Universität St. Gallen

Hochschule für Wirtschafts-,

Rechts- und Sozialwissenschaften sowie Internationale Beziehungen

Hausarbeit

Data Analytics II: Empirische Wirtschaftsforschung

FS 2020

Verfasser:

Niklas Leander Kampe

Ebnetstrasse 6

8406 Winterthur

Bachelor Volkswirtschaftslehre (BVWL)

16-611-618

Universität St. Gallen

01. Juni 2020

Inhaltsverzeichnis

Tabellenverzeichnis

Oregon Health Insurance Experiment

Der vorgegebene Datensatz der Umfrage im Rahmen des Oregon Health Insurance Experiments beinhaltet folgende Variablen :

- *Winner*: Dummyvariable. Eins, wenn Person in der Lotterie gewonnen hatte
- *insured*: Dummyvariable. Eins, wenn Person krankenversichert ist
- *good_health*: Dummyvariable. Eins, wenn Person angibt bei guter Gesundheit zu sein
- *female*: Dummyvariable. Eins, wenn Person weiblich ist
- *hhinc_pctfpl*: Haushaltseinkommen in Prozent der Armutsgrenze
- *race*: Kategorievariable, die anzeigt zu welcher ethnischen Gruppe die Person gehört

Die folgenden Teilaufgaben in Kapitel 1 beziehen sich auf den obenstehend definierten Datensatz mit 21'099 Objekten und beruht auf einer freiwilligen Umfrage. Die Prüfung auf fehlende Werte ergibt, dass alle darin enthaltenen Objekte vollständige Teildatensätze enthalten (Kampe, 2020).

1.1 Univariate Regression

Die univariate Regression mit der Dummyvariablen *good_health* als abhängige und *insured* als erklärende Variable hat zum Ziel den Einfluss von einer vorhandenen Krankenversicherung bei den Umfrageteilnehmern auf den bestehenden, selbst eingeschätzten Gesundheitszustand zu untersuchen. Dies stellt gerade in Ländern ohne bestehende oder nur teils obligatorische Krankenversicherungspflicht wie den USA eine zentrale Fragestellung dar, um den effektiven Nutzen einer Versicherungspflicht auf den Gesundheitszustand der Bevölkerung schätzen zu können und somit eine dahingehende valide Diskussionsgrundlage zu definieren. Innerhalb dieser Schätzung ist jedoch zu beachten, dass zwischen Korrelations- und Kausalitätseffekten unterschieden werden muss, die in der anschliessenden Interpretation der Ergebnisse, vor allem hinsichtlich solch zentraler Fragestellungen, von grosser Bedeutung ist (Eugster & Knaus, 2020A, S. 4).

Die univariate Regression $good\_health = \beta_0 + \beta_1 * insured + u$ ergibt folgende Ergebnisse:

Min	1Q	Median	3Q	Max
- 0.5769	- 0.5769	0.4231	0.4231	0.4268

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	0.576865	0.003990	144.586	< 2e-16 ***
insured	- 0.003699	0.007640	- 0.484	0.628

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4942 on 21097 degrees of freedom
Multiple R-squared: 1.111e-05, Adjusted R-squared: - 3.629e-05
F-statistic: 0.2344 on 1 and 21097 DF, p-value: 0.6283

Tabelle 1: Univariate OLS-Regression good_health ~ insured (Kampe, 2020)

Aus der obenstehenden Tabelle ist ersichtlich, dass im Rahmen der oben definierten univariaten Regression die Koeffizienten β_0 als $\hat{\beta}_0 = 0.576865$ und β_1 als $\hat{\beta}_1 = -0.003699$ geschätzt werden. Die Schätzung des Achsenabschnittsparameters (Intercept/$\hat{\beta}_0$) definiert hierbei den Wert von *good_health*, wenn die erklärende Variable *insured* den Wert 0 annimmt bzw. die befragte Person

keine Krankenversicherung besitzt, was aus der Definition als Dummyvariable hervorgeht. Der Steigungsparameter (insured/$\hat{\beta}_1$) zeigt hingegen die Steigung der Regressionsgeraden und somit die relative Beziehung zwischen den Variablen *good_health* und *insured* ($\Delta insured \rightarrow -0.003699 * \Delta good\_health$). Im Rahmen dieser Regression ergibt sich ein negativer Steigungsparameter, was bedeutet, dass sich die Variable *good_health* mit steigender erklärender Variable *insured* verringert. Qualitativ interpretiert bedeutet dies, dass der persönlich eingeschätzte Gesundheitszustand der Versicherten leicht schlechter ausfällt als der der Nicht-Versicherten. Die Standardfehler von $\hat{\beta}_0$ und $\hat{\beta}_1$ fallen mit den Werten von 0.003990 bzw. 0.007640 hierbei sehr klein aus, was auf jeweilig kleine Residuen schliessen lässt.

