Klausur zu Ökonometrie (Master)

Technische Universität Dortmund | Fakultät Wirtschaftswissenschaften | 1. Oktober 2025

Bitte tragen Sie Ihre Daten sorgfältig und leserlich ein:

Matrikelnummer ☐☐☐☐☐☐ Nachname _____

Studiengang _____ Vorname _____

Bearbeitungshinweise:

Diese Klausur besteht aus fünf Aufgaben, von welchen **vier Aufgaben Ihrer Wahl** zu bearbeiten sind.
Bearbeiten Sie alle Aufgaben, so werden nur die ersten vier bewertet.
Alle Antworten sind zu begründen.
Bitte verwenden Sie einen Kugelschreiber oder nicht zu starken Filzstift.
Für jede der fünf Aufgaben sind maximal je 18 Punkte zu erreichen.
Die Bearbeitungszeit beträgt 90 Minuten.

Erlaubte Hilfsmittel:

- Taschenrechner (nicht programmierbar)
- Ein DIN-A4 Blatt mit handschriftlichen Notizen (Vorder- und Rückseite)

Viel Erfolg!

Vom Prüfer auszufüllen:

Punkte Aufgabe 1 ☐ / 18

Punkte Aufgabe 2 ☐ / 18

Punkte Aufgabe 3 ☐ / 18

Punkte Aufgabe 4 ☐ / 18

Punkte Aufgabe 5 ☐ / 18

Gesamtpunkte ☐ / 72 **Note:**

Aufgabe 1

Es gelte die Parameter $\beta_0, \beta_1, \beta_2$ des Regressionsmodells („Modell 1“)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i,\ i = 1, \ldots, n$$

zu schätzen. Die Modellannahmen **MLR 1** bis **MLR 4** seien für Modell 1 erfüllt.

a) Benennen Sie diese vier Annahmen und geben Sie für eine dieser Annahmen Ihrer Wahl ein Beispiel, inwiefern diese Annahme bei tatsächlich erhobenen Daten verletzt sein könnte.

Die Variable x_{i2} sei leider nicht beobachtbar. Es würden nun die Parameter δ_0 und δ_1 des Modells („Modell 2“)

$$y_i = \delta_0 + \delta_1 x_{i1} + v_i$$

geschätzt, wobei $v_i = \beta_2 x_{i2} + u_i$ der unbeobachtete Störterm sei.

b) Sind in diesem Modell die Annahmen **MLR 1** bis **MLR 4** erfüllt?

c) Geben Sie den OLS-Schätzer $\hat{\delta}_1$ für δ_1 für Modell 2 an und bestimmen Sie dessen Erwartungswert.

Stellen Sie sich nun vor, Sie führen eine Studie durch, um den Einfluss von Bildung auf das Einkommen zu analysieren. Sie haben Daten über die Jahre an Schule und das jährliche Einkommen einer Gruppe von Personen gesammelt. Bei der Analyse stellen Sie fest, dass ein höherer Bildungsgrad mit einem höheren Einkommen korreliert ist.

d) Erläutern Sie, was unter „omitted variable bias“ zu verstehen ist.

e) Identifizieren Sie mindestens zwei Variablen, die möglicherweise nicht in Ihrem Modell berücksichtigt wurden und die sowohl das Einkommen als auch den Bildungsgrad beeinflussen könnten.

f) Diskutieren Sie, wie das Fehlen dieser Variablen Ihre Schätzung des Effekts von Bildung auf das Einkommen verzerren könnte.

Aufgabe 2

Betrachten Sie das lineare Regressionsmodell mit multiplen Regressoren:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \ ,$$

wobei $\mathbf{y} \in \mathbb{R}^n$ der Regressand, $\mathbf{X} \in \mathbb{R}^{n \times K+1}$ die Regressormatrix, $\boldsymbol{\beta} \in \mathbb{R}^{K+1}$ der Parametervektor und $\mathbf{u} \in \mathbb{R}^n$ der Störterm ist. Es gelten die Annahmen **MLR 1** bis **MLR 5**.

a) Wie lautet der OLS-Schätzer für $\boldsymbol{\beta}$ und dessen Erwartungswert und Varianz-Kovarianzmatrix?

b) Definieren Sie die Prognose $\hat{\mathbf{y}}$ und die Residuen $\hat{\mathbf{u}}$ und begründen Sie, warum $\sum_{i=1}^{n} \hat{u}_i = 0$ gilt.

