Klausur zu Ökonometrie (Master)

Technische Universität Dortmund | Fakultät Wirtschaftswissenschaften | 2. Oktober 2026

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, geben Sie bei Rechenaufgaben Ihren Rechenweg an.
Bitte verwenden Sie einen Kugelschreiber oder einen nicht zu stark durchdrückenden Filzstift.
Für jede der fünf Aufgaben sind maximal je 18 Punkte zu erreichen.
Die Bearbeitungszeit beträgt 90 Minuten.

Erlaubte Hilfsmittel:

- Taschenrechner (nicht programmierbar)

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 zu Kapitel 3 Multiple Lineare Regression

Eine Forscherin untersucht den Zusammenhang zwischen dem monatlichen Nettoeinkommen y, der Anzahl der Bildungsjahre x und der Berufserfahrung z. Sie betrachtet das Modell

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i, \qquad i = 1, \ldots, n.$$

Für $n = 6$ Beobachtungen liegen folgende beobachtete Werte und Prognosen vor:

$$\mathbf{y} = \begin{pmatrix} 4 \\ 5 \\ 6 \\ 7 \\ 9 \\ 11 \end{pmatrix}, \qquad \hat{\mathbf{y}} = \begin{pmatrix} 4{,}5 \\ 5{,}0 \\ 6{,}5 \\ 7{,}0 \\ 8{,}5 \\ 10{,}5 \end{pmatrix}.$$

Zusätzlich ist bekannt, dass

$$SST = 34.$$

a) Schreiben Sie das Modell in Matrixnotation. Geben Sie außerdem die Dimensionen von $\mathbf{X}$ und $\boldsymbol{\beta}$ an.

b) Formulieren Sie die Annahme des vollen Spaltenrangs der Regressormatrix $\mathbf{X}$ sowohl in Worten als auch mathematisch.

c) Es gelte für alle Beobachtungen $i = 1, \ldots, n$:

$$x_i + z_i = 10.$$

Erläutern Sie kurz, warum in diesem Fall die Annahme des vollen Spaltenrangs verletzt ist.

d) Angenommen, Berufserfahrung $\mathbf{z}$ wird nicht beobachtet und daher im geschätzten Modell weggelassen. Gleichzeitig sind $\mathbf{x}$ und $\mathbf{z}$ korreliert. Die Annahme $x_i + z_i = 10$ aus c) gelte weiterhin. Welche Modellannahme ist dadurch möglicherweise verletzt? Begründen Sie Ihre Antwort kurz.

e) Berechnen Sie den Residuenvektor $\hat{\mathbf{u}}$ und die Summe der quadrierten Residuen SSR.

f) Berechnen Sie das Bestimmtheitsmaß R^2 sowie den unverzerrten Schätzer $\hat{\sigma}^2$ für die Varianz der Störterme.

Aufgabe 2 zu Kapitel 4 Inferenz

Ein Unternehmen untersucht, welche Faktoren den monatlichen Umsatz $sales_i$ seiner Filialen beeinflussen. Das unrestringierte Modell lautet:

$$sales_i = \beta_0 + \beta_1 advertising_i + \beta_2 employees_i + \beta_3 location_i + u_i.$$

Dabei misst $advertising_i$ die Werbeausgaben, $employees_i$ die Anzahl der Beschäftigten und $location_i$ die Attraktivität des Standorts.

Für $n = 44$ Filialen ergeben sich die folgenden Schätzergebnisse:

$$SSR_u = 720, \qquad R_u^2 = 0{,}64.$$

Für ein restringiertes Modell, in dem die Variablen $advertising_i$ und $location_i$ nicht enthalten sind, gilt:

$$SSR_r = 900, \qquad R_r^2 = 0{,}55.$$

a) Wie viele Regressoren K enthält das unrestringierte Modell? Bestimmen Sie die Freiheitsgrade im Nenner der F-Statistik.

b) Formulieren Sie die Nullhypothese und die Alternativhypothese, die durch den Vergleich des restringierten und des unrestringierten Modells überprüft werden.

c) Wie viele Restriktionen J werden in diesem Test überprüft? Begründen Sie Ihre Antwort kurz.

d) Berechnen Sie die F-Statistik auf Grundlage der Summen der quadrierten Residuen.

e) Treffen Sie zum Signifikanzniveau $\alpha = 5\%$ eine Testentscheidung. Interpretieren Sie das Ergebnis im Kontext der Erklärung des Umsatzes.

f) Testen Sie nun mit einem globalen F-Test, ob die drei Regressoren im unrestringierten Modell gemeinsam keinen Erklärungsgehalt besitzen. Formulieren Sie zunächst die Nullhypothese, berechnen Sie anschließend die globale F-Statistik und treffen Sie zum Signifikanzniveau $\alpha = 5\%$ eine Entscheidung.