Nachdem die Korrelation zwischen *good_health* und *insured* berechnet und interpretiert wurde, stellt sich folglich die Frage der Kausalität bzw. des kausalen Effekts. Dieser lässt sich jedoch im Vergleich zur Korrelation nicht quantitativ berechnen, wodurch eine qualitative Interpretation vor dem Hintergrund der Variablendefinition als auch den Korrelationsergebnissen erfolgen muss. Die abhängige Dummyvariable *good_health* gibt an, ob sich die befragte Person bei guter Gesundheit befindet, und die erklärende Dummyvariable, ob die befragte Person krankenversichert ist. Der Blick auf diese Definition lässt darauf schliessen, dass die Exogenitätsannahme nicht erfüllt ist und *insured* somit eine endogene Variable definiert. Im Rahmen von Regressionsanalysen werden folgende zwei Arten von Endogenität definiert (Eugster & Knaus, 2020E, S. 5):

1. Ommited Variables
2. Reverse Causality

Im bestehenden Fall kann auf Ommited Variables geschlossen werden, wobei eine Verzerrung durch ausgelassene Variablen entsteht (Eugster & Knaus, 2020D, S. 5 ff.). Die Annahme, dass der Gesundheitszustand lediglich durch den Krankenversicherungsstatus erklärt werden kann, ist sehr unwahrscheinlich, da durchaus weitere Variablen in Betracht kommen, die einen Einfluss auf den Gesundheitszustand einer Person haben. Dies könnten beispielsweise die gesundheitliche Grundveranlagung, die familiäre Vorgeschichte, das Konsumverhalten und viele weitere sein. Als Folge dessen entsteht der Zustand, dass $Cov(insured, u) \neq 0$ und die erklärende Variable somit mit dem Fehlerterm der Regressionsgleichung korreliert (Eugster & Knaus, 2020D, S. 5 ff.). Dies führt schlussendlich dazu, dass die Schätzung der Koeffizienten verzerrt und inkonsistent ist und zudem, dass Korrelationen entstehen, wo keine Kausalitäten vorhanden sind, was in diesem Fall zutrifft.

1.2 Exogenes Instrument

Im Zuge dieser Teilaufgabe wird die Lotterie (*winner*) des OHIE als Instrumentalvariable für die Dummyvariable der Krankenversicherung (*insured*) auf Exogenität geprüft, da in Kapitel 1.1 erarbeitet wurde, dass im Rahmen der Regression $good\_health \sim insured$ Endogenität in Form von Ommited Variables auftritt und somit keine unverzerrte und konsistente Schätzung möglich ist. Trotz Endogenität gewährt die Hinzunahme einer Instrumentalvariable für die bestehende erklärende Variable die konsistente Schätzung der Regressionsparameter. Hierzu muss die gegebene Instrumentalvariable *winner* folgende zwei Bedingungen erfüllen, um als valides Instrument zu gelten (Hanck, Arnold, Gerber, & Schmelzer, 2019):

1. Relevanz: *insured* und die IV *winner* müssen korreliert sein → $Cov(winner, insured) \neq 0$

2. Exogenität: *z* darf nicht mit Fehlerterm *u* korreliert sein → $Cov(winner, u) = 0$

Mit Rückblick auf die Fragestellung dieser Teilaufgabe stellt die Exogenität die zentrale zu untersuchende Annahme dar, die auch als «Exclusion Restriction» bezeichnet wird. Die Definition besagt hierbei, dass die Instrumentalvariable der Lotterie nicht mit dem Fehlerterm korreliert sein darf und somit analog eine Kovarianz mit dem Fehlerterm von null aufweisen muss ($Cov(z, v) = 0$) (Eugster & Knaus, 2020E, S. 5). Gemäss der Definition des Variable *winner* (Gewinner in der Lotterie), welche auf einer Zufallsverteilung beruht und sehr wahrscheinlich nicht mit den existierenden erklärenden Variablen für den Gesundheitszustand einer Person korreliert ist, kann somit schlussendlich gesagt werden, dass die Instrumentalvariable *winner* ein exogenes Instrument darstellt und somit eine der zwei Annahmen erfüllt. Wichtig ist jedoch zu beachten, dass die erste Annahme der Relevanz nicht untersucht wurde und somit nicht abschliessend feststeht, ob *winner* ein valides Instrument definiert.

1.3 Kreuztabelle

Nachdem in der vorherigen Teilaufgabe die Frage beantwortet wurde, ob die Instrumentalvariable *winner* ein exogenes Instrument darstellt, prüft diese Teilaufgabe die Validitätsannahme der Relevanz. Aus der Erstellung einer Kreuztabelle zwischen dem randomisierten Angebot eines OHP (*winner*) und dem Versichertenstatus (*insured*) ergeht folgende Grafik:

A tibble: 4 x 4

	winner	insured	n	percentage*
	<dbl>	*<dbl>*	*<int>*	*<dbl>*
1	0	0	9284	44.00
2	0	1	1334	6.32
3	1	0	6061	28.73
4	1	1	4420	20.95

** gerundet auf 2 Dezimalstellen*

Tabelle 2: Kreuztabelle winner x insured (Kampe, 2020)

Die obenstehende Tabelle untersucht hierbei im vorhandenen Datensatz, in welcher absoluten und relativen Anzahl die Randomisierung der zwei Dummyvariablen vorhanden ist. Den grössten Wert definieren hierbei knapp 44% der Befragten, die weder Gewinner der Lotterie noch krankenversichert sind, und den niedrigsten Wert ca. 6.32% der Befragten, die nicht in der Lotterie gewonnen haben, jedoch krankenversichert sind. Die Bedingung der Relevanz einer Instrumentalvariable stellt neben der in Kapitel 1.2 behandelten Exogenität die zweite Bedingung für die Validität einer Instrumentalvariable dar. Diese definiert sich darüber, dass die Instrumentalvariable *z* bzw. *winner* nicht mit dem Regressionskoeffizienten *x1* bzw. *insured* korrelieren darf ($Cov(z, x_1) \neq 0$) (Hanck, Arnold, Gerber, & Schmelzer, 2019). Mit Rückblick auf die Kreuztabelle in Tabelle 2 stellt sich nun die Frage, ob die Dummyvariable *winner* ein relevantes/starkes Instrument für die einhergehende Regression aus Kapitel 1.1 definiert und somit analog, ob *winner* einen Teil der Variation von *insured* abbildet (Eugster & Knaus, 2020E, S. 5). Aus der Tabelle wird ersichtlich, dass eine ungleichmässige Verteilung zwischen *winner* und *insured* besteht, was wiederum darauf schliessen lässt, dass $Cov(winner, insured) \neq 0$ ist und somit ein relevantes Instrument in Form der Instrumentalvariable *winner* besteht. Die unterstreicht auch quantitativ die mathematische Berechnung der Kovarianz auf Grundlage des vorhandenen Datensatzes, wobei sich der Wert $Cov(winner, insured) = 0.07402033$ ergibt,