c) Impliziert eine der Annahmen **MLR 1** bis **MLR 5**, dass jeder Regressor $\mathbf{x}_j$ mit dem Störterm $\mathbf{u}$ unkorreliert ist? Bestimmen Sie $Cov(x_{ij}, u_i) = \mathbb{E}[(x_{ij} - \mathbb{E}[x_{ij}])(u_i - \mathbb{E}[u_i])]$ für beliebige $j = 1, \ldots, K+1$ und $i = 1, \ldots, n$.

d) Verwenden Sie die Ausdrücke $\mathbf{X}'\mathbf{X}$ und $\mathbf{X}'\mathbf{y}$ um die notwendige Bedingung erster Ordnung des Minimierungsproblems für die kleinsten Quadrate Schätzer $\hat{\boldsymbol{\beta}}$ zu definieren.

e) Benutzen Sie diese Gleichung um zu begründen, dass $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$ gilt.

f) Wie ist der Zusammenhang der Aussage $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$ aus e) zu $\sum_{i=1}^{n} \hat{u}_i = 0$ aus b) und wie lassen sich diese Aussagen geometrisch interpretieren?

Aufgabe 3

Sie arbeiten empirisch an der Forschungsfrage: „Was macht eine erfolgreiche deutsche Getränkemarke aus?“ und analysieren dafür einen Datensatz mit einer linearen Regression:

$$\texttt{sales}_i = \beta_0 + \beta_1 \texttt{cost\_adv}_i + \beta_2 \texttt{price}_i + \beta_3 \texttt{sports\_drink}_i + \beta_4 \texttt{soft\_drink}_i + u_i \ ,$$

wobei $i = 1, \ldots, 80$ die erhobenen Marken sind. Die betrachteten Variablen sind:

- `sales` Absatz Deutschlandweit (in Liter) in 2024
- `cost_adv` Kosten für Werbung (in €) in 2024
- `price` Preisniveau (in €) in 2024
- `sports_drink` Dummyvariable. Ist 1 wenn die Marke ein Sportgetränk repräsentiert
- `soft_drink` Dummyvariable. Ist 1 wenn die Marke ein Softdrink repräsentiert.
- *Hinweis:* Alkoholika sind weder Sportgetränke noch Softdrinks.

Außerdem dürfen Sie annehmen, dass die klassischen MLR Annahmen (bis auf MLR 3) für diese Regression erfüllt ist.

a) Erklären Sie, was Multikollinearität bedeutet.

b) Welche Eigenschaft muss Ihr Datensatz erfüllen damit Sie bei der OLS Schätzung keine Probleme mit Multikollinearität bekommen?
Nehmen Sie bei Ihrer Antwort Bezug auf den konkreten Anwendungsfall.

Nehmen Sie für den Rest der Aufgabe an, dass keine exakte Multikollinearität, d.h. keine lineare Abhängigkeit der Regressoren, vorliegt.

Es bezeichne σ^2 die Varianz der homoskedastischen Störterme u_i, n den Stichprobenumfang, $s_{x_j x_j} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ die mittlere quadratische Abweichung des Regressors $\mathbf{x}_j$ und R_j^2 das Bestimmtheitsmaß der Regression des Regressors $\mathbf{x}_j$ auf die übrigen Regressoren.

c) Geben Sie die Varianz des OLS-Schätzers $\hat{\beta}_j$ eines Parameters β_j als Formel an.
Geben Sie vier Gründe an, warum diese Varianz hoch sein könnte.

d) Sehen Sie in der in Aufgabe 3 betrachteten Regression potential für eine hohe Varianz eines Schätzers? Begründen Sie Ihre Aussage.

Sie möchten testen, ob Sportgetränke genauso erfolgreich sind wie Softdrinks, also die Hypothese, dass:

$$H_0 : \beta_3 = \beta_4 \quad H_1 : \beta_3 \neq \beta_4$$

e) Beschreiben Sie, wie Sie obige Hypothese anhand eines t-Tests testen können.

f) Angenommen, Sie wollen die Hypothese $\beta_3 = \beta_4$ stattdessen anhand eines F-Tests testen und formulieren die Nullhypothese $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. Nennen Sie $\mathbf{R}$, $\mathbf{r}$, die Anzahl $n - K - 1$ der Freiheitsgrade und die Anzahl J der Restriktionen.