Aufgabe 3 zu Kapitel 8 Heteroskedastizität

Ein Unternehmen untersucht den Zusammenhang zwischen Werbeausgaben x_i (in Tausend Euro) und Umsatz y_i (in Tausend Euro). Es wird das Modell

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

betrachtet.

Für drei Filialen liegen die folgenden Daten vor:

i	1	2	3
x_i	1	2	3
y_i	2	3	5
σ_i	1	2	1

Dabei ist σ_i die bekannte Standardabweichung des Störterms der jeweiligen Beobachtung. Verwenden Sie für die Transformation

$$\tilde{y}_i = \frac{y_i}{\sigma_i}, \qquad \tilde{x}_{i0} = \frac{1}{\sigma_i}, \qquad \tilde{x}_{i1} = \frac{x_i}{\sigma_i}.$$

Für die transformierten Daten gilt:

$$\tilde{X}'\tilde{X} = \begin{pmatrix} 2{,}25 & 4{,}50 \\ 4{,}50 & 11{,}00 \end{pmatrix}, \qquad \tilde{X}'\tilde{y} = \begin{pmatrix} 7{,}75 \\ 18{,}50 \end{pmatrix}.$$

a) Erläutern Sie kurz, was unter Heteroskedastizität verstanden wird. Gehen Sie dabei auf die bedingten Varianzen der Störterme ein.

b) Welche Annahme des klassischen multiplen linearen Regressionsmodells ist bei Heteroskedastizität verletzt?

c) Berechnen Sie die transformierten Werte $\tilde{y}_i$, $\tilde{x}_{i0}$ und $\tilde{x}_{i1}$ für alle drei Beobachtungen.

d) Erläutern Sie, warum Beobachtung $i = 2$ in der GLS-Schätzung ein geringeres Gewicht erhält als die Beobachtungen $i = 1$ und $i = 3$.

e) Berechnen Sie den GLS-Schätzer $\tilde{\beta}_1$.

f) Interpretieren Sie den geschätzten Koeffizienten $\tilde{\beta}_1$ im Kontext der Aufgabe.

Aufgabe 4 zu
Kapitel 12 Serielle Korrelation in Zeitreihen

Eine Forscherin untersucht den Zusammenhang zwischen dem Konsum y_t und dem verfügbaren Einkommen x_t in $n = 5$ Jahren:

$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

Die Störterme folgen einem AR(1)-Prozess:

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad \rho = 0{,}5.$$

Für die beobachteten Daten gilt:

t	1	2	3	4	5
y_t	10	13	15	18	20
x_t	2	3	4	5	6

a) Erläutern Sie kurz, was unter serieller Korrelation der Störterme in einer Zeitreihe verstanden wird.

b) Geben Sie die Bedingungen $E[u_t]$ und $Var(u_t)$ an, unter denen ein AR(1)-Prozess stationär ist. Welche Bedingung an ρ wird durch Stationärität impliziert?

c) Bestimmen Sie den Erwartungswert $E[u_t]$ und die Varianz $\mathrm{Var}(u_t)$, falls $\mathrm{Var}(\varepsilon_t) = 3$ gilt.

d) Berechnen Sie $\mathrm{Cov}(u_t, u_{t-1})$, wenn $\rho = 0{,}5$ und $\mathrm{Var}(u_t) = 4$ gilt.

e) Führen Sie für $t = 2, \ldots, 5$ die Quasidifferenzierung mit $\rho = 0{,}5$ durch.

f) Erläutern Sie kurz, warum die Quasidifferenzierung bei einem AR(1)-Prozess hilfreich für die Schätzung des Regressionsmodells ist. Gehen Sie dabei auf den transformierten Störterm ein.