wodurch die Relevanz gegeben ist, auch wenn sie nicht besonders hochausfällt. Die führt dazu, dass die Instrumentalvariable zwar ein relevantes, jedoch schwaches Instrument definiert, wodurch die Schätzerergebnisse extrem gross oder auch extrem klein ausfallen können, was wiederum zu einer Verzerrung der Ergebnisse führen kann und die Varianz im Vergleich zum OLS-Schätzer umso grösser ausfällt (Eugster & Knaus, 2020E, S. 11).

1.4 Kausaler Effekt mit Instrumentalvariable

Wie bereits in den einhergehenden Teilaufgaben erläutert wurde, beruht der Einsatz von Instrumentalvariablen grundsätzlich auf der konsistenten Schätzbarkeit der Regressionskoeffizienten trotz bestehender Endogenität und der daraus folgenden Verletzung der Annahme MLR.4 durch Ommited Variables und/oder Reverse Causality. Die Durchführung der Regression $good\_health = \beta_0 + \beta_1 * insured + u$ unter Hinzunahme der Instrumentalvariable *winner* erfolgt nach dem Konzept des Two Stage Least Squares (TSLS) und erfolgt im Gegensatz zur OLS-Regression in zwei Schritten. In Schritt 1 der TSLS-Methode wird die exogene Variation von *insured*, die mit dem Fehlerterm *v* korreliert, mit Hilfe einer OLS-Regression isoliert (Eugster & Knaus, 2020E, S. 6 ff.):

$$insured = \pi_0 + \pi_1 * winner + \varepsilon \quad \text{(i)}$$

Im zweiten Schritt wird der mit der OLS-Regression geschätzte Wert $\widehat{insured}$ aus (i) in die ursprüngliche Regression eingesetzt (Eugster & Knaus, 2020E, S. 6 ff.):

$$good\_health = \beta_0 + \beta_1 * \widehat{insured} + \tilde{v}\,, \quad \tilde{v} = \beta_1 * \hat{\varepsilon} + v \quad \text{(ii)}$$

Aus der TSLS-Regression mit *winner* als Instrumentalvariable ergehen folgende Ergebnisse:

Min	1Q	Median	3Q	Max
- 0.4217	- 0.4217	- 0.1256	0.5783	0.8744

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	0.125636	0.004076	30.82	< 2e-16 ***
winner	0.296080	0.005784	51.19	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4201 on 21097 degrees of freedom
Multiple R-squared: 0.1105, Adjusted R-squared: 0.1104
F-statistic: 2621 on 1 and 21097 DF, p-value: < 2.2e-16

Tabelle 3: TSLS-Regression Schritt 1 - Isolierung exogene Variation von insured korreliert mit v (= tsls1.1) (Kampe, 2020)

Min	1Q	Median	3Q	Max
- 0.6014	- 0.5507	0.3986	0.4493	0.4493

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	0.529153	0.007123	74.291	< 2e-16 ***
tsls1.1_fittedvalues	0.171255	0.022954	7.461	8.94e-14

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4936 on 21097 degrees of freedom
Multiple R-squared: 0.002631, Adjusted R-squared: 0.002584
F-statistic: 55.66 on 1 and 21097 DF, p-value: 8.936e-14

Tabelle 4: TSLS-Regression Schritt 2 - Schätzer aus OLS-Regression in ursprüngliche Regression (Kampe, 2020)

Aus der obenstehenden Tabelle ist ersichtlich, dass die Regressionskoeffizienten von $\widehat{\beta_0} = 0.529153$ und $\widehat{\beta_1} = 0.171255$ geschätzt werden. Mit Rückblick auf die einhergehende OLS-Regression aus Teilaufgabe 1.1 weist die IV-Schätzung somit einen ähnlichen Intercept bzw. Achsenabschnitt auf, jedoch fällt der Steigungsparameter grösser aus und ist sowohl positiv und als auch proportionaler ($\Delta insured \rightarrow 0.171255 * \Delta good\_health$). Die effektiven Differenzen der IV-Schätzung zur OLS-Schätzung ($\widehat{\beta_{0/1\,IV}} - \widehat{\beta_{0/1\,OLS}}$) liegen für $\widehat{\beta_0}$ bei -0.047712 und für $\widehat{\beta_1}$ bei 0.174954.

