

Monte-Carlo-Simulation der Lohngleichheitsanalyse in Logib Modul 1 Schlussbericht



BEREICH ARBEIT

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



Eidgenössisches Departement des Innern EDI
Eidgenössisches Büro für die Gleichstellung von Frau und Mann EBG

IMPRESSUM

TITEL

Monte-Carlo-Simulation der Lohngleichheitsanalyse in Logib
Modul 1

AUTOR

Dr. Boris Kaiser

HERAUSGEBERIN

Eidgenössisches Büro für die Gleichstellung
von Frau und Mann EBG

VERTRIEB

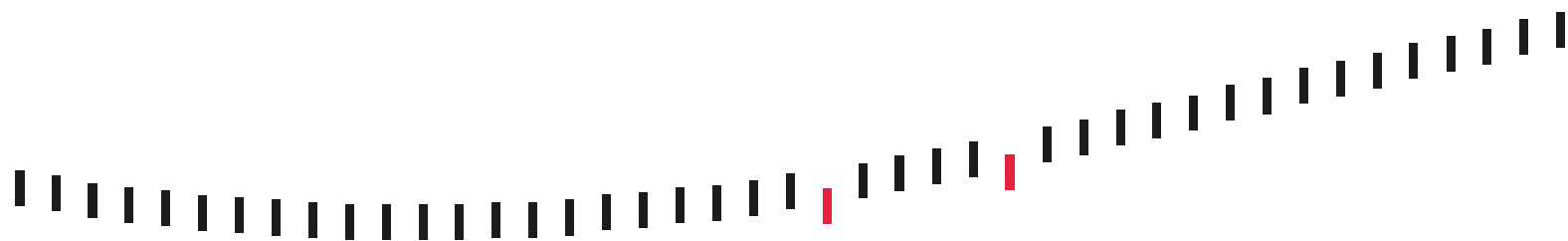
Eidgenössisches Büro für die Gleichstellung
von Frau und Mann EBG
Schwarztorstrasse 51, 3003 Bern
ebg@ebg.admin.ch
www.ebg.admin.ch

Dieser Bericht wurde im Auftrag des EBG verfasst. Die darin enthaltenen Einschätzungen und Interpretationen entsprechen nicht zwingend der Sicht des EBG.

Schlussbericht

Monte-Carlo-Simulation der Lohnungleichheitsanalyse in Logib Modul 1

Basel | 09.06.2023



Impressum

Monte-Carlo-Simulation der Lohnungleichheitsanalyse in Logib Modul 1

Schlussbericht

09.06.2023 (Version 2)

Auftraggeberin: Eidgenössisches Büro für die Gleichstellung von Frau und Mann EBG

Autor: Dr. Boris Kaiser

Verantwortlich seitens Auftraggeberin: Dr. Jeremy Kolly, Dr. Simon Hilber

Projektleitung seitens Auftragnehmerin: Dr. Boris Kaiser

Zitierweise:

Kaiser, Boris (2023). Monte-Carlo-Simulation der Lohnungleichheitsanalyse in Logib Modul 1. BSS Volkswirtschaftliche Beratung. Studie im Auftrag des EBG.

BSS Volkswirtschaftliche Beratung AG

Aeschengraben 9

4051 Basel

T +41 61 262 05 55

contact@bss-basel.ch

www.bss-basel.ch

© 2022 BSS Volkswirtschaftliche Beratung AG

Inhalt

Tabellen	iii
Abbildungen.....	iv
Zusammenfassung	1
1 Einleitung.....	2
2 Statistische Grundlagen.....	3
2.1 Hypothesentests	3
2.2 Implementierung in Logib	4
3 Simulationsdesign.....	5
3.1 Datenaufbereitung	5
3.2 Populationsmodell.....	6
3.3 Simulation.....	10
4 Ergebnisse.....	11
4.1 Geschlechter-Koeffizient und Standardfehler	11
4.2 Mächtigkeit nach Grösse und Frauenanteil	13
4.3 Tiefere Toleranzschwelle.....	15
4.4 Modifizierter Test mit numerischer Toleranzschwelle.....	16
4.5 Heteroskedastizität und robuste Standardfehler	20
4.6 Spezifikation der Lohnfunktion.....	22
4.7 Zusätzliche Analysen	23
5 Fazit.....	25
6 Quellenangaben	28
A. Anhang.....	29
A.1 Populationsdaten	29
A.2 Weitere Simulationsergebnisse	37

Tabellen

Tabelle 1: Relevante Variablen für Logib Modul 1.....	5
Tabelle 2: Populationsregression mit echten und simulierten Lohndaten.....	9
Tabelle 3: Geprüfte Parameterwerte im Algorithmus.....	11
Tabelle 4: Ausschlüsse aus den LSE-Daten nach Filter.....	29
Tabelle 5: Verteilung der Kontrollvariablen nach Geschlecht.....	29
Tabelle 6: Informationskriterien für den Fit der drei Verteilungen.....	35
Tabelle 7: Populationsregression mit echten und simulierten Lohndaten, vollständig	35