Aufgabe 5 zu Kapitel 15 Instrumente

Eine Forscherin untersucht den kausalen Effekt der Ausbildungsdauer auf den logarithmierten Stundenlohn. Das strukturelle Modell lautet:

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + u_i.$$

Dabei bezeichnet $educ_i$ die Anzahl der Bildungsjahre einer Person. Die Forscherin vermutet, dass unbeobachtete Fähigkeiten einer Person sowohl die Ausbildungsdauer als auch den Lohn beeinflussen.

Als mögliche Instrumentalvariable für $educ_i$ wird die Entfernung $distance_i$ zwischen dem Wohnort der Person im Alter von 16 Jahren und der nächstgelegenen Universität vorgeschlagen.

Für eine Stichprobe liegen die folgenden empirischen Kovarianzen vor:

$$\widehat{\mathrm{Cov}}(distance, educ) = -1{,}20, \qquad \widehat{\mathrm{Cov}}(distance, \ln(wage)) = -0{,}096.$$

Die OLS-Schätzung des ursprünglichen Modells ergibt:

$$\widehat{\beta}_1^{OLS} = 0{,}10.$$

a) Erläutern Sie, wann ein Regressor als endogen bezeichnet wird. Formulieren Sie Ihre Antwort mithilfe einer Kovarianz.

b) Begründen Sie kurz, warum die Variable $educ_i$ im vorliegenden Modell möglicherweise endogen ist.

c) Nennen Sie die beiden Bedingungen, die eine Variable erfüllen muss, um ein gültiges Instrument für den endogenen Regressor $educ_i$ zu sein.

d) Erläutern Sie, warum die Entfernung zur nächstgelegenen Universität möglicherweise ein relevantes Instrument für $educ_i$ sein könnte.

e) Berechnen Sie den IV-Schätzer $\widehat{\beta}_1^{IV}$ für den Parameter β_1.

f) Vergleichen Sie den IV-Schätzer aus Teilaufgabe e) mit dem gegebenen OLS-Schätzer. Welcher Schätzer ist zur Bestimmung eines kausalen Effekts geeigneter, falls $educ_i$ tatsächlich endogen ist? Begründen Sie Ihre Antwort kurz.

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

Musterlösung

Für jede Teilaufgabe werden 3 Punkte vergeben, sodass jede Aufgabe insgesamt 18 Punkte umfasst.

Aufgabe 1 – Multiple lineare Regression

Gegeben ist das Modell

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i, \qquad i = 1, \dots, n.$$

a) Das Modell in Matrixnotation lautet:

$$y = X\beta + u.$$

Dabei gilt:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_6 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_6 & z_6 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_6 \end{pmatrix}.$$

Da $n = 6$ Beobachtungen und zwei Regressoren neben der Konstanten vorliegen, gilt:

$$X \in \mathbb{R}^{6\times 3}, \qquad \beta \in \mathbb{R}^3.$$

b) Die Annahme des vollen Spaltenrangs besagt, dass alle Spalten der Regressormatrix X linear unabhängig sind.

Mathematisch lautet die Annahme:

$$\mathrm{rk}(X) = K + 1.$$

Da das Modell zwei Regressoren enthält, gilt $K = 2$ und somit:

$$\mathrm{rk}(X) = 3.$$

c) Es gelte:

$$x_i + z_i = 10$$

für alle Beobachtungen i.

In Vektorschreibweise gilt damit:

$$x + z = 10\iota.$$

Folglich lässt sich beispielsweise der Regressor z als Linearkombination der übrigen Spalten darstellen:

$$z = 10\iota - x.$$

Die Spalten ι, x und z sind daher linear abhängig.

Folglich hat X keinen vollen Spaltenrang:

$$\mathrm{rk}(X) < 3.$$

d) Die Annahme aus Teilaufgabe c) gelte weiterhin, also:

$$x_i + z_i = 10 \qquad \Leftrightarrow \qquad z_i = 10 - x_i.$$

Das ursprüngliche Modell kann damit umgeschrieben werden zu:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i \\ &= \beta_0 + \beta_1 x_i + \beta_2 (10 - x_i) + u_i \\ &= (\beta_0 + 10\beta_2) + (\beta_1 - \beta_2) x_i + u_i. \end{aligned}$$

Definiert man

$$\alpha_0 := \beta_0 + 10\beta_2 \qquad \text{und} \qquad \alpha_1 := \beta_1 - \beta_2,$$

so lautet das äquivalente Modell:

$$y_i = \alpha_0 + \alpha_1 x_i + u_i.$$

Das Weglassen von z_i führt in diesem speziellen Fall nicht zu einem Omitted-Variable-Bias und verletzt nicht automatisch die Annahme der strikten Exogenität.