1.5 Probleme bei Nichtbeantwortung

Die Nichtbeantwortung im Rahmen einer nicht verpflichtenden Umfrage, welche wir in diesem Experiment und Datensatz vorliegen haben, kann zu Problemen in der Schätzung des Regressionskoeffizienten führen. Dies beruht darauf, dass fehlende Werte innerhalb des Datensatzes für die anschliessende Regression zu einer selektiven Stichprobe führen können, die wiederum zu Heterogenität bei der Selektion auf die erklärende Variable oder zu Inkonsistenz aller Parameter bei der Selektion auf die abhängige Variable führen (Eugster & Knaus, 2020C, S. 4 ff.). Solange die fehlenden Werte jedoch zufällig verteilt sind, stellt dies dahingehend kein grundlegendes Problem dar, da die Schätzer weiterhin über die um die Anzahl fehlender Werte reduzierte zufällige Stichprobe konsistent geschätzt werden können (Eugster & Knaus, 2020C, S. 4). Somit gilt grundsätzlich, dass eine Selektion auf die abhängige Variable in allen Fällen verboten ist und die Selektion auf die erklärende Variable unter Einhaltung der Anpassungsschritte möglich ist (Eugster & Knaus, 2020C, S. 6).

Die Berücksichtigung der zusätzlichen Kontrollvariablen erbringen folgendes Ergebnis:

Coefficients:

(Intercept)	insured	female	hhinc_pctfpl	racehisp	raceother	racewhite
0.467382	0.036567	0.008704	0.001162	-0.004733	0.044912	0.001332

Tabelle 5: Berücksichtigung zusätzlicher Kontrollvariablen female, hhinc_pctfpl und race (Kampe, 2020)

Die Hinzunahme der zusätzlichen Kontrollvariablen verfolgt den Zweck, dass im Rahmen fehlender Werte der Selection Bias bestimmt werden kann. Gegeben der Definition, dass eine unverzerrte Schätzung nur dann möglich ist, wenn der Fehlerterm keine Confounding beinhaltet, zeigt die Schätzung mit den zusätzlichen Variablen, ob fehlende Werte aufgrund der einhergehend erläuterten möglichen selektiven Stichprobe zu dem genannten Bias führen. Aus der Teilaufgabe 1.1 ist ersichtlich, dass der OLS-Schätzer $\widehat{\beta_1}$ für die erklärende Variable *insured* auf die unabhängige Variable *good_health* - 0.003699 beträgt. Aus Tabelle 5 ist nun ersichtlich, dass der OLS-Schätzer im Rahmen der multiplen Regression unter Hinzunahme der Kontrollvariablen +0.036567 beträgt. Somit zeigt sich ein Unterschied in der Schätzerergebnissen von absolut 0.040266. Dies lässt darauf schliessen, da sich der OLS-Schätzer $\widehat{\beta_1}$ verändert hat, dass die Nichtbeantwortung der Umfrage aufgrund der Freiwilligkeit der Teilnahme zu einem Selection Bias führt.

2 Simulation IV

Im Rahmen der Simulationsstudie wird folgender Data Generating Process (DGP) angewandt:

- Setze Anzahl der Beobachtungen auf n = 1'000.
- Ziehe zwei korrelierte Fehlerterme u und v aus einer multivariaten Normalverteilung, $u, v \sim N(\mu, \Sigma)$ mit $\mu = [0\ 0]$ und $\Sigma = \begin{bmatrix} 100 & -7 \\ -7 & 1 \end{bmatrix}$
- Ziehe die Instrumentalvariable als Dummyvariable z mit $z \sim Binominal(1, 0.5)$
- Generiere die endogene erklärende Variable Krankenversicherung, KV, in zwei Schritten:
  - Generiere eine temporäre Variable, $temp = -1 + 1 * z + v$
  - Generiere die endogene Dummyvariable, $KV = 1(temp > 0)$.
- Generiere die abhängige Variable, $G = 50 + 10 * KV + u$. Sie soll einen Gesundheitsindex darstellen, der Werte zwischen 0 und 100 annimmt.

Die folgenden Teilaufgaben bauen auf dem obenstehend definierten DGP auf.

2.1 Analyse der synthetischen Stichprobe

Zu Beginn der Simulation wird die synthetische Stichprobe auf Grundlage des einhergehend definierten DGPs analysiert. Dies geschieht durch die Erstellung geeigneter deskriptiver Statistiken, die Untersuchung der Kovarianzen sowie die Erstellung von OLS- als auch einer IV-Regression.

2.1.1 Geeignete deskriptive Statistiken

Aus der Erstellung der geeigneten deskriptiven Statistiken für die Variablen *G (Gesundheitsindex)*, *KV* (Krankenversicherung) und *z* (Instrumentalvariable) ergeht folgende Tabelle:

G	summary(G)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
		24.57	47.10	53.08	53.04	58.99	79.81
KV	table(KV)	0	1				
		650	350				
z	table(z)	0	1				
		503	497				

Tabelle 6: Deskriptive Statistiken für G, KV und z (Kampe, 2020)

Die Wahl der summary-Funktion für die Variable G beruht auf dem Attribut, dass die Variable einen Wert zwischen 0 und 100 annehmen kann. Durch die ersichtliche Aufteilung in die vier Quantile, als auch die min- und max-Werte sowie den Durchschnitt der synthetischen Stichprobe erhält man einen geeigneten Überblick über einerseits die vorhandenen Grenzen der Daten und eine erste Ahnung, wo sich der durchschnittliche Gesundheitszustand der Stichprobe befindet. Da KV und z jeweils eine Dummyvariable definieren, die nur die Werte null oder eins annehmen können, wird hierbei eine Tabellenfunktion verwendet, um die Verteilung der Dummies innerhalb der Stichprobe aufzuzeigen. Detaillierte deskriptive Statistiken eignen sich für Dummyvariablen im ersten Schritt nicht.