Abbildungen

Abbildung 1: Konstruktion der Variablen des Populationsmodells	7
Abbildung 2: Boxplot Geschlechter-Koeffizient und durchschnittliche Standardfehler	13
Abbildung 3: Mächtigkeit des zweiseitigen t-Tests, relativer Frauenanteil.....	14
Abbildung 4: Mächtigkeit des einseitigen t-Tests (-0.05), relativer Frauenanteil.....	15
Abbildung 5: Mächtigkeit des einseitigen t-Tests (-0.03), relativer Frauenanteil.....	16
Abbildung 6: Mächtigkeit des modifizierten Tests, Schranke 0.05, relativer Frauenanteil.....	18
Abbildung 7: Mächtigkeit des modifizierten und herkömmlichen Tests, ($b=-0.08$)	19
Abbildung 8: Alpha-Fehler des modifizierten und herkömmlichen Tests, ($b=0$).....	20
Abbildung 9: Bias im Standardfehler nach Varianzfunktion und Schätzverfahren	21
Abbildung 10: Mächtigkeit nach Varianzfunktion und Schätzverfahren für Standardfehler	22
Abbildung 11: Bias im Geschlechter-Koeffizienten nach Lohnfunktion	23
Abbildung 12: Signifikante Ergebnisse bei mehrfachem Testen	25
Abbildung 13: Quantil-Quantil-Grafik der OLS-Residuen mit drei Verteilungen.....	35
Abbildung 14: Boxplot Geschlechter-Koeffizient und Standardfehler, $\beta = -0.08$	37
Abbildung 15: Mächtigkeit des zweiseitigen t-Tests, absoluter Frauenanteil.....	38
Abbildung 16: Mächtigkeit des einseitigen t-Tests (-0.03), absoluter Frauenanteil	39
Abbildung 17: Mächtigkeit des einseitigen t-Tests (-0.05), absoluter Frauenanteil	40
Abbildung 18: Mächtigkeit des modifizierten Tests, Schranke 0.025, relativer Frauenanteil.....	41
Abbildung 19: Mächtigkeit des modifizierten und herkömmlichen Tests, Toleranzschwelle 0.025, ($b=-0.08$).....	42
Abbildung 20: Alpha-Fehler des modifizierten und herkömmlichen Tests, Toleranzschwelle 0.025, ($b=0$).....	43
Abbildung 21: Alpha-Fehler nach Varianzfunktion und Schätzverfahren für Standardfehler	44
Abbildung 22: Mächtigkeit nach Spezifikation der Lohnfunktion	45
Abbildung 23: Mächtigkeit nach Signifikanzniveau	46
Abbildung 24: Häufigkeit des Alpha-Fehlers nach Signifikanzniveau	47
Abbildung 25: Mächtigkeit zweiter t-Test bei Ablehnung des ersten t-Tests.....	48

Zusammenfassung

Für die Analyse der Lohnungleichheit zwischen Frauen und Männern innerhalb von Unternehmen stellt der Bund den Arbeitgebenden das Standard-Analyse-Tool «Logib» zur Verfügung. Das Modul 1 von Logib richtet sich an Unternehmen ab 50 Beschäftigten und besteht im Kern aus einer OLS-Regression. Anhand von statistischen Tests wird beurteilt, ob in einem Unternehmen ein signifikanter, geschlechterspezifischer Lohneffekt vorliegt oder nicht. Dabei gelangt eine Toleranzschwelle von 0.05 (d.h. 5%) zur Anwendung.

Das Ziel der vorliegenden Studie besteht darin, zu untersuchen, wie gut die durchgeführten Tests in Logib Modul 1 unerklärbare Lohnunterschiede zwischen Frauen und Männern innerhalb von Unternehmen detektieren können. Die zentrale statistische Kennzahl ist die sogenannte *Mächtigkeit*: Diese sagt aus, wie häufig ein Test korrekte Ergebnisse liefert, wenn tatsächlich ein geschlechtsspezifischer Lohneffekt existiert. Konkret wird in der Studie die Frage adressiert, wie die Mächtigkeit des Tests von verschiedenen Faktoren, namentlich Unternehmensgrösse, Geschlechterverhältnis und Grösse des Geschlechter-Koeffizienten, abhängt. Weiter wird untersucht, wie sich bestimmte methodische Modifikationen auf die Performance der Tests in Logib auswirken.

Das methodische Vorgehen besteht aus einer sogenannten *Monte-Carlo-Simulation*. In dieser Simulation werden eine Vielzahl künstlicher Lohndatensätze generiert, mit denen die Logib-Analyse durchgeführt wird. Da der wahre geschlechtsspezifische Effekt experimentell kontrolliert wird, lassen sich Aussagen zur Performance der Tests ableiten. Damit die Simulation möglichst realitätsnah ist, wird für die Kalibrierung der Simulation die Schweizerische Lohnstrukturerhebung (LSE) des Jahres 2018 genutzt.

Die Resultate der Monte-Carlo-Simulation zeigen, dass die Mächtigkeit des Tests massgeblich von drei Faktoren abhängt. Der erste und naheliegende ist die Unternehmensgrösse: In mittleren und grösseren Unternehmen werden unerklärbare Lohnunterschiede zwischen den Geschlechtern mit Logib weitaus häufiger aufgedeckt als in kleinen Unternehmen. Zweitens spielt auch das Geschlechterverhältnis eine entscheidende Rolle. Sind in einem kleinen Unternehmen mit 100 Mitarbeitenden 50% Frauen beschäftigt, erkennt der erste t-Test einen Lohneffekt von -10% nahezu immer (97%); bei 5% Frauen im Unternehmen sind es nur noch 40%. Drittens, und wenig überraschend, ist entscheidend, wie gross das Ausmass der Lohnunterschiede im Unternehmen ist. Bei sehr grossen Geschlechter-Koeffizienten ab -0.2 (-20%) liefert der zweite t-Test bei einer Toleranzschwelle von -5% in fast allen Szenarien das korrekte Ergebnis. Bei Geschlechter-Koeffizienten von -10% ist das hingegen nicht der Fall.

Weitere Resultate zeigen unter anderem, wie sich die Performance der Tests bei bestimmten Modifikationen verändern würde. Wird zum Beispiel die Toleranzschwelle des zweiten Tests auf -3% reduziert, steigt die Wahrscheinlichkeit, dass geschlechterspezifische Lohneffekte von -10% korrekt detektiert werden, während falsch-positive Ergebnisse bei einem wahren Lohneffekt von null nur geringfügig zunehmen. Dies gilt vor allem für kleinere und mittlere Unternehmen. Zweitens wurde ein modifizierter zweistufiger Test mit einer *numerischen* Toleranzschwelle von 0.05 ge-

prüft. Das heisst, der Lohngleichheitstest gilt als nicht bestanden, wenn der Geschlechter-Koeffizient signifikant von null verschieden ist und der Schätzwert im Betrag grösser ist als 0.05. Der modifizierte Test detektiert Lohnungleichheit vor allem bei kleinen und mittleren Unternehmen deutlich häufiger als der zweistufige t-Test. Die grössere Mächtigkeit des Tests geht jedoch mit einer markant höheren Rate an falsch-positiven Ergebnissen einher. Drittens kommt es zu einem Performanceverlust, weil die konventionellen Standardfehler, die in Logib verwendet werden, in bestimmten Szenarien nach oben verzerrt sind. Durch die Verwendung robuster Standardfehler könnte die Mächtigkeit etwas gesteigert werden. Viertens könnte eine Erhöhung des Signifikanzniveaus bei kleinen Unternehmen die Performance des ersten t-Tests – unter Inkaufnahme eines häufigeren Alpha-Fehlers – moderat steigern.