Aufgabe 4

Betrachten Sie folgende Regression (für welche die Annahmen **MLR 1** bis **MLR 6** erfüllt seien):

$$\underset{}{\widehat{\mathrm{y}}} = \underset{(0{,}91037)}{10{,}2765} + \underset{(10{,}595)}{8{,}43804}\,\mathrm{x}_1 - \underset{(7{,}7238)}{4{,}51749}\,\mathrm{x}_2 - \underset{(7{,}9037)}{2{,}87635}\,\mathrm{x}_3$$

(Standardfehler in Klammern)

Der Stichprobenumfang betrage $n = 100$, die Standardabweichung von y betrage

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 14{,}03017$$

und der Standardfehler der Regression betrage

$$\hat{\sigma} = \sqrt{\frac{1}{n-K-1}\sum_{i=1}^{n}\hat{u}_i^2} = 8{,}948813\ .$$

Runden Sie bei allen Berechnungen bitte auf zwei Nachkommastellen.

a) Sie möchten testen, ob für einen der drei Regressoren x_j, $j = 1, 2, 3$ die Nullhypothese gilt:

$$H_0:\ \beta_j = 0\ ,\ H_1:\ \beta_j \neq 0$$

Erläutern Sie, wie Sie diese Hypothese valide mit $\alpha = 5\%$ testen können und führen Sie den Test für einen der drei Regressoren durch. Begründen Sie Ihre Testentscheidung.

b) Geben Sie das symmetrische Konfidenzintervall für ein β_j, $j = 1, 2, 3$ mit $\alpha = 5\%$ an (nicht für alle drei).

c) Formulieren Sie die Nullhypothese des globalen F-Tests und erläutern sie diesen.

d) Wie lautet die Summe der quadrierten Residuen SSR und die „total sum of squares“ SST?

e) Wie lautet das Bestimmtheitsmaß R^2?

f) Führen Sie den globalen F-Test für obige Regression durch und begründen Sie Ihre Testentscheidung.

Aufgabe 5

In einer Studie möchten Sie den Einfluss von Bildung x auf das Einkommen y untersuchen. Es wird jedoch angenommen, dass Bildung endogen ist, da unberücksichtigte Faktoren wie Motivation oder familiärer Hintergrund sowohl die Bildung als auch das Einkommen beeinflussen können. Um dieses Problem zu lösen, entscheiden Sie sich für die Anwendung der Methode der Instrumentalvariablen.

a) Erläutern Sie, was unter dem Begriff „Instrumentalvariable“ zu verstehen ist und welche Bedingungen erfüllt sein müssen, damit eine Variable als Instrument betrachtet werden kann.

b) Nehmen Sie an, dass die Anzahl der Geschwister z als Instrument für die Bildung verwendet wird. Diskutieren Sie, ob diese Variable die notwendigen Bedingungen erfüllt und begründen Sie Ihre Antwort.

c) Skizzieren Sie den Ansatz zur Schätzung des Einflusses von Bildung auf das Einkommen mithilfe der Methode der Instrumentalvariablen. Welche Schritte sind dabei notwendig?

d) Angenommen, die geschätzte Gleichung lautet $y = \beta_0 + \beta_1 x + u$. Die Schätzung mit OLS ergibt $\hat{\beta}_1 = 0,5$. Wenn Sie nun eine IV-Schätzung durchführen und $\hat{\beta}_1^{IV} = 0,8$ erhalten, interpretieren Sie diesen Unterschied im Kontext der Endogenität.

e) Angenommen, nach Durchführung einer IV-Schätzung stellen Sie fest, dass Ihr Instrument z nicht signifikant mit x korreliert ist (p-Wert > 0,05). Erklären Sie die Implikationen dieser Beobachtung für Ihre Schätzung von β_1 und diskutieren Sie mögliche Schritte zur Verbesserung Ihrer Analyse.

f) Stellen Sie dar, welche möglichen Probleme bei der Verwendung von z als Instrument auftreten können und wie diese Probleme die Schätzung beeinflussen könnten.

