkswirtschaftslehre oder der speziellen Betriebswirtschaftslehre handelt. Ist dies der Fall, steigt die Note des Studenten um 0,57 Punkte.
   Die Teilnahme an der Vorlesung hat keinen signifikant von Null verschiedenen Einfluss auf die Note ($|t| = 1,5$).
   Die Anzahl der wöchentlichen Stunden, die der Student neben dem Studium arbeitet, verschlechtert die Note hingegen signifikant. Sowohl die Anzahl der wöchentlichen Stunden als auch das Quadrat der wöchentlichen Arbeitsstunden sind signifikant verschieden von Null. Dies bedeutet, dass die Note des Studenten mit dem Arbeitspensum exponentiell fällt. Arbeitet er z.B. vier Stunden die Woche, so verringert sich die Note um 3,7 Punkte, während ein Arbeitspensum von 8 Stunden zu einer Verschlechterung von 7,5 führt.
   Schließlich ist der Einfluss der abgeschlossenen Fachsemester ebenfalls nicht signifikant verschieden von Null.
   Insgesamt können die erklärenden Variablen 48% der gesamten Varianz der Noten erklären ($R^2 = 0,480$).

b) Eine Verzerrung des Koeffizienten des Anteils der Semesterwochenstunden, die der Student an der Vorlesung teilgenommen hat A_{iv} kann auftreten, wenn die Teilnahme z.B. mit der unbeobachteten Motivation des

Studenten korreliert ist. Besonders motivierte Studenten sind wahrscheinlich häufiger anwesend und haben im Schnitt bessere Noten.
Diese positive Korrelation zwischen der unbeobachteten Motivation mit der Anwesenheit sowie mit der Note führt zu einer Verzerrung des Koeffizienten nach oben.
Formal kann dies folgendermaßen ausgedrückt werden:

$$E(\hat{\beta}_A) = \beta_A + \beta_X \frac{Cov(AX)}{Var(A)}.$$

Die Variable B_i, Anzahl der Schuljahre der Mutter, wäre als Proxy-Variable zur Lösung dieses Problems geeignet, wenn sie erstens mit der unbeobachteten Motivation des Studenten korreliert wäre. Zweitens müsste sie redundant für das Modell sein. Dies bedeutet, dass die Anzahl der Schuljahre der Mutter keinen Einfluss auf die Note haben darf, sobald für die unbeobachtete Motivation und die anderen erklärenden Variablen kontrolliert wurde, und dass zwischen der unbeobachteten Motivation und den anderen erklärenden Variablen kein Zusammenhang mehr bestehen darf sobald für die Proxy-Variable kontrolliert wird.
Wenn diese Annahmen erfüllt wären, wäre B_i eine geeignete Proxy-Variable.

c) In dem vorliegenden Fall liegen tatsächlich falsche Standardfehler vor. In der Stichprobe liegen für denselben Studenten i mehrere Beobachtungen vor, da dieser wahrscheinlich mehr als eine Vorlesung v besucht. Die Fehler der Studenten sind über die von ihnen besuchten Vorlesungen miteinander korreliert.

d) Es gibt verschiedene Möglichkeiten die Panelstruktur der Daten zu nutzen, um via Fixed-Effects-Schätzungen zeitinvariante unbeobachtbare Heterogenität zu eliminieren. Es muss jedoch beachtet werden, dass in einem solchen Modell keine Variablen verwendet werden können, die über die Vorlesungen hinweg konstant sind. Das ursprüngliche Modell sieht folgendermaßen aus:

$$N_{iv} = \beta_1 S_{iv} + \beta_2 A_{iv} + \alpha_i + \varepsilon_{iv},$$

Ein möglicher Ansatz ist die Fixed-Effects-Transformation. Hierfür bildet man für den Studenten den Durchschnitt über alle besuchten Kurse.

$$\overline{N}_i = \beta_0 + \beta_1 \overline{S_i} + \beta_2 \overline{A_i} + \alpha_i + \overline{\varepsilon}_i$$

Subtrahiert man diese Gleichung nun von Gleichung (8.55) ergibt sich:

$$\widetilde{N}_i = \beta_1 \widetilde{S}_i + \beta_2 \widetilde{A}_i + \widetilde{\varepsilon}_i$$

Das Problem der verzerrten Koeffizienten kann nur mit Hilfe der Fixed-Effects-Transformation gelöst werden, sofern die Ursache der Verzerrung in einem Faktor liegt, der über die Vorlesungen hinweg konstant ist. Eine Voraussetzung ist dementsprechend, dass ε_{iv} strikt exogen ist.