Insgesamt liefert die Monte-Carlo-Simulation wertvolle Erkenntnisse, die in die Beurteilung von Logib Modul 1 sowie dessen potenzielle Weiterentwicklung einfließen können. Zum einen können die Ergebnisse als Informationsgrundlage zur Frage, ab welcher Unternehmensgrösse bzw. ab welchem Mindest-Frauenanteil das Modul 1 empfohlen wird, hinzugezogen werden. Zum anderen zeigen die Ergebnisse auf, wie sich allfällige Modifikationen in Bezug auf die Wahl der Toleranzschwelle, das gewählte Signifikanzniveau oder die Schätzmethode für die Standardfehler auf die Performance der t-Tests auswirken würden. Als Limitation ist die im Rahmen der Simulation getroffene *Annahme* zu nennen, dass das Lohnmodell alle relevanten Kontrollvariablen enthält. Der Einfluss möglicher fehlender Variablen («Omitted Variable Bias») und deren Implikationen auf die Tests sind somit nicht Gegenstand dieser Studie.

1 Einleitung

Für die Analyse der Lohnungleichheit zwischen Frauen und Männern innerhalb von Unternehmen stellt der Bund den Arbeitgebenden das Standard-Analyse-Tool «Logib» zur Verfügung. Logib besteht aus zwei Modulen: Modul 1 basiert auf einer statistischen Regressionsmethode und richtet sich an mittlere und grössere Unternehmen ab 50 Beschäftigten, während Modul 2 für kleine Unternehmen konzipiert ist. Konkret besteht Modul 1 aus einer linearen OLS-Regression der standardisierten, logarithmierten Löhne auf einen Geschlechter-Indikator und eine Reihe von Kontrollvariablen. Anhand von statistischen t-Tests wird beurteilt, ob in einem Unternehmen ein signifikanter, geschlechterspezifischer Lohneffekt vorliegt oder nicht. Dabei gelangt eine Toleranzschwelle von 0.05 (d.h. 5%) zur Anwendung.¹

Das Ziel der vorliegenden Studie besteht darin, zu untersuchen, wie gut die durchgeführten t-Tests in Logib Modul 1 unerklärbare Lohnunterschiede zwischen Frauen und Männern innerhalb von Unternehmen detektieren können. Dazu wird evaluiert, wie häufig die t-Tests korrekte Ergebnisse liefern, wenn tatsächlich ein geschlechtsspezifischer Lohneffekt existiert. Diese Eigenschaft eines Tests wird als *Mächtigkeit* oder *Trennschärfe* (englisch: *power*) bezeichnet. Konkret wird in der Studie folgende zentrale Frage adressiert: Wie hängt die *Mächtigkeit* des Tests von der Stichprobengrösse (Unternehmensgrösse), vom Geschlechterverhältnis, von der Grösse des Geschlechter-Koeffizienten sowie von der Form der Fehlervarianz (Homo- bzw. Heteroskedasti-

¹ Eine ausführliche Beschreibung findet sich in EBG (2021).

zität) ab? Ergänzend wird untersucht, wie stark ein von der Stichprobengrösse abhängiges Signifikanzniveau die Mächtigkeit beeinflussen kann. Zudem wird die Frage adressiert, inwieweit die sequenzielle Durchführung zweiter t-Tests ein Problem darstellt.

Das methodische Vorgehen besteht aus einer sogenannten *Monte-Carlo-Simulation*. In dieser Simulation werden eine Vielzahl künstlicher Datensätze generiert, mit denen die Logib-Analyse durchgeführt wird. Dabei wird der wahre geschlechtsspezifische Effekt experimentell kontrolliert, sodass die Eigenschaften der t-Tests unter bekannten Bedingungen beurteilt und verglichen werden können. Damit die simulierten Daten möglichst ähnliche statistische Eigenschaften wie reale Lohndaten aufweisen, wird für die Kalibrierung der Simulation die Schweizerische Lohnstrukturerhebung (LSE) des Jahres 2018 genutzt.

Dieser Bericht ist wie folgt strukturiert. Zunächst wird in Kapitel 2 auf einige statistische Konzepte eingegangen. Zweitens wird in Kapitel 3 aufgezeigt, wie die Monte-Carlo-Simulation konzipiert und durchgeführt wird. Kapitel 4 widmet sich den wichtigsten Ergebnissen der Simulation. Kapitel 5 rundet den Bericht mit einem Fazit ab. Ergänzende Informationen und weitere Ergebnisse sind im Anhang A aufgeführt.

2 Statistische Grundlagen

2.1 Hypothesentests

Hypothesen in Bezug auf einzelne statistische Parameter – häufig Regressionskoeffizienten – werden meist mit sogenannten t-Tests durchgeführt. Im vorliegenden Fall ist der statistische Parameter (β) der Geschlechter-Koeffizient in einer betrieblichen Lohnregression. Der t-Test bezieht sich auf die Nullhypothese, dass der Parameter β einem bestimmten Wert, meist null, entspricht. Die Alternativhypothese besagt, dass die Nullhypothese falsch ist.

Die *Mächtigkeit* eines t-Tests beantwortet die Frage, wie häufig der t-Test die Nullhypothese tatsächlich ablehnt, wenn sie falsch ist. Im vorliegenden Kontext heisst das: Die Mächtigkeit ist die Häufigkeit eines signifikanten, geschlechterspezifischen Effekts, wenn tatsächlich ein solcher Effekt vorliegt. Allgemein hängt die Mächtigkeit eines t-Tests von folgenden Eigenschaften der *Daten* ab²:

- *Wahre Effektgrösse (β)*: Je grösser der wahre Geschlechter-Koeffizient, desto grösser die Mächtigkeit des Tests, Effekte grösser als null (oder grösser als c%) korrekt zu detektieren.