Denn z_i enthält wegen

$$z_i = 10 - x_i$$

keine zusätzliche, von der Konstante und x_i unabhängige Information.

Vielmehr ist es notwendig, eine der perfekt kollinearen Variablen wegzulassen, damit die Annahme des vollen Spaltenrangs erfüllt werden kann.

Aus der ursprünglichen strikten Exogenität

$$E[u_i \mid x_i, z_i] = 0$$

folgt wegen $z_i = 10 - x_i$ auch:

$$E[u_i \mid x_i] = 0.$$

Somit bleibt MLR 4 im reduzierten Modell erfüllt.

Nicht getrennt identifiziert werden können allerdings die ursprünglichen Parameter β_1 und β_2.

Identifiziert werden kann nur ihre Differenz:

$$\alpha_1 = \beta_1 - \beta_2.$$

e) Die Residuen ergeben sich aus:

$$\hat{u} = y - \hat{y}.$$

Damit folgt:

$$\hat{u} = \begin{pmatrix} 4 \\ 5 \\ 6 \\ 7 \\ 9 \\ 11 \end{pmatrix} - \begin{pmatrix} 4{,}5 \\ 5{,}0 \\ 6{,}5 \\ 7{,}0 \\ 8{,}5 \\ 10{,}5 \end{pmatrix} = \begin{pmatrix} -0{,}5 \\ 0 \\ -0{,}5 \\ 0 \\ 0{,}5 \\ 0{,}5 \end{pmatrix}.$$

Die Summe der quadrierten Residuen beträgt:

$$\begin{aligned} SSR &= \hat{u}'\hat{u} \\ &= (-0{,}5)^2 + 0^2 + (-0{,}5)^2 + 0^2 + (0{,}5)^2 + (0{,}5)^2 \\ &= 0{,}25 + 0{,}25 + 0{,}25 + 0{,}25 \\ &= 1. \end{aligned}$$

Alternativ: $\bar{y} = 7 \Rightarrow SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 25 \Rightarrow SSR = 34 - 25 = 9$
Anmerkung: In dem Zahlenbeispiel der Aufgabe gilt leider $\hat{y} \cdot \hat{u} = 4 \neq 0$. Daher führen die zwei Lösungswege nicht zum gleichen Ergebnis.

f) Für das Bestimmtheitsmaß gilt:

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

Mit $SSR = 1$ und $SST = 34$ folgt:

$$R^2 = 1 - \frac{1}{34} = \frac{33}{34} \approx 0{,}9706.$$

Das Modell erklärt somit ungefähr $97{,}06\%$ der Variation des Nettoeinkommens in der Stichprobe.

Der unverzerrte Schätzer für die Varianz der Störterme lautet:

$$\hat{\sigma}^2 = \frac{SSR}{n-K-1}.$$

Da $n = 6$ und $K = 2$ gilt:

$$\hat{\sigma}^2 = \frac{1}{6-2-1} = \frac{1}{3} \approx 0{,}3333.$$

Mit dem alternativen Lösungsweg über $SSE = 25$ und $SST = 34$ folgt $R^2 = \frac{25}{34} \approx 0,735$ und mit $SSR = 9$ folgt $\hat{\sigma}^2 = 9/3 = 3$.

Aufgabe 2 – Inferenz

a) Das unrestringierte Modell enthält drei Regressoren neben der Konstante: $advertising_i$, $employees_i$ und $location_i$.

Daher gilt:

$$K = 3.$$

Die Nennerfreiheitsgrade der F-Statistik betragen:

$$n - K - 1 = 44 - 3 - 1 = 40.$$

b) Im restringierten Modell werden die Variablen $advertising_i$ und $location_i$ weggelassen.