Kritische Werte der t-Verteilung

		Signifikanzniveau				
einseitig:		10%	5%	2,5%	1%	0,5%
zweiseitig:		20%	10%	5%	2%	1%
Freiheits-	1	3,078	6,314	12,706	31,821	63,657
grade	2	1,886	2,920	4,303	6,965	9,925
	3	1,638	2,353	3,182	4,541	5,841
	4	1,533	2,132	2,776	3,747	4,604
	5	1,476	2,015	2,571	3,365	4,032
	6	1,440	1,943	2,447	3,143	3,707
	7	1,415	1,895	2,365	2,998	3,499
	8	1,397	1,860	2,306	2,896	3,355
	9	1,383	1,833	2,262	2,821	3,250
	10	1,372	1,812	2,228	2,764	3,169
	11	1,363	1,796	2,201	2,718	3,106
	12	1,356	1,782	2,179	2,681	3,055
	13	1,350	1,771	2,160	2,650	3,012
	14	1,345	1,761	2,145	2,624	2,977
	15	1,341	1,753	2,131	2,602	2,947
	16	1,337	1,746	2,120	2,583	2,921
	17	1,333	1,740	2,110	2,567	2,898
	18	1,330	1,734	2,101	2,552	2,878
	19	1,328	1,729	2,093	2,539	2,861
	20	1,325	1,725	2,086	2,528	2,845
	21	1,323	1,721	2,080	2,518	2,831
	22	1,321	1,717	2,074	2,508	2,819
	23	1,319	1,714	2,069	2,500	2,807
	24	1,318	1,711	2,064	2,492	2,797
	25	1,316	1,708	2,060	2,485	2,787
	26	1,315	1,706	2,056	2,479	2,779
	27	1,314	1,703	2,052	2,473	2,771
	28	1,313	1,701	2,048	2,467	2,763
	29	1,311	1,699	2,045	2,462	2,756
	30	1,310	1,697	2,042	2,457	2,750
	40	1,303	1,684	2,021	2,423	2,704
	60	1,296	1,671	2,000	2,390	2,660
	90	1,291	1,662	1,987	2,368	2,632
	120	1,289	1,658	1,980	2,358	2,617
	∞	1,282	1,645	1,960	2,326	2,576

Kritische Werte der F-Verteilung zum Signifikanzniveau von 1%

		Anzahl der Restriktionen									
		1	2	3	4	5	6	7	8	9	10
$n-k-1$	10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
	11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
	12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
	13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
	14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
	15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
	16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
	17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
	18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
	19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
	20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
	21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
	22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
	23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
	24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
	25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
	26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
	27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
	28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
	29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
	30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
	40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
	60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
	90	6,93	4,85	4,01	3,54	3,23	3,01	2,84	2,72	2,61	2,52
	120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
	∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Kritische Werte der F-Verteilung zum Signifikanzniveau von 5%

		Anzahl der Restriktionen									
		1	2	3	4	5	6	7	8	9	10
$n-k-1$	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
	11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
	13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
	14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
	16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
	17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
	18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
	19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
	21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
	22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
	23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
	24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
	25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
	26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
	27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
	28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
	29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
	90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94
	120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
	∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Kritische Werte der χ^2-Verteilung

		Signifikanzniveau		
		10%	5%	1%
Freiheits-	1	2,71	3,84	6,63
Grade	2	4,61	5,99	9,21
	3	6,25	7,81	11,34
	4	7,78	9,49	13,28
	5	9,24	11,07	15,09
	6	10,64	12,59	16,81
	7	12,02	14,07	18,48
	8	13,36	15,51	20,09
	9	14,68	16,92	21,67
	10	15,99	18,31	23,21
	11	17,28	19,68	24,72
	12	18,55	21,03	26,22
	13	19,81	22,36	27,69
	14	21,06	23,68	29,14
	15	22,31	25,00	30,58
	16	23,54	26,30	32,00
	17	24,77	27,59	33,41
	18	25,99	28,87	34,81
	19	27,20	30,14	36,19
	20	28,41	31,41	37,57
	21	29,62	32,67	38,93
	22	30,81	33,92	40,29
	23	32,01	35,17	41,64
	24	33,20	36,42	42,98
	25	34,38	37,65	44,31
	26	35,56	38,89	45,64
	27	36,74	40,11	46,96
	28	37,92	41,34	48,28
	29	39,09	42,56	49,59
	30	40,26	43,77	50,89

Lösung für Aufgabe 1

a)