² Die Mächtigkeit hängt von der t-Statistik ab, diese ist definiert als $t = (\hat{\beta} - c) / se(\hat{\beta})$, wobei $\hat{\beta}$ den geschätzten Geschlechter-Koeffizienten, c den Toleranzwert und $se(\hat{\beta})$ den Standardfehler darstellen. Aus Wooldridge (2015, Theorem 3.2) geht hervor, dass sich der Standardfehler wie folgt in mehrere Bestandteile zerlegen lässt: $se(\hat{\beta}) = \sigma / \sqrt{(N - 1)Var(D)(1 - R_D^2)}$. $\sigma = Var(u) = Var(u|X) = Var(Y|X)$ ist die Streuung der Löhne gegeben die erklärenden Variablen, N die Stichprobengrösse, $Var(D)$ die Varianz des Geschlechter-Indikators und R_D^2 misst, wie stark das Geschlecht mit den Kontrollvariablen zusammenhängt.

- *Stichprobenumfang (N)*: Mit wachsender Unternehmensgrösse steigt die Mächtigkeit des Tests.
- *Geschlechterverhältnis*: Die Mächtigkeit ist umso grösser, je stärker die Varianz im Geschlechter-Indikator (D). Bei binären Variablen wie dem Geschlecht ist die Varianz am grössten, wenn der Anteil der beiden Gruppen jeweils 50% beträgt. Folglich ist die Mächtigkeit umso grösser, je ausgeglichener das Geschlechterverhältnis ist.³
- *Streuung in der abhängigen Variable (Y)*: Mit zunehmender Streuung der Löhne (gegeben die erklärenden Variablen) nimmt die Mächtigkeit des Tests ab.

Nebst den datenbezogenen Faktoren spielt auch eine Eigenschaft des *Tests* eine massgebliche Rolle: das Signifikanzniveau (α). Das Signifikanzniveau ist die *theoretische* Fehlerhäufigkeit des t-Tests, wenn die Nullhypothese wahr ist (sogenannter «Alpha-Fehler» oder «Fehler 1. Art»). Es beschreibt also, wie oft ein geschlechtsspezifischer Effekt fälschlicherweise detektiert wird, wenn kein solcher vorliegt. Zwischen Signifikanzniveau und Mächtigkeit besteht ein Zielkonflikt: Je tiefer das Signifikanzniveau gewählt wird, desto niedriger ist auch die Mächtigkeit eines Tests.

Die Mächtigkeit eines t-Tests lässt sich empirisch nicht direkt überprüfen, da der wahre Wert eines Parameters nicht bekannt ist. In Monte-Carlo-Simulationen wird der interessierende Parameter experimentell festgelegt, sodass Aussagen zur Mächtigkeit ermöglicht werden.

2.2 Implementierung in Logib

Logib Modul 1 besteht aus einer linearen OLS-Regression auf Ebene eines Unternehmens, wobei die standardisierten, logarithmierten Löhne auf einen Geschlechter-Indikator und eine Reihe von Kontrollvariablen (Ausbildungsjahre, potenzielle Erwerbserfahrung, Dienstjahre, Kompetenzniveau und berufliche Stellung) regressiert werden. Die Kontrollvariablen werden im vorliegenden Kontext auch als «objektive, lohnrelevante und nichtdiskriminierende Faktoren» bezeichnet.⁴ Die Standardfehler werden mit dem «konventionellen» Schätzverfahren berechnet, welches unterstellt, dass der Störterm homoskedastisch ist.⁵ Zunächst wird mit einem zweiseitigen t-Test geprüft, ob der Geschlechter-Koeffizient signifikant von null verschieden ist. Wenn ja, wird ein zweiter, einseitiger t-Test durchgeführt, ob der Geschlechter-Koeffizient im Betrag grösser ist als 0.05. Auf beiden Stufen wird ein Signifikanzniveau von 5% verwendet; dies ist der gängigste Wert in der wissenschaftlichen Literatur. Für eine detaillierte Beschreibung von Modell und Methode wird auf EBG (2021, Kapitel 3.2) verwiesen.

³ Zwei Beispiele: Bei einem Frauenanteil von 50% ist $sd(D) = \sqrt{0.5(1 - 0.5)} = 0.5$. Bei einem Frauenanteil von 10% ist $sd(D) = \sqrt{0.1(1 - 0.1)} = 0.3$. Die t-Statistik steigt somit um 67%, wenn der Frauenanteil, ceteris paribus, von 10% auf 50% steigt.

⁴ Vgl. dazu die Konformitätserklärung des Standard-Analyse-Tools Logib (Seite 2). Link: [https://www.ebg.admin.ch/dam/ebg/de/dokumente/lohnungleichheit/logib/konformitaetserklaerung_logib.pdf.download.pdf/Konformit%C3%A4tserkl%C3%A4rung_Standard-Analyse-Tool%20\(Logib\)_V2020.1.pdf](https://www.ebg.admin.ch/dam/ebg/de/dokumente/lohnungleichheit/logib/konformitaetserklaerung_logib.pdf.download.pdf/Konformit%C3%A4tserkl%C3%A4rung_Standard-Analyse-Tool%20(Logib)_V2020.1.pdf)

⁵ Die konventionellen Standardfehler sind definiert als $se(\hat{\beta})_{konv} = \hat{\sigma} \sqrt{(X'X)^{-1}}$, wobei $\hat{\sigma}$ die geschätzte Standardabweichung des Störterms und X die Matrix der erklärenden Variablen des Modells darstellen (Hansen 2022, Kapitel 4.13).

3 Simulationsdesign

Das Kernstück dieser Studie bildet eine Monte-Carlo-Simulation. In einer solchen Simulation werden Datengenerierungsprozesse definiert und anhand von wiederholt gezogenen Zufallsstichproben die relevanten Koeffizienten geschätzt und die t-Tests durchgeführt. Durch eine grosse Anzahl Wiederholungen lassen sich die empirischen Eigenschaften eines Schätzverfahrens, unter anderem die Mächtigkeit eines statistischen Tests, bei gegebenen Eigenschaften der Daten ermitteln.