2.1.2 Kovarianz & Verletzung von MLR.4

Die Berechnung der Kovarianz zwischen der Variablen *KV* (Krankenversicherung) und dem Fehlerterm *u* ergibt folgendes Ergebnis:

$$Cov(KV, u) = -2.475349$$

Tabelle 7: Kovarianz zwischen der Variablen Krankenversicherung KV und Fehlerterm u (Kampe, 2020)

Das lineare Regressionsmodell zur Schätzung der Regressionskoeffizienten beruht auf vier Annahmen, die kumulativ erfüllt sein müssen, um die Koeffizienten unverzerrt schätzen zu können (Eugster & Knaus, 2020B, S. 6 ff.):

MLR.1 Linearität der Parameter
MLR.2 Zufallsstichprobe
MLR.3 Keine perfekte Multikollinearität
MLR.4 Bedingter Erwartungswert von Null

Die Annahme MLR.4 besagt, dass der Fehlerterm *u* einen Erwartungswert von 0 gegeben aller Regressoren aufweisen muss, damit die Schätzer konsistent schätzbar sind. Dies hat zur Konsequenz, dass $Cov(x_j, u) = E(x_j u) - E(x_j) * E(u) = E(x_j u) = 0$ ist (Eugster & Knaus, 2020B, S. 8). Aus der obenstehenden Tabelle im Rahmen dieser Simulation geht jedoch hervor, dass die Kovarianz zwischen dem Regressor *KV* und dem Fehlerterm *u* knapp -2,48 beträgt. Somit kann hinsichtlich dieses Ergebnisses gesagt werden, dass im Kontext der Simulation die Annahme MLR.4 verletzt ist und somit die Schätzer nicht konsistent schätzbar sind.

2.1.3 OLS Regression I

Aus der Regression $KV = \pi_0 + \pi_1 * z + \varepsilon$ auf Basis des eingeführten DGP ergeht folgendes Resultat:

Min	1Q	Median	3Q	Max
-0.5171	-0.1849	-0.1849	0.4829	0.8151

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	0.18489	0.01996	9.265	< 2e-16 ***
z	0.33221	0.02831	11.736	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4476 on 998 degrees of freedom
Multiple R-squared: 0.1213, Adjusted R-squared 0.1204
F-statistic: 137.7 on 1 and 998 DF, p-value: < 2.2e-16

Tabelle 8: Univariate OLS-Regression OLS KV ~ z (Kampe, 2020)

Aus der obenstehenden Tabelle ist ersichtlich, dass der Schätzer für die Variable *z* ($\widehat{\pi_1}$) als Steigungsparameter der Regression einen Wert von 0.33221 aufweist. Eine Interpretation dieses Wertes besagt, dass mit steigendem Wert von der erklärenden Variable *z* die abhängige Variable *KV* unproportional um $\widehat{\pi_1}$ steigt ($\Delta z \rightarrow 0.33221 * \Delta KV$), wodurch eine positive Beziehung zwischen *KV* und *z* besteht. Eine Analyse, ob es sich im Rahmen dieser Regression bei *z* um ein starkes Instrument handelt, hängt wiederum von der Annahme der Relevanz ab ($Cov(z, KV) \neq 0$). Mit Blick auf die

obenstehende Tabelle zeigt sich, dass die Annahme der Relevanz erfüllt ist und somit die Kovarianz zwischen der unabhängigen Variable *KV* und der Instrumentalvariable *z* grösser null ist. Des Weiteren ergibt die Berechnung der Kovarianz einen Wert von $Cov(z, KV) = 0.08313313$ (Kampe, 2020), was die zuvor interpretierten Ergebnisse unterstreicht, dass es sich bei *z* um ein starkes Instrument handelt.

2.1.4 OLS Regression II

Aus der Regression $G = \beta_0 + \beta_1 * KV + u$ auf Basis des eingeführten DGP ergeht folgendes Resultat:

Min	1Q	Median	3Q	Max
-28.772	-5.889	-0.027	5.907	26.465

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	53.3455	0.3463	154.042	< 2e-16 ***
KV	-0.8698	0.5854	-1.486	0.138

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.829 on 998 degrees of freedom
Multiple R-squared: 0.002207, Adjusted R-squared: 0.001208
F-statistic: 2.208 on 1 and 998 DF, p-value: 0.1376

Tabelle 9: Univariate OLS-Regression G ~ KV (Kampe, 2020)

Aus der obenstehenden Tabelle ist ersichtlich, dass der Schätzer für die Variable *KV* ($\widehat{\beta_1}$) als Steigungsparameter der Regression einen Wert von -0.8698 ergibt. Somit besteht ein negativer, unproportionaler Zusammenhang zwischen der erklärenden und der unabhängigen Variable ($\Delta KV \rightarrow -0.8698 * \Delta KV$). Die Abweichung vom wahren Wert β_0 beträgt 3.3455 und vom wahren Wert β_1 -10.8698.