- MLR 1, Linearität in $\boldsymbol{\beta}$: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
  Verletzung: $y_i = \beta_0 + x_{i1}^{\beta_1} + u_i$
- MLR 2 Zufallsstichprobe $\mathbf{X}, \mathbf{y}$
  Verletzung: $\mathbf{X}$ oder $\mathbf{y}$ konstant
- MLR 3 Linear unabhängige Regressoren $rk(\mathbf{X}) = K + 1$
  Verletzung: Dummyvariablen $\mathbf{x}_1$ und $\mathbf{x}_2$ mit $x_{i1} + x_{i2} = 1$ für alle $i = 1, \ldots, n$
- MLR 4 Strikte Exogenität $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbb{E}[\mathbf{u}] = \mathbf{0}$
  Verletzung: ausgelassene Variable, die mit vorhandenem Regressor korreliert.

b) Mit $v_i = \beta_2 x_{i2} + u_i$ gilt $\mathbb{E}[(x_{i1} - \bar{x}_1)v_i] = \mathbb{E}[(x_{i1} - \bar{x}_1)(\beta_2 x_{i2} + u_i)] = \beta_2 \mathbb{E}[(x_{i1} - \bar{x}_1)x_{i2}] + \mathbb{E}[(x_{i1} - \bar{x}_1)u_i]$. Wegen MLR 4 in Modell 1 gilt $\mathbb{E}[(x_{i1} - \bar{x}_1)u_i] = 0$. Annahme MLR 4 ist also in Modell 2 verletzt, falls $\beta_2 \neq 0$ und $Cov(x_1, x_2) = \mathbb{E}[(x_{i1} - \bar{x}_1)x_{i2}] \neq 0$.

c) $\hat{\delta}_1 = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)y_i}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$. Mit $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$ gilt:

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2)} = \beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)x_{i2}}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} + \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)u_i}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

Also gilt

$$\mathbb{E}[\hat{\delta}_1] = \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)}$$

d) Der omitted variable bias ist eine Verzerrung, welche auftreten kann, wenn bei der OLS-Schätzung eine erklärenden Variable ausgelassen wird.

e) Das Talent oder Beziehungen der entsprechenden Person können ebenfalls einen Einfluss auf das Einkommen haben.

f) Wenn Talent positiv mit Bildung korreliert und Talent ein Teil der unbeobachteten Störterme ist, so ist also Bildung mit den unbeobachteten Störtermen korreliert. Damit ist schwache Exogenität und damit auch strikte Exogenität der Regressoren verletzt. Dies ist aber eine Voraussetzung für unverzerrte OLS-Schätzer. Wenn zusätzlich Talent positiv mit dem Einkommen zusammenhängt, überschätzt der Schätzer $\hat{\delta}_1$ den tatsächlichen Wert β_1.

Lösung zu Aufgabe 2

a) Der OLS-Schätzer lautet $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; es gilt $\mathbf{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ und $\mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

b) $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ und $\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$
Begründung 1: Es gilt $\sum_{i=1}^{n} \hat{u}_i = \boldsymbol{\iota}'\hat{\mathbf{u}}$. Wegen $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = [\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = [\mathbf{X}' - \mathbf{X}']\mathbf{y} = \mathbf{N}$ gilt auch $\boldsymbol{\iota}'\hat{\mathbf{u}} = 0$, da $\boldsymbol{\iota}'$ die erste Zeile von $\mathbf{X}'$ ist.
Begründung 2: Der OLS-Schätzer $\hat{\boldsymbol{\beta}}$ erfüllt die Bedingungen erster Ordnung $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$. Mit $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$ gilt also $\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\mathbf{y} \Leftrightarrow \mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = 0 \Leftrightarrow \mathbf{X}'\hat{\mathbf{u}} = 0$.
Begründung 3:

$$\begin{aligned} \sum_{i=1}^{n} \hat{y}_i &= \sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_K x_{iK} \\ \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} + \ldots + \hat{\beta}_K \frac{1}{n}\sum_{i=1}^{n} x_{iK} \\ \Leftrightarrow \bar{\hat{y}} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_K \bar{x}_K \end{aligned}$$

Da außerdem gilt $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_K \bar{x}_K$, woraus folgt $\bar{y} = \bar{\hat{y}} \Leftrightarrow \bar{y} - \bar{\hat{y}} = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i = 0 \Leftrightarrow \sum_{i=1}^{n} \hat{u}_i = 0$.