Die Nullhypothese lautet daher:

$$H_0 : \beta_1 = 0 \quad \text{und} \quad \beta_3 = 0.$$

Die Alternativhypothese lautet:

$$H_1 : \beta_1 \neq 0 \quad \text{oder} \quad \beta_3 \neq 0.$$

Unter der Nullhypothese haben Werbeausgaben und Standortattraktivität gemeinsam keinen zusätzlichen Erklärungsgehalt für den Umsatz, wenn für die Beschäftigtenzahl kontrolliert wird.

c) Es werden die zwei Restriktionen

$$\beta_1 = 0 \qquad \text{und} \qquad \beta_3 = 0$$

überprüft.

Damit gilt:

$$J = 2.$$

d) Die F-Statistik lautet:

$$F = \frac{(SSR_r - SSR_u)/J}{SSR_u/(n - K - 1)}.$$

Mit $SSR_r = 900$, $SSR_u = 720$, $J = 2$ und $n - K - 1 = 40$ folgt:

$$\begin{aligned} F &= \frac{(900 - 720)/2}{720/40} \\ &= \frac{180/2}{18} \\ &= \frac{90}{18} \\ &= 5. \end{aligned}$$

Damit beträgt die F-Statistik:

$$F = 5{,}00.$$

e) Für $J = 2$ Zählerfreiheitsgrade und 40 Nennerfreiheitsgrade beträgt der kritische Wert zum Signifikanzniveau $\alpha = 5\%$:

$$c_{F;2;40;0,95} = 3{,}23.$$

Da gilt:

$$F = 5{,}00 > 3{,}23,$$

wird die Nullhypothese verworfen.

Die Werbeausgaben und die Standortattraktivität besitzen somit gemeinsam einen statistisch signifikanten zusätzlichen Erklärungsgehalt für den monatlichen Umsatz.

f) Beim globalen F-Test lautet die Nullhypothese:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$$

Die Alternativhypothese lautet:

$$H_1 : \text{Mindestens einer der Parameter } \beta_1, \beta_2, \beta_3 \text{ ist ungleich null.}$$

Die Teststatistik des globalen F-Tests lautet:

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

Mit $R^2 = 0{,}64$, $K = 3$ und $n - K - 1 = 40$ folgt:

$$\begin{aligned} F &= \frac{0{,}64/3}{(1 - 0{,}64)/40} \\ &= \frac{0{,}2133}{0{,}009} \\ &\approx 23{,}70. \end{aligned}$$

Für 3 Zählerfreiheitsgrade und 40 Nennerfreiheitsgrade beträgt der kritische Wert zum Signifikanzniveau $\alpha = 5\%$:

$$c_{F;3;40;0{,}95} = 2{,}84.$$

Da gilt:

$$23{,}70 > 2{,}84,$$

wird die Nullhypothese verworfen.

Die drei Regressoren erklären den Umsatz somit gemeinsam statistisch signifikant.

Aufgabe 3 – Heteroskedastizität

a) Heteroskedastizität liegt vor, wenn die bedingten Varianzen der Störterme nicht für alle Beobachtungen gleich sind.

Formal gilt:

$$\mathrm{Var}(u_i \mid X) = \sigma_i^2,$$

wobei für mindestens ein Paar von Beobachtungen $i \neq j$ gilt:

$$\sigma_i^2 \neq \sigma_j^2.$$

b) Bei Heteroskedastizität ist die Annahme der Homoskedastizität verletzt.

Dies ist Teil der Annahme MLR 5, welche fordert:

$$\Sigma(u \mid X) = \sigma^2 I.$$

Im vorliegenden Fall gilt stattdessen:

$$\Sigma(u \mid X) = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

c) Die Transformation lautet:

$$\tilde{y}_i = \frac{y_i}{\sigma_i}, \qquad \tilde{x}_{i0} = \frac{1}{\sigma_i}, \qquad \tilde{x}_{i1} = \frac{x_i}{\sigma_i}.$$

Für $i = 1$ gilt:

$$\tilde{y}_1 = \frac{2}{1} = 2, \qquad \tilde{x}_{10} = \frac{1}{1} = 1, \qquad \tilde{x}_{11} = \frac{1}{1} = 1.$$

Für $i = 2$ gilt:

$$\tilde{y}_2 = \frac{3}{2} = 1{,}5, \qquad \tilde{x}_{20} = \frac{1}{2} = 0{,}5, \qquad \tilde{x}_{21} = \frac{2}{2} = 1.$$

Für $i = 3$ gilt:

$$\tilde{y}_3 = \frac{5}{1} = 5, \qquad \tilde{x}_{30} = \frac{1}{1} = 1, \qquad \tilde{x}_{31} = \frac{3}{1} = 3.$$

Damit ergeben sich die transformierten Daten:

i	$\tilde{x}_{i0}$	$\tilde{x}_{i1}$	$\tilde{y}_i$
1	1	1	2
2	0,5	1	1,5
3	1	3	5

d) Die zweite Beobachtung besitzt mit $\sigma_2 = 2$ die größte Standardabweichung des Störterms.