2.1.5 IV Regression

Die Verwendung von *z* als Instrumentalvariable für *KV* im Rahmen der Regression $G = \beta_0 + \beta_1 * KV + u$, unterteilt in die zwei Schritte der TSLS, ergibt folgende Ergebnisse:

Min	1Q	Median	3Q	Max
-0.5171	-0.1849	-0.1849	0.4829	0.8151

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	0.18489	0.01996	9.265	< 2e-16 ***
z	0.33221	0.02831	11.736	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4476 on 998 degrees of freedom
Multiple R-squared: 0.1213, Adjusted R-squared: 0.1204
F-statistic: 137.7 on 1 and 998 DF, p-value: < 2.2e-16

Tabelle 10: IV Regression Schritt 1 - Isolierung exogene Variation von KV korreliert mit v (Kampe, 2020)

Min	1Q	Median	3Q	Max
-28.4243	-5.4909	-0.1233	5.8837	28.2614

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	49.215	0.638	77.141	< 2e-16 ***
tsls2.1_fittedvalues	10.931	1.647	6.637	5.23e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.65 on 998 degrees of freedom
Multiple R-squared: 0.04228, Adjusted R-squared: 0.04132
F-statistic: 44.06 on 1 and 998 DF, p-value: 5.23e-11

Tabelle 11: IV Regression Schritt 2 - Schätzer aus OLS-Regression in ursprüngliche Regression (Kampe, 2020)

Aus der obenstehenden Tabelle ist ersichtlich, dass die Schätzung von β_0 ($= \widehat{\beta_0}$) einen Wert von 49.215 und die Schätzung von β_1 ($= \widehat{\beta_1}$) einen Wert von 10.931 ergibt. Mit Rückblick auf die OLS-Schätzergebnisse aus Teilaufgabe 2.1.4 zeigt sich, dass sich die Schätzabweichungen von den wahren Werten für β_0 von -0.785 und für β_1 von 0.931 ergeben. Dies zeigt, dass im Vergleich zu den Abweichungen aus der OLS-Regression deutlich bessere Schätzergebnisse erzielt werden unter Hinzunahme des Instruments *z* für die erklärende Variable *KV* in der ursprünglichen Regressionsgleichung.

2.2 Simulationsstudie

In diesem Kapitel erfolgt eine Simulationsstudie auf Grundlage des einhergehend definierten DGPs. Folgende Simulationswerte werden hierbei angenommen:

- Stichprobengrösse: 775
- Anzahl Stichproben: 1'000

In den folgenden Teilaufgaben erfolgen die Berechnungen der Kernergebnisse der Simulationsstudie.

2.2.1 OLS & IV Regression

Die OLS-Regression $G = \beta_0 + \beta_0 * KV + u$ sowie die IV-Regression mit *z* als Instrumentalvariable für *KV*, basierend auf den Teilaufgaben 3.1.4 und 3.1.5, unter Hinzunahme der Stichprobengrösse von 775 und der Anzahl Stichproben von 1'000, ergeben für die Schätzer von β_1 folgende Ergebnisse:

	OLS-Schätzer	**IV-Schätzer**
[1,]	-0.96188185	10.145926
[2,]	-0.49828324	10.787635
[3,]	-1.16404508	12.272298
[4,]	-1.04702189	10.108299
[5,]	-1.00352198	9.820997
[6,]	-0.91003443	10.490237
[7,]	-1.53364862	10.451865
[8,]	-0.63884543	10.643476
[9,]	-1.10139443	9.302657
[10,]	-1.05964462	11.503647
[…,]	…	…
[1000,]	-1.11113893	8.261429
mean	-0.8769117	10.91347

Tabelle 12: OLS- und IV-Schätzer aus Simulationsstudie (Kampe, 2020)

Die obenstehende Tabelle zeigt auf, dass die OLS-Regression einen Durchschnittsschätzer für β_1 von -0.8768117 und die IV-Regression von 10.91347 schätzt. Da der wahre Wert β_1 im DGP mit dem Wert 10 definiert ist, zeigt sich bereits hier, dass sich die Werte aus der Regression unter Hinzunahme der Instrumentalvariablen deutlich näher am wahren Wert befindet.

2.2.2 Histogramme

Aus der Erstellung der Histogramme für die Verteilung der OLS- und IV-Schätzer für β_1, aufbauend aus den Ergebnissen aus Aufgabe 3.2.2, ergehen folgende Grafiken:

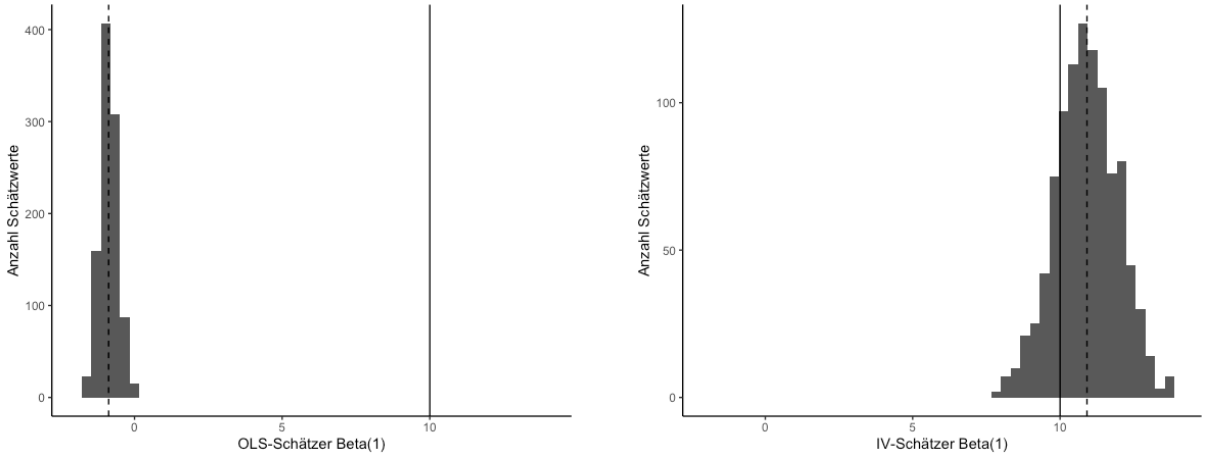

Tabelle 13: Histogramm mit IV- und OLS-Schätzer mit wahrem Wert β = 10 (Kampe, 2020)