Damit die Simulation in Bezug auf den Untersuchungsgegenstand möglichst realitätsnah ist, nutzen wir als Grundlage die Schweizerische Lohnstrukturerhebung (LSE) des Jahres 2018. Entsprechend handelt es sich um eine *empirisch kalibrierte* Monte-Carlo-Simulation. Zunächst wird nachfolgend die Aufbereitung der LSE erläutert (Abschnitt 3.1). Zweitens wird aufgezeigt, wie basierend auf der LSE ein Populationsmodell spezifiziert wird, das die Grundlage für die Simulationen bildet (Abschnitt 3.2). Drittens wird das Simulationsverfahren beschrieben (Abschnitt 3.3).

3.1 Datenaufbereitung

In einem ersten Schritt werden in der LSE die für das Logib-Modell relevanten Variablen definiert. Tabelle 1 liefert einen Überblick. Da zur Validierung das R-Paket «logib» des EBG verwendet wird, sind auch die dort verwendeten, englischen Bezeichnungen aufgeführt. Wo notwendig, werden die LSE-Variablen transformiert bzw. umkodiert, sodass sie den Definitionen in Logib entsprechen.

Tabelle 1: Relevante Variablen für Logib Modul 1

Variable in Logib	Name in R-Paket «logib»	Variablen in der LSE
Geschlecht (Mann=0, Frau = 1)	sex	geschle
Anzahl Bildungsjahre	years_of_training	ausbild (Kodierung gemäss EBG 2021, Kap. 3.2.2)
Potenzielle Erwerbserfahrung in Jahren	years_of_earning	$\max(\text{alter} - \text{years_of_training} - 6, 0)$
Anzahl Dienstjahre	years_of_service	dienstja
Kompetenzniveau (4 Stufen)	skill_level	va_ps07
Berufliche Stellung (5 Stufen)	professional_position	berufst
Standardisierter Bruttomonatslohn, logarithmiert	$\log(\text{standardized_salary})$	$\log(\text{mbls})$

In einem zweiten Schritt wird die relevante Stichprobe definiert. Die Ausgangslage bildet die Gesamtwirtschaft, also der private und öffentliche Sektor, da Logib in beiden Sektoren zur Anwen-

dung kommt. Für die Simulation ist es notwendig, unvollständige Beobachtungen auszuschliessen. Dazu werden erstens nur Beobachtungen beibehalten, bei denen die für das Logib-Modell relevanten Variablen keine fehlenden Werte enthalten (Filter A). Zweitens wird eine Reihe weiterer Filter gesetzt, um Beobachtungen mit extremen oder unplausiblen Werten auszuschliessen (Filter B bis E).⁶ Diese Filter entsprechen jenen aus der BFS-Studie zur Analyse der gesamtwirtschaftlichen Lohnunterschiede (vgl. Seite 15 in Kaiser und Möhr 2021). Insgesamt werden durch die Filterung rund 25% der Beobachtungen der LSE 2018 ausgeschlossen. Nähere Angaben zu den einzelnen Filtern finden sich in Tabelle 4 im Anhang.

Drittens wird ein Datensatz generiert, der von der Struktur her der Grundgesamtheit der Beschäftigten im Schweizer Arbeitsmarkt entspricht. Dazu wird aus dem LSE-Datensatz eine Bootstrap-Ziehung durchgeführt, deren Grösse der Summe der Gewichte entspricht.⁷ Die Wahrscheinlichkeiten für die Ziehung einzelner Beobachtungen sind entsprechend proportional zu den Gewichten in der LSE. Im Ergebnis resultiert ein *Populationsdatensatz* mit rund 2.97 Mio. Beobachtungen; dieser stellt die Grundlage für die Monte-Carlo-Simulation dar.

3.2 Populationsmodell

Im Populationsdatensatz wird nun ein sogenanntes *Populationsmodell* definiert. Dieses legt fest, wie die einzelnen Variablen miteinander zusammenhängen. Die Annahmen, die dem Populationsmodell zugrunde liegen, sowie die Konstruktion der Variablen werden nachfolgend erläutert.

3.2.1 Annahmen

Dem Populationsmodell liegen drei zentrale Annahmen zugrunde:

- *Keine fehlenden Kontrollvariablen*: Diese Annahme besagt, dass für die Lohnbestimmung keine weiteren Variablen relevant sind, als jene die im Logib-Modell bereits enthalten sind. Die Analyse von «Omitted Variable Bias» ist somit nicht Gegenstand dieser Studie.
- *Korrekt spezifizierte Lohnfunktion*: Logib besteht aus einer *linearen* Lohngleichung, wobei der Geschlechter-Indikator und die Kontrollvariablen untereinander nicht interagiert werden.⁸ Im Populationsmodell wird die Annahme unterstellt, dass die funktionale Form des Logib-Modells die Lohnbestimmung der Unternehmen korrekt abbildet. Das bedeutet unter anderem, dass der geschlechtsspezifische Lohnunterschied für alle Ausprägungen der Merkmale der Beschäftigten genau gleich gross ist. Ein Beispiel: Die Grösse des geschlechtsspezifischen Effekts

⁶ Dazu gehören Beobachtungen mit sehr tiefem oder hohem Beschäftigungsgrad von <20% oder >150% (Filter B); standardisierte Bruttolöhne unterhalb 1/3 des Medianlohns oder oberhalb des 15-fachen des Medianlohns (Filter C und D) sowie unplausible Kombinationen von Beschäftigungsgrad und Arbeitsstunden (Filter E)

⁷ Bei einer Bootstrap-Ziehung werden zufällig Beobachtungen aus einem Datensatz gezogen, wobei einzelne Beobachtungen mehrfach gezogen werden können. Der Bootstrap ist somit eine «Zufallsziehung mit Zurücklegen».

⁸ Konkret hat das lineare Lohnmodell die folgende Form: $Y = \beta_0 + D\beta + X\beta_1 + u$, wobei Y der logarithmierte Lohn, D der Geschlechter-Indikator, X der Vektor der Kontrollvariablen und u der Störterm darstellen. Der Koeffizient β_0 ist eine Konstante, β ist der Geschlechter-Koeffizient und β_1 ist der Vektor der Koeffizienten der Kontrollvariablen.

ist also auf allen Kompetenz- und Kaderstufen gleich gross. Diese Annahme wird in einem Teil der Simulation gelockert (siehe Abschnitt 3.2.4).