Sie wird deshalb bei der GLS-Transformation durch 2 geteilt und erhält ein geringeres Gewicht als die Beobachtungen $i = 1$ und $i = 3$.

Beobachtungen mit einer höheren Unsicherheit sollen einen geringeren Einfluss auf die Schätzung ausüben.

e) Der GLS-Schätzer lautet:

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}.$$

Gegeben sind:

$$\tilde{X}'\tilde{X} = \begin{pmatrix} 2{,}25 & 4{,}50 \\ 4{,}50 & 11{,}00 \end{pmatrix}, \qquad \tilde{X}'\tilde{y} = \begin{pmatrix} 7{,}75 \\ 18{,}50 \end{pmatrix}.$$

Die Determinante von $\tilde{X}'\tilde{X}$ beträgt:

$$\begin{aligned}\det(\tilde{X}'\tilde{X}) &= 2{,}25 \cdot 11 - 4{,}5 \cdot 4{,}5 \\ &= 24{,}75 - 20{,}25 \\ &= 4{,}5.\end{aligned}$$

Damit ergibt sich:

$$(\tilde{X}'\tilde{X})^{-1} = \frac{1}{4{,}5}\begin{pmatrix} 11 & -4{,}5 \\ -4{,}5 & 2{,}25 \end{pmatrix}.$$

Folglich gilt:

$$\tilde{\beta} = \frac{1}{4{,}5}\begin{pmatrix} 11 & -4{,}5 \\ -4{,}5 & 2{,}25 \end{pmatrix}\begin{pmatrix} 7{,}75 \\ 18{,}50 \end{pmatrix}.$$

Für den Achsenabschnitt ergibt sich:

$$\begin{aligned}\tilde{\beta}_0 &= \frac{11 \cdot 7{,}75 - 4{,}5 \cdot 18{,}5}{4{,}5} \\ &= \frac{85{,}25 - 83{,}25}{4{,}5} \\ &= \frac{2}{4{,}5} \\ &\approx 0{,}4444.\end{aligned}$$

Für den Steigungskoeffizienten ergibt sich:

$$\begin{aligned}\tilde{\beta}_1 &= \frac{-4{,}5 \cdot 7{,}75 + 2{,}25 \cdot 18{,}5}{4{,}5} \\ &= \frac{-34{,}875 + 41{,}625}{4{,}5} \\ &= \frac{6{,}75}{4{,}5} \\ &= 1{,}5.\end{aligned}$$

Der GLS-Schätzer lautet somit:

$$\tilde{\beta} = \begin{pmatrix} 0{,}4444 \\ 1{,}5 \end{pmatrix}.$$

f) Der geschätzte Steigungskoeffizient beträgt:

$$\tilde{\beta}_1 = 1{,}5.$$

Erhöhen sich die Werbeausgaben um eine Einheit, steigt der erwartete Umsatz im Durchschnitt um $1{,}5$ Einheiten.

Da beide Variablen in Tausend Euro gemessen werden, bedeutet dies: Eine Erhöhung der Werbeausgaben um 1.000 Euro ist mit einem geschätzten Umsatzanstieg von 1.500 Euro verbunden.

Aufgabe 4 – Serielle Korrelation in Zeitreihen

a) Serielle Korrelation liegt vor, wenn Störterme aus unterschiedlichen Zeitpunkten miteinander korreliert sind.

Bei positiver serieller Korrelation gilt beispielsweise:

$$\mathrm{Cov}(u_t, u_{t-1}) > 0.$$

Ein positiver oder negativer Störterm in einer Periode hängt dann systematisch mit dem Störterm der vorherigen Periode zusammen.

b) Ein AR(1)-Prozess

$$u_t = \rho u_{t-1} + \epsilon_t$$

ist stationär, wenn sein Erwartungswert für alle Zeitpunkte identisch ist:

$$E[u_t] = \mu \qquad \text{für alle } t.$$

Zusätzlich muss seine Varianz über die Zeit konstant sein:

$$\mathrm{Var}(u_t) = \sigma_u^2 \qquad \text{für alle } t.$$

Für ϵ_t gilt iid mit $E[\epsilon_t] = 0$ und $Var(\epsilon_t) = \sigma_\epsilon^2$ für alle t.