c) Annahme **MLR 4**: $\mathbf{E}[\mathbf{u}|\mathbf{x}] = 0$ (strikte Exogenität) impliziert schwache Exogenität. Begründung: $Cov(x_{ij}, u_i) = \mathbf{E}[(x_{ij} - \mathbf{E}[x_{ij}])u_i] = \mathbf{E}_{\mathbf{x}}\big[(x_{ij} - \mathbf{E}[x_{ij}])\underbrace{\mathbf{E}_u[u_i|\mathbf{X}]}_{=0}\big] = 0$

d)

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

e) Mit $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ ist die Gleichung aus d) identisch zu

$$\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\mathbf{y} \Leftrightarrow \mathbf{X}'(\hat{\mathbf{y}} - \mathbf{y})$$

Mit $\hat{\mathbf{u}} = \hat{\mathbf{y}} - \mathbf{y}$ folgt $\mathbf{X}'\hat{\mathbf{u}} = 0$.

f) Die erste Spalte von $\mathbf{X}$ lautet $\boldsymbol{\iota}$. Damit entspricht die erste Zeile von $\mathbf{X}'\hat{\mathbf{u}}$: $\boldsymbol{\iota}'\hat{\mathbf{u}} = \sum_{i=1}^{n} 1 \cdot \hat{u}_i = \sum_{i=1}^{n} \hat{u}_i$.
Geometrische Interpretation: Das Skalarprodukt jeder Zeile von $\mathbf{X}'$, also das Skalarprodukt jeder Spalte von $\mathbf{X}$ und dem Vektor der Residuen $\hat{\mathbf{u}}$ ist gleich null. Damit ist der Vektor der Residuen senkrecht (orthogonal) zu jeder Spalte der Regressormatrix. Der Vektor der Residuen liegt damit im orthogonalen Komplement des Spaltenraums der Regressormatrix, $\hat{\mathbf{u}} \in \mathcal{R}^{\perp}(\mathbf{X})$.

Lösung zu Aufgabe 3

a) Multikollinearität bedeutet, dass zwei oder mehr Regressoren stark korrelieren.

b) Dummyfalle! In dem Datensatz müssen Marken vorkommen, die kein Sportgetränk und auch keinen Softdrink repräsentieren (Wassermarken, Milchmarken, Biermarken etc.).

c)

$$Var(\hat{\beta}_j) = \sigma^2 \frac{1}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \frac{1}{1 - R_j^2} \ ,$$

wobei R_j^2 das Bestimmtheitsmaß der Regression von $\mathbf{x}_j$ auf die übrigen Regressoren ist.

d) Korrelation zwischen Regressoren: Man könnte vermuten, dass zum Beispiel `price` und `cost_adv` korrelieren. Dann wäre die entsprechenden Bestimmtheitsmaße R_1^2 und R_2^2 nahe 1. Daher erhöht sich die Varianz für beide Schätzungen $(\hat{\beta}_1, \hat{\beta}_2)$.

e) Zunächst addieren wir $\beta_1 \texttt{price}_i$ zu der Regression und subtrahieren es wieder:

$$\texttt{sales}_i = \beta_0 + \beta_1 \texttt{cost\_adv}_i + \beta_1 \texttt{price} - \beta_1 \texttt{price} + \beta_2 \texttt{price}_i + \beta_3 \texttt{sports\_drink}_i + \beta_4 \texttt{soft\_drink}_i + u_i$$

Wir erhalten dann den neuen Regressor $z_i = \texttt{cost\_adv}_i + \texttt{price}$ und den neuen Parameter $\theta = \beta_2 - \beta_1$:

$$\texttt{sales}_i = \beta_0 + \beta_1 z_i + \theta \texttt{price}_i + \beta_3 \texttt{sports\_drink}_i + \beta_4 \texttt{soft\_drink}_i + u_i$$

Nun können wir die Hypothese $H_0 : \ \theta = 0$ anhand eines t-Tests testen.

f)

$$\begin{aligned} R &= \begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix} \\ r &= 0 \\ K &= 4 \\ \Rightarrow n - K - 1 &= 80 - 4 - 1 = 75 \\ J &= 1 \end{aligned}$$