- *Homoskedastischer Störterm*: Weiter wird unterstellt, dass der Störterm homoskedastisch ist, was bedeutet, dass dessen Varianz nicht mit den Modellvariablen zusammenhängt. Diese Annahme hat keinen Einfluss auf den geschlechtsspezifischen Effekt, potenziell jedoch auf den Standardfehler und die Mächtigkeit des t-Tests. Die Annahme wird in einem Teil der Simulation gelockert (siehe Abschnitt 3.2.4).

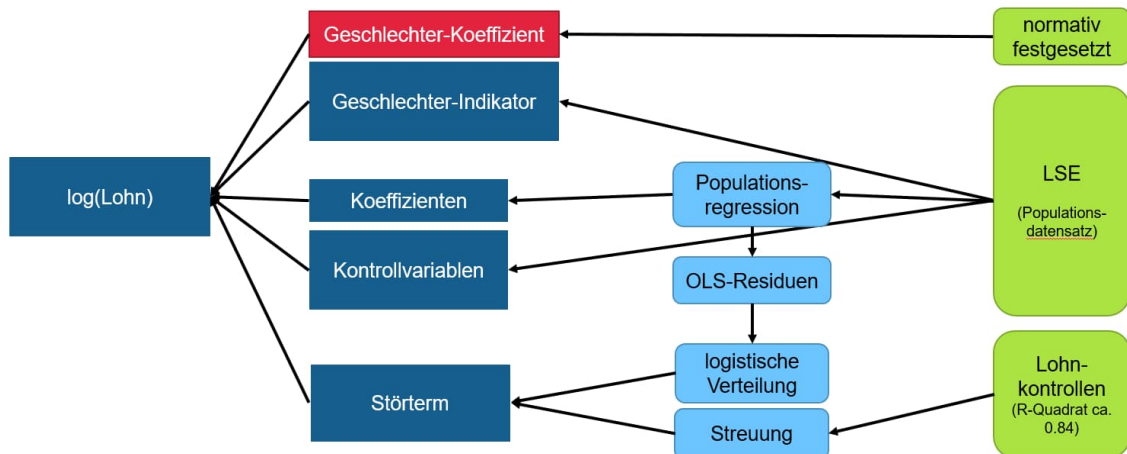
3.2.2 Konstruktion der Variablen

Abbildung 1 stellt schematisch dar, wie die Variablen im Populationsdatensatz und die Koeffizienten der Lohnfunktion konstruiert werden, um die simulierten Lohnvariablen zu bilden. Diese haben drei additive Bestandteile:

- *«Exogener» Teil*: Die Kontrollvariablen multipliziert mit den jeweiligen Regressionskoeffizienten.
- *Geschlechtsspezifischer Teil*: Der Geschlechter-Indikator multipliziert mit dem experimentell festgelegten Geschlechter-Koeffizienten.
- *Störterm*: Eine aus einer Wahrscheinlichkeitsverteilung gezogene Zufallsvariable.

Nachfolgend gehen wir der Reihe nach auf die einzelnen Bestandteile ein.

Abbildung 1: Konstruktion der Variablen des Populationsmodells



Anmerkungen: Die Grafik zeigt schematisch, wie die einzelnen Variablen im Populationsdatensatz und die Koeffizienten konstruiert werden, um den simulierten logarithmierten Lohn zu bilden.

Kontrollvariablen und Geschlechter-Indikator: Der Geschlechter-Indikator und die Kontrollvariablen im Populationsdatensatz basieren auf den LSE-Daten und fließen direkt in die Monte-Carlo-Simulation ein. Abweichende manuelle Anpassungen der Verteilungen werden keine vorgenommen. Die relativen Häufigkeiten der Ausprägungen der Kontrollvariablen sind differenziert nach Geschlecht in Tabelle 5 im Anhang ersichtlich.

Regressionskoeffizienten: Die Regressionskoeffizienten der Kontrollvariablen werden einer OLS-Populationsregression (Logib-Spezifikation) entnommen (vgl. Tabelle 2 weiter unten). Der Geschlechter-Koeffizient wird normativ auf einen bestimmten Wert festgesetzt. Um die Eigenschaften der t-Tests bei unterschiedlich ausgeprägten Lohnunterschieden zu testen, werden folgende Werte des Geschlechter-Koeffizienten verwendet: 0, -0.01, -0.02, ..., -0.08, -0.1, -0.2, -0.3.⁹ Mit anderen Worten wird für jeden dieser Werte eine separate, simulierte Lohnvariable gebildet.

Störterm: Für die Simulation von logarithmierten Löhnen wird ein *Störterm* basierend auf einer stetigen Wahrscheinlichkeitsfunktion benötigt. Damit die simulierten Löhne möglichst ähnliche Eigenschaften wie die echten Löhne aufweisen, muss eine möglichst passende Verteilung verwendet werden. Wir evaluieren drei mögliche Verteilungen auf ihre Eignung hin: die Normalverteilung, die Student-t-Verteilung und die logistische Verteilung. Es zeigt sich, dass die logistische Verteilung die Daten am besten beschreibt; Details zum Vorgehen finden sich in Anhang A.1.3. Die Standardabweichung des Störterms wird so kalibriert, dass die simulierten Lohnvariablen in etwa die gleiche Streuung wie in den Lohnkontrollen des EBG aufweisen. Dazu wird die Standardabweichung so gewählt, dass in der OLS-Populationsregression ungefähr ein R-Quadrat von 0.84 resultiert; dabei handelt es sich um den durchschnittlichen Wert in den Lohnkontrollen des EBG (vgl. Tabelle 2 weiter unten).¹⁰

3.2.3 Populationsregression

Anhand einer OLS-Regression im Populationsdatensatz lässt sich zeigen, dass der Geschlechter-Koeffizient auf Populationsebene «korrekt» geschätzt wird, siehe Tabelle 2. Die Zeile «sexF» zeigt den Geschlechter-Koeffizienten für die reale Lohnvariable aus der LSE (Spalte «real») sowie für ausgewählte *simulierte* Lohnvariablen (Spalten «simuliert») mit einem vorgängig festgesetzten Koeffizienten von 0, -0.05 und -0.10. In den simulierten Lohndaten ist das R-Quadrat ungefähr 0.84, was bedeutet, dass der vom Modell erklärte Variationsanteil gut vergleichbar ist mit den durchgeführten Lohnkontrollen des EBG. Die vollständigen Resultate der Populationsregression sind in Tabelle 7 im Anhang aufgeführt.