Im linken Teil von Tabelle 13 ist die Verteilung des OLS-Schätzers, im rechten Teil die des IV-Schätzers ersichtlich. Als Kernergebnis ergibt sich, dass sich der IV-Schätzer deutlich näher an den wahren Wert von $\beta_1 = 10$ annähert. Die jeweiligen Durchschnittsschätzer sind grafisch durch die gestrichelte Linie dargestellt. Des Weiteren zeigt sich innerhalb der IV-Schätzung eine höhere Varianz. Somit kann abschliessend gesagt werden, dass auf Grundlage der Regression im Rahmen der Simulationsergebnisse die Hinzunahme der Instrumentalvariablen *z* zu deutlich besseren Schätzergebnissen führt. Dies obliegt der Verletzung von MLR.4 und der vorhandenen Endogenität innerhalb der ursprünglichen Regressionsgleichung.

2.2.3 Verzerrung der Schätzer

Die Berechnung der Verzerrung der zuvor berechneten OLS- und IV-Schätzer, ausgehend von der Berechnungsgrundlage $Verzerrung_{OLS/IV} = \emptyset\, \widehat{\beta_1} - \beta_1$ mit $\beta_1 = 10$, ergibt folgende Ergebnisse:

Verzerrung OLS-Schätzer	**Verzerrung IV-Schätzer**
-10.8769116995876	0.91347460091176

Tabelle 14: Verzerrung OLS- und IV-Schätzer (Kampe, 2020)

Wie bereits in den vorangehenden Teilaufgaben qualitativ interpretiert wurde, erreicht der IV-Schätzer deutliche bessere Ergebnisse im Vergleich zum wahren Wert β_1. Die Verzerrung des OLS-Schätzers beläuft sich hierbei auf -10.8769116995876 und die des IV-Schätzers auf 0.91347460091176. Somit wird der wahre Wert durch die OLS-Regressions-Simulation deutlich unterschätzt und durch die IV-Regression leicht überschätzt, wobei der IV-Schätzer jedoch deutlich besser schätzt.

2.2.4 Wahl des Schätzers

Wie bereits in den einhergehenden Teilaufgaben erläutert und quantitativ auf Grundlage der Simulation aufgezeigt wurde, erreicht der IV-Schätzer deutlich bessere Ergebnisse zur Schätzung des wahren Wertes β_1. Dies liegt der Verletzung von Annahme MLR.4 zu Grunde, im Rahmen dessen die ursprüngliche Regressionsgleichung durch die zugrundeliegende Endogenität keine unverzerrte Schätzung zulässt. Unter Hinzunahme des Instrumentalvariablen kann die Exogenitätsverletzung dahingehend umgangen werden, dass ein valides Instrument für die erklärende Variable verwendet wird, um die exogene Variation von x1, die mit v unkorreliert ist, zu isolieren. Es zeigt sich, dass, wie bereits erläutert, der Einbezug der Instrumentalvariable zu besseren Ergebnissen im Vergleich zum wahren Wert von β_1 führen, wodurch die Wahl auf den IV-Schätzer fällt.

2.3 Simulationsstudie mit schwachem Instrument

Im Rahmen dieser Simulationsstudie erfolgt die Wiederholung der Teilaufgaben 3.2.1-3.2.3 unter Anwendung eines schwachen Instruments. Dies wird, unterschiedlich zum eingangs definierten DGP, wie folgt definiert: $temp = -1 + 0.1 * z + v$

Zu Untersuchung der sich dadurch verändernden Schätzerverteilung werden folgend die OLS- als auch die IV-Regression wiederholt, die Histogramme erneut aufgezeigt und die Verzerrung der Schätzer berechnet. Folgende Ergebnisse sind hierbei entstanden:

	OLS-Schätzer	**IV-Schätzer**
[1,]	-4.113948	-52.112907
[2,]	-3.909071	-612.216124
[3,]	-3.845677	-60.767523
[4,]	-4.380212	-195.450636
[5,]	-4.142595	-168.846929
[6,]	-3.533642	66.878585
[7,]	-3.633914	-78.078746
[8,]	-4.684463	749.243554
[9,]	-3.584256	-3727.834877
[10,]	-3.791026	340.018183
[...,]	...	...
[1000,]	-3.041821	107.60954
colMeans	-3.853479	-11.10858

Tabelle 15: OLS- und IV-Schätzer aus Simulationsstudie – Schwaches Instrument (Kampe, 2020)