Mit

$$\underbrace{E[u_t]}_{\mu} = E[\rho u_{t-1} + \epsilon_t] = \rho \underbrace{E[u_{t-1}]}_{\mu} + \underbrace{E[\epsilon_t]}_{0} \Leftrightarrow \mu = \rho\mu$$

folgt $\mu = 0$ oder $\rho = 1$.

Mit

$$\underbrace{Var(u_t)}_{\sigma_u^2} = Var(\rho u_{t-1} + \epsilon_t) = \rho^2 \underbrace{Var(u_{t-1})}_{\sigma_u^2} + \rho \underbrace{Cov(u_{t-1}, \epsilon_t)}_{0} + \underbrace{Var(\epsilon_t)}_{\sigma_\epsilon^2} \Leftrightarrow \sigma_u^2 = \rho^2\sigma_u^2 + \sigma_\epsilon^2$$

$$\Rightarrow \sigma_u^2 = \frac{1}{1-\rho^2}\sigma_\epsilon^2, \text{ falls } |\rho| < 1$$

Für einen stationären AR(1)-Prozess muss also gelten:

$$|\rho| < 1 \text{ und } \mu = 0$$

c) Da die Innovationen ε_t einen Erwartungswert von null besitzen, gilt bei Stationarität:

$$E[u_t] = 0.$$

Die Varianz eines stationären AR(1)-Prozesses lautet:

$$\mathrm{Var}(u_t) = \frac{\sigma_\varepsilon^2}{1-\rho^2}.$$

Mit $\sigma_\varepsilon^2 = 3$ und $\rho = 0{,}5$ folgt:

$$\begin{aligned}\mathrm{Var}(u_t) &= \frac{3}{1-(0{,}5)^2} \\ &= \frac{3}{1-0{,}25} \\ &= \frac{3}{0{,}75} \\ &= 4.\end{aligned}$$

Damit gilt:

$$E[u_t] = 0, \qquad \mathrm{Var}(u_t) = 4.$$

d) Für einen AR(1)-Prozess gilt:

$$\mathrm{Cov}(u_t, u_{t-1}) = \rho \,\mathrm{Var}(u_{t-1}).$$

Wegen der Stationarität gilt:

$$\mathrm{Var}(u_{t-1}) = \mathrm{Var}(u_t) = 4.$$

Daher folgt:

$$\mathrm{Cov}(u_t, u_{t-1}) = 0{,}5 \cdot 4 = 2.$$

e) Die Quasidifferenzierung lautet:

$$\tilde{y}_t = y_t - \rho y_{t-1}$$

und

$$\tilde{x}_t = x_t - \rho x_{t-1}.$$

Mit $\rho = 0{,}5$ folgt für $t = 2$:

$$\tilde{y}_2 = 13 - 0{,}5 \cdot 10 = 8, \qquad \tilde{x}_2 = 3 - 0{,}5 \cdot 2 = 2.$$

Für $t = 3$ folgt:

$$\tilde{y}_3 = 15 - 0{,}5 \cdot 13 = 8{,}5, \qquad \tilde{x}_3 = 4 - 0{,}5 \cdot 3 = 2{,}5.$$

Für $t = 4$ folgt:

$$\tilde{y}_4 = 18 - 0{,}5 \cdot 15 = 10{,}5, \qquad \tilde{x}_4 = 5 - 0{,}5 \cdot 4 = 3.$$

Für $t = 5$ folgt:

$$\tilde{y}_5 = 20 - 0{,}5 \cdot 18 = 11, \qquad \tilde{x}_5 = 6 - 0{,}5 \cdot 5 = 3{,}5.$$

Damit ergibt sich:

t	2	3	4	5
$\tilde{y}_t$	8	8,5	10,5	11
$\tilde{x}_t$	2	2,5	3	3,5

f) Der transformierte Störterm lautet:

$$\tilde{u}_t = u_t - \rho u_{t-1}.$$

Mit dem AR(1)-Prozess

$$u_t = \rho u_{t-1} + \varepsilon_t$$

folgt:

$$\begin{aligned} \tilde{u}_t &= (\rho u_{t-1} + \varepsilon_t) - \rho u_{t-1} \\ &= \varepsilon_t. \end{aligned}$$

Der transformierte Störterm entspricht damit der Innovation ε_t, welche nicht seriell korreliert ist.