⁹ Für die Analyse des modifizierten zweistufigen Tests in Abschnitt 4.4 gelangen leicht andere Werte zur Anwendung.

¹⁰ Das R-Quadrat ist der Anteil der vom Modell erklärten Variation an der gesamten Variation in den Löhnen. Der Zusammenhang zwischen dem adjustierten R-Quadrat und dem Störterm ist wie folgt: $adj. R^2 = 1 - Var(u)/Var(Y)$, siehe Wooldridge (2015, Kap. 6.3). Das heisst, je kleiner $Var(u)$, desto grösser ist das adjustierte und das normale R-Quadrat.

Tabelle 2: Populationsregression mit echten und simulierten Lohndaten

	real	simuliert (b=0)	simuliert (b=-0.05)	simuliert (b=-0.10)
sexF	-0.114*** (0.0003)	0.00001 (0.0001)	-0.050*** (0.0001)	-0.100*** (0.0001)
Beobachtungen	2,970,032	2,970,032	2,970,032	2,970,032
R-Quadrat	0.524	0.839	0.842	0.848
Standardfehler Residuen	0.270	0.120	0.120	0.120

Anmerkungen: Die Signifikanzniveaus sind *p<0.05; **p<0.01; ***p<0.001. Standardfehler werden in Klammern ausgewiesen.

3.2.4 Erweiterungen

Das oben spezifizierte Populationsmodell unterstellt teilweise relativ starke Annahmen. Mit erweiterten simulierten Lohnvariablen werden zwei zugrundeliegende Annahmen kritisch geprüft.

Flexiblere Lohnfunktion

Die simulierten Lohnvariablen in Abschnitt 3.2.2 unterstellen, dass die Lohnfunktion exakt der Logib-Modellspezifikation entspricht. Diese Annahme wird etwas abgeschwächt, indem Lohnvariablen simuliert werden, die auf einer flexibleren Spezifikation beruhen. Diese erlaubt, dass sämtliche Einflüsse der erklärenden Variablen geschlechterspezifisch sein können. Bei dieser flexibleren Spezifikation kann sich das Ausmass der Lohnungleichheit somit je nach Merkmalen der Beschäftigten unterscheiden.¹¹

Heteroskedastizität

Die Lohnvariablen in Abschnitt 3.2.2 haben einen homoskedastischen Störterm, was die Verwendung konventioneller Standardfehler (wie zurzeit in Logib implementiert) rechtfertigt. Dies ist allerdings keine realistische Eigenschaft von Lohndaten, sondern stellt eine Vereinfachung dar. Eine realistischere Annahme ist Heteroskedastizität: Das heisst, die Varianz des Störterms hängt vom Geschlecht und den Kontrollvariablen ab. Dazu spezifizieren wir bei der Zufallsziehung aus der logistischen Verteilung eine heteroskedastische Varianzfunktion, die wir mit einer nichtlinearen Regression der quadrierten OLS-Residuen auf die Kontrollvariablen schätzen. Die Standardabweichung des Störterms wird zudem geschlechterspezifisch gewählt, wobei der Faktor aus den LSE-Daten entnommen wird. Im Fall von Heteroskedastizität ist es empfehlenswert, robuste Standardfehler zu verwenden.

¹¹ Konkret hat das flexiblere Lohnmodell die folgende Form: $Y = \beta_0 + D\beta + X\beta_1 + DX\beta_2 + u$, wobei Y der logarithmierte Lohn, D der Geschlechter-Indikator, X der Vektor der Kontrollvariablen und u der Störterm darstellen. Dies ist leicht flexibler als die Logib-Spezifikation $Y = \beta_0 + D\beta + X\beta_1 + u$. Das flexiblere Lohnmodell ist mathematisch äquivalent zu zwei separaten Lohnregressionen, so wie sie in der Oaxaca-Blinder-Zerlegung verwendet werden (Blinder 1973; Oaxaca 1973).

3.3 Simulation

Nachdem alle Variablen im Populationsdatensatz generiert wurden, lässt sich die Monte-Carlo-Simulation durchführen. Der verwendete Algorithmus besteht aus vier Schritten:

1. *Stichprobenziehung*: Aus der Population wird eine Zufallsstichprobe der Grösse N und einem vorgängig festgelegten Frauenanteil gezogen. Diese Stichprobe repräsentiert die Daten eines fiktiven Unternehmens. Bei der Stichprobenziehung werden die beiden Parameter je nach Szenario variiert, um deren Einfluss auf die Mächtigkeit des t-Tests zu untersuchen (siehe Tabelle 3 weiter unten).
2. *Modellschätzung*: Das Logib-Modell wird für jede Lohnvariable mit OLS geschätzt. Der Standardfehler wird mit der konventionellen Schätzmethode oder der robusten Schätzmethode (Eicker 1963; White 1980) berechnet.¹²
3. *p-Werte*: Auf Basis der Koeffizienten, Standardfehler und Freiheitsgrade werden die t-Tests aller gewünschter Nullhypothesen berechnet und daraus die p-Werte gebildet.¹³ Anhand des p-Werts lässt sich direkt erkennen, ob eine Nullhypothese bei einem bestimmten Signifikanzniveau verworfen wird oder nicht. Konkret werden folgende Nullhypothesen getestet:
 - 3.1 $H_0: \beta = 0, H_1: \beta \neq 0$ (zweiseitig)
 - 3.2 $H_0: \beta = -0.03, H_1: \beta < -0.03$ (einseitig)
 - 3.3 $H_0: \beta = -0.05, H_1: \beta < -0.05$ (einseitig)
4. *Signifikanz*: Anhand der p-Werte und der gewünschten Signifikanzniveaus wird ermittelt, ob der geschlechtsspezifische Effekt signifikant ist oder nicht. Nebst dem konventionellen Signifikanzniveau von 5% (gegenwärtiger Wert in Logib) werden auch zwei von der Stichprobengrösse abhängige Signifikanzniveaus geprüft: $5\%/\sqrt{N/200}$ und $5\%/\sqrt{N/400}$. Die Formeln implizieren, dass das gewählte Signifikanzniveau mit zunehmender Stichprobengrösse abnimmt.¹⁴ Dies bedeutet, dass in kleineren Unternehmen ein grösserer Alpha-Fehler akzeptiert wird als in grösseren Unternehmen.
5. *Modifizierter zweistufiger Test*: Nebst den herkömmlichen Signifikanztests führen wir auch einen modifizierten zweistufigen Test durch. Im ersten Schritt wird die Nullhypothese $H_0: \beta = 0$ gegen die Alternativhypothese $H_1: \beta \neq 0$ (zweiseitig) auf einem Signifikanzniveau von 5% getestet. Sofern ein signifikantes Ergebnis vorliegt, wird als Zweites getestet, ob der numerische Schätzwert im Betrag grösser ist als 0.05, d.h., $|\hat{\beta}| > 0.05$. Der Lohngleichheitstest gilt dann als nicht bestanden, wenn die numerische Toleranzschwelle überschritten wird.