Lösung zu Aufgabe 4

a) t-Test: $t_j = \frac{\hat{\beta}_j - \gamma}{se(\hat{\beta}_j)}$ t-verteilt. Lehne $H_0: \ \beta_j = \gamma$ ab, falls $|t_j| > c_{t_{n-K-1,1-\frac{\alpha}{2}}}$.
Hier:

$$\begin{aligned} t_1 &= \frac{8{,}43804}{10{,}595} = 0{,}7964 \\ t_2 &= \frac{-4{,}51749}{7{,}7238} = -0{,}5849 \\ t_3 &= \frac{-2{,}87635}{7{,}90365} = -0{,}3639 \end{aligned}$$

Der kritische Wert für einen zweiseitigen Test zum Niveau $\alpha = 5\%$ für eine t-verteilte Zufallsvariable mit 96 Freiheitsgraden lautet $c_{96,0.975} \approx 1{,}987$. Für alle $j = 1, 2, 3$ gilt $|t_j| < 1{,}987$. Daher kann $H_0: \ \beta_j = 0$ für keines der $j = 1, 2, 3$ abgelehnt werden.

b) Das symmetrische Konfidenzintervall für $\alpha = 5\%$ lautet im Allgemeinen

$$\left[\hat{\beta}_j - c_{96,0.975} \cdot se(\hat{\beta}_j), \hat{\beta}_j + c_{96,0.975} \cdot se(\hat{\beta}_j)\right]$$

Für β_1:

$$[8{,}43804 - 1{,}987 \cdot 10{,}595; 8{,}43804 + 1{,}987 \cdot 10{,}595] = [-12{,}61; 29{,}49]$$

Für β_2:

$$[-4{,}51749 - 1{,}987 \cdot 7{,}7238; -4{,}51749 + 1{,}987 \cdot 7{,}7238] = [-24{,}38; 15{,}35]$$

Für β_3:

$$[-2{,}87635 - 1{,}987 \cdot 7{,}9037; -2{,}87635 + 1{,}987 \cdot 7{,}9037] = [-18{,}58; 12{,}83]$$

c) Der globale F-Test prüft die Nullhypothese

$$H_0: \ \beta_1 = \ldots = \beta_{K+1} = 0$$

Die Teststatistik des globalen F-Tests ist definiert durch

$$F = \frac{(SST - SSR)/K}{SSR/(n-K-1)} = \frac{R^2/K}{(1-R^2)/(n-K-1)}$$

Falls $F > c_{F_{n-K-1,K}}$, wird die Nullhypothese abgelehnt.

d) Der Standardfehler der Regression ist definiert durch

$$\hat{\sigma} = \sqrt{\frac{1}{n-K-1}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{1}{n-K-1}SSR}$$

Laut Output ist $n = 100$, $K = 3$ und $SSR = 7687,8$. Demnach gilt:

$$SSR = \sum_{i=1}^{n}\hat{u}_i^2 = 96 \cdot \hat{\sigma}^2 = 96 \cdot 8{,}948813^2 = 7687,80$$

Die Standardabweichung der abhängigen Variablen ist definiert durch

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{n-1}SST}$$

Demnach gilt:

$$SST = (s_y)^2 (n-1) = 14,03017^2 \cdot 99 = 19487,72$$

e) Das Bestimmtheitsmaß ist definiert durch

$$R^2 = 1 - \frac{SSR}{SST}$$

Demnach gilt hier:

$$R^2 = 1 - \frac{7687{,}8}{19487{,}72} = 1 - 0{,}394495 = 0{,}61$$

f) Der Wert Teststatistik lautet hier:

$$F = \frac{(19487,72 - 7687,80)/3}{7687{,}8/96} = 49,12$$

oder

$$F = \frac{0{,}61/3}{(1-0{,}61)/96} = 50{,}05$$

Mit $c_{F_{n-K-1,K}} = 2,71$ ist der kritische Wert deutlich unter dem Wert der Teststatistik und die Nullhypothese $H_0: \ \beta_1 = \beta_2 = \beta_3 = 0$ kann abgelehnt werden.