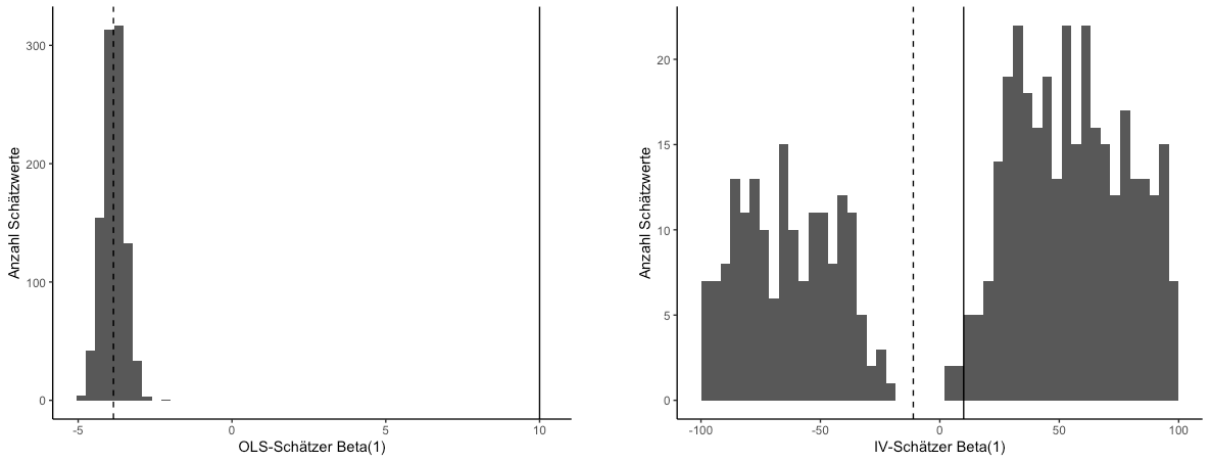

Tabelle 16: Histogramm mit IV- und OLS-Schätzer mit wahrem Wert $\beta = 10$ – Schwaches Instrument (Kampe, 2020)

Verzerrung OLS-Schätzer	**Verzerrung IV-Schätzer**
-13.853478823683	-21.108576627934

Tabelle 17: Verzerrung OLS- und IV-Schätzer – Schwaches Instrument (Kampe, 2020)

Aus den Ergebnissen in den Tabellen 13-15 ist ersichtlich, dass Anpassung hin zu einem schwachen Instrument deutliche Effekte auf die IV-Schätzung von β_1 mit sich bringt. So verändert sich das Kernergebnis der Schätzer-Verzerrung von 0.91347460091176 auf -21.108576627934 ($\Delta \approx 22.02205$) ändert. Somit ist aufgrund der Schwäche respektive der fehlenden Relevanz des Instruments *z* keine zuverlässige Schätzung mehr möglich und kann als nicht-valides Instrument definiert werden. Der OLS-Schätzer verändert sich zwar durch die Veränderung des Instruments nicht (effektiv schon, da in der Berechnung ein neuer Seed gesetzt wurde), jedoch befindet sich der OLS-Durchschnittsschätzer deutlich näher am wahren Wert $\beta_1 = 10$. Somit stellt der OLS-Schätzer unter Hinzunahme eines schwachen Instruments einen besseren Schätzer mit näheren Schätzerverteilung dar.

Literaturverzeichnis

Eugster, B., & Knaus, M. (3. März 2020A). Univariates lineares Regressionsmodell. St. Gallen, St. Gallen, Schweiz.

Eugster, B., & Knaus, M. (4. März 2020B). Multiples lineares Regressionsmodell. St. Gallen, St. Gallen, Schweiz.

Eugster, B., & Knaus, M. (5. Mai 2020C). Stichprobenselektion. St. Gallen, St. Gallen, Schweiz.

Eugster, B., & Knaus, M. (12. Mai 2020D). Endogenität I: Proxyvariablen. St. Gallen, St. Gallen, Schweiz.

Eugster, B., & Knaus, M. (13. Mai 2020E). Endogenität II: Instrumentalvariablen. St. Gallen, St. Gallen, Schweiz.

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (30. August 2019). *The IV Estimator with a Single Regressor and a Single Instrument.* Von Introduction to Econometrics with R: https://www.econometrics-with-r.org/12-1-TIVEWASRAASI.html abgerufen

Kampe, N. L. (25. Mai 2020). Kampe_Code.R. Winterthur, Zürich, Schweiz.

Eigenständigkeitserklärung

Ich erkläre hiermit,

- dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Verwendung anderer als der angegebenen Hilfsmittel verfasst habe;
- dass ich sämtliche verwendeten Quellen erwähnt und gemäss gängigen wissenschaftlichen Zitierregeln korrekt zitiert habe;
- dass ich sämtliche immateriellen Rechte an von mir allfällig verwendeten Materialien wie Bilder oder Grafiken erworben habe oder dass diese Materialien von mir selbst erstellt wurden;
- dass das Thema, die Arbeit oder Teile davon nicht bereits Gegenstand eines Leistungsnachweises einer anderen Veranstaltung oder Kurses waren, sofern dies nicht ausdrücklich mit dem Referenten /der Referentin im Voraus vereinbart wurde und in der Arbeit ausgewiesen wird;
- dass ich ohne schriftliche Zustimmung der Universität keine Kopien dieser Arbeit an Dritte aushändigen oder veröffentlichen werde, wenn ein direkter Bezug zur Universität St. Gallen oder ihrer Dozierenden hergestellt werden kann;
- dass ich mir bewusst bin, dass meine Arbeit elektronisch auf Plagiate überprüft wer- den kann und ich hiermit der Universität St. Gallen laut Prüfungsordnung das Urheberrecht soweit einräume, wie es für die Verwaltungshandlungen notwendig ist;
- dass ich mir bewusst bin, dass die Universität einen Verstoss gegen diese Eigenständigkeitserklärung sowie insbesondere die Inanspruchnahme eines Ghostwriter-Service verfolgt und dass daraus disziplinarische wie auch strafrechtliche Folgen resultieren können, welche zum Ausschluss von der Universität resp. zur Titelaberkennung führen können.

.....

Mit Einreichung der schriftlichen Arbeit stimme ich mit konkludentem Handeln zu, die Eigenständigkeitserklärung abzugeben, diese gelesen sowie verstanden zu haben und, dass sie der Wahrheit entspricht.