Die Quasidifferenzierung neutralisiert somit die Autokorrelation und ermöglicht eine effizientere FGLS-Schätzung.

Aufgabe 5 – Instrumentalvariablen

a) Ein Regressor ist endogen, wenn er mit dem Störterm korreliert ist.

Für die Ausbildungsjahre $educ_i$ lautet die Bedingung:

$$\mathrm{Cov}(educ_i, u_i) \neq 0.$$

In diesem Fall ist die Annahme der strikten Exogenität verletzt.

b) Die unbeobachteten Fähigkeiten einer Person beeinflussen möglicherweise sowohl die Ausbildungsdauer als auch den Lohn.

Da diese Fähigkeiten nicht beobachtet werden, sind sie Teil des Störterms u_i.

Wenn Personen mit höheren Fähigkeiten im Durchschnitt mehr Ausbildungsjahre wählen, gilt:

$$\mathrm{Cov}(educ_i, ability_i) \neq 0.$$

Dann gilt auch:

$$\mathrm{Cov}(educ_i, u_i) \neq 0.$$

Die Variable $educ_i$ ist in diesem Fall endogen, weil eine relevante Variable ausgelassen wurde.

c) Ein gültiges Instrument z_i für den endogenen Regressor $educ_i$ muss zwei Bedingungen erfüllen.

Die erste Bedingung ist die Instrumentrelevanz:

$$\mathrm{Cov}(z_i, educ_i) \neq 0.$$

Das Instrument muss also mit dem endogenen Regressor korreliert sein.

Die zweite Bedingung ist die Instrumentexogenität:

$$\mathrm{Cov}(z_i, u_i) = 0.$$

Das Instrument darf also nicht mit dem Störterm der strukturellen Lohngleichung korreliert sein.

d) Die Entfernung zur nächstgelegenen Universität kann die Ausbildungsentscheidung beeinflussen.

Eine größere Entfernung kann beispielsweise die finanziellen, zeitlichen oder organisatorischen Kosten eines Studiums erhöhen.

Daher ist es plausibel, dass Personen mit einer größeren Entfernung im Durchschnitt weniger Ausbildungsjahre aufweisen.

Die Relevanzbedingung

$$\mathrm{Cov}(distance_i, educ_i) \neq 0$$

könnte daher erfüllt sein.

Die gegebene empirische Kovarianz

$$\widehat{\mathrm{Cov}}(distance, educ) = -1{,}20$$

ist ungleich null und spricht in der Stichprobe für Instrumentrelevanz.

e) Der einfache IV-Schätzer lautet:

$$\hat{\beta}_1^{IV} = \frac{\widehat{\mathrm{Cov}}(distance, \ln(wage))}{\widehat{\mathrm{Cov}}(distance, educ)}.$$

Einsetzen der gegebenen Werte liefert:

$$\hat{\beta}_1^{IV} = \frac{-0{,}096}{-1{,}20} = 0{,}08.$$

Damit lautet der IV-Schätzer:

$$\boxed{\hat{\beta}_1^{IV} = 0{,}08.}$$

f) Der OLS-Schätzer beträgt:

$$\hat{\beta}_1^{OLS} = 0{,}10.$$

Der IV-Schätzer beträgt:

$$\hat{\beta}_1^{IV} = 0{,}08.$$

Der OLS-Schätzer ist somit größer als der IV-Schätzer:

$$0{,}10 > 0{,}08.$$

Falls $educ_i$ tatsächlich endogen ist, ist der OLS-Schätzer verzerrt und inkonsistent.

Falls $distance_i$ ein relevantes und exogenes Instrument ist, ist der IV-Schätzer dagegen konsistent für den kausalen Effekt der Ausbildungsjahre auf den logarithmierten Stundenlohn.

In diesem Fall ist der IV-Schätzer zur Bestimmung des kausalen Effekts vorzuziehen.