Vollständiger Output der Regressionsanalyse:

Modell 1: KQ, benutze die Beobachtungen 1–100
Abhängige Variable: y

	Koeffizient	Std. Fehler	t-Quotient	p-Wert
const	10,2765	0,910371	11,29	0,0000
x1	8,43804	10,5952	0,7964	0,4278
x2	−4,51749	7,72378	−0,5849	0,5600
x3	−2,87635	7,90365	−0,3639	0,7167

Mittel abhängige Var.	11,35109	Stdabw. abhängige Var.	14,03017
Summe quad. Residuen	7687,800	Stdfehler Regression	8,948813
R^2	0,605505	Korrigiertes R^2	0,593178
$F(3, 96)$	49,11646	P-Wert(F)	2,51e–19

Lösung zu Aufgabe 5

a) Eine Instrumentalvariable (IV) ist eine Variable, die verwendet wird, um das Problem der Endogenität in einem Regressionsmodell zu lösen. Damit eine Variable als Instrument betrachtet werden kann, müssen zwei Bedingungen erfüllt sein:

- Relevanz: Die Instrumentalvariable z muss signifikant mit der erklärenden Variablen x korreliert sein. Das bedeutet, dass z einen Einfluss auf x haben sollte.
- Exogenität: Die Instrumentalvariable z darf nicht direkt mit dem Fehlerterm u korreliert sein. Dies stellt sicher, dass z keinen direkten Einfluss auf die abhängige Variable y hat, außer über x.

b) Die Anzahl der Geschwister z kann als potenzielle Instrumentalvariable für Bildung x betrachtet werden.

- Relevanz: Es gibt Hinweise darauf, dass die Anzahl der Geschwister einen Einfluss auf Bildungsentscheidungen haben kann (z.B. Ressourcenverteilung innerhalb einer Familie).
- Exogenität: Allerdings könnte die Anzahl der Geschwister auch mit anderen unberücksichtigten Faktoren korreliert sein, wie z.B. dem sozioökonomischen Status oder den Erziehungsstilen der Eltern. Daher könnte es problematisch sein, diese Variable als exogen anzunehmen.

c) Der Ansatz zur Schätzung des Einflusses von Bildung auf Einkommen mithilfe von IV umfasst folgende Schritte:

1. Überprüfen Sie die Relevanz und Exogenität des Instruments.
2. Schätzen Sie zunächst ein Regressionsmodell für x (Bildung) unter Verwendung von z (Anzahl der Geschwister) als Instrument: $x = \pi_0 + \pi_1 z + v$.
3. Verwenden Sie die geschätzten Werte von x ($\hat{x}$) aus dem ersten Schritt und schätzen Sie dann das endgültige Modell für y: $y = \beta_0 + \beta_1 \hat{x} + u$.
4. Interpretieren Sie das Ergebnis und prüfen Sie die Robustheit Ihrer Schätzungen.

d) Der Unterschied zwischen $\hat{\beta}_1 = 0,5$ aus OLS und $\hat{\beta}_{IV} = 0,8$ aus IV deutet darauf hin, dass die OLS-Schätzung möglicherweise nach unten verzerrt ist aufgrund von Endogenitätsproblemen (z.B. omitted variable bias). Die IV-Schätzung liefert einen höheren Wert für den Einfluss von Bildung auf das Einkommen und legt nahe, dass bei Berücksichtigung endogener Effekte ein stärkerer positiver Zusammenhang besteht.

e) Wenn das Instrument z nicht signifikant mit x korreliert ist (p-Wert > 0,05), hat dies schwerwiegende Implikationen für die Schätzung von β_1:

- Es weist darauf hin, dass das gewählte Instrument möglicherweise ungeeignet ist und somit keine relevante Information zur Erklärung von x liefert.
- Dies kann zu einer schwachen Identifikation führen und dazu führen, dass unsere IV-Schätzungen ungenau oder sogar irreführend sind.
- Mögliche Schritte zur Verbesserung könnten beinhalten: Suche nach alternativen Instrumenten mit stärkerer Korrelation zu x oder Durchführung zusätzlicher Tests zur Validierung des Instruments.

f) Mögliche Probleme bei der Verwendung von z als Instrument können Folgendes umfassen:

- Wenn z nicht ausreichend relevant ist (schwache Instrumente), können die Schätzungen ungenau werden.
- Wenn z tatsächlich mit dem Fehlerterm u korreliert ist (Verletzung der Exogenitätsannahme), führt dies zu verzerrten Schätzungen.
- Ein weiteres Problem könnte Heteroskedastizität oder Multikollinearität in den Daten sein, was ebenfalls die Schätzgenauigkeit beeinträchtigen kann.