¹² Genau genommen gibt es mehrere heterokedastizitätsrobuste Schätzmethoden für Standardfehler (Hansen 2022, Kap. 4.14). Wir verwenden die Methode «HC1», welche wie folgt definiert ist: $\hat{V}(\hat{\beta}) = \left(\frac{N}{N-k}\right) (X'X)^{-1} \left(\sum_i X_i X_i' \hat{e}_i^2\right) (X'X)^{-1}$, wobei k die Anzahl Freiheitsgrade, X die Matrix aller erklärenden Variablen und $\hat{e}_i$ die OLS-Residuen darstellen. In der Simulation hat sich gezeigt, dass die anderen Methoden «HC2» bis «HC4» mit dem R-Package «sandwich» in kleinen Stichproben nicht in jedem Fall ein endliches Ergebnis liefern, während «HC1» immer ein endliches Ergebnis produziert. Zu erwähnen ist, dass «HC1» die Standardeinstellung via Option «robust» in Stata ist.

¹³ Der p-Wert ist die Wahrscheinlichkeit, dass der Koeffizient im Betrag mindestens so gross ist wie der beobachtete Wert, wenn die Nullhypothese wahr ist. Der p-Wert entspricht dem kleinsten Signifikanzniveau, bei dem die Nullhypothese gerade noch verworfen werden kann.

¹⁴ Zwei Beispiele: Bei 100 Beschäftigten sind die beiden Signifikanzniveaus 7.1% bzw. 10%; bei 1000 Beschäftigten sind es 2.2% bzw. 3.2%.

Tabelle 3: Geprüfte Parameterwerte im Algorithmus

Parameter	Geprüfte Werte
Stichprobenumfang (N)	50, 75, 100, 250, 500, 1000, 3000, 5000
Geschlechterverhältnis, Anzahl oder Anteil Frauen	Anzahl Frauen: 3, 6, 10, 20. Anteil Frauen: 5%, 10%, 25%, 50%

Für jedes Szenario in Bezug auf die Parameterwerte wird der Algorithmus viele Male wiederholt und die Ergebnisse werden jedes Mal abgespeichert. Bei Monte-Carlo-Simulationen ist es allgemein sinnvoll, die Anzahl Replikationen in Abhängigkeit der Stichprobengrösse zu wählen. Wir verwenden je nach Stichprobenumfang (N) zwischen 2000 und 8000 Replikationen.¹⁵ Unter Einbezug aller Szenarien werden total rund 15.4 Mio. Monte-Carlo-Replikationen durchgeführt.

4 Ergebnisse

In diesem Kapitel präsentieren wir die wichtigsten Ergebnisse der Monte-Carlo-Simulation; ergänzende Ergebnisse sind in Anhang A.2 enthalten, auf die wir jeweils an den entsprechenden Stellen verweisen. Zunächst gehen wir auf die Streuung der geschätzten Geschlechter-Koeffizienten und die Grösse der Standardfehler ein (Abschnitt 4.1), bevor die Hauptergebnisse zur Mächtigkeit der t-Tests diskutiert werden (Abschnitt 4.2). Im Anschluss folgen Analysen hinsichtlich einer tieferen Toleranzschwelle (Abschnitt 4.3) und einem modifizierten zweistufigen Test mit einer numerischen Toleranzschwelle (Abschnitt 4.4). Weitere Analysen widmen sich der Heteroskedastizität und robusten Standardfehler (Abschnitt 4.5) sowie einer flexibleren Spezifikation der Lohnfunktion (Abschnitt 4.6). Zusätzliche Ausführungen zu alternativen Signifikanzniveaus und multiplem Testen finden sich in Abschnitt 4.7.

4.1 Geschlechter-Koeffizient und Standardfehler

Abbildung 2 zeigt die Verteilung der geschätzten Geschlechter-Koeffizienten als Boxplots¹⁶ differenziert nach Unternehmensgrösse (50 bis 5000) und Frauenanteil (absolut und relativ). In Klammern ist jeweils der durchschnittliche Standardfehler über alle Replikationen angegeben. Die Ergebnisse beziehen sich hier auf das Szenario mit einem wahren Geschlechter-Koeffizienten von null.¹⁷ Abbildung 2 zeigt erstens, dass der Geschlechter-Koeffizient in allen Szenarien erwartungstreu geschätzt wird: Im *Durchschnitt* entspricht der geschätzte Koeffizient in etwa dem wahren Wert. Dies ist Ausdruck davon, dass das Populationsmodell keine fehlenden Kontrollvariablen beinhaltet und somit kein «Omitted Variable Bias» vorliegt. Zweitens ist zu beobachten, dass die Streuung bei einer kleinen, *absoluten* Anzahl Frauen gross ist und mit steigender Unterneh-

¹⁵ Bei $N \leq 250$ werden 8000 Replikationen durchgeführt, bei $N \in \{500, 1000\}$ sind es 4000 Replikationen und bei $N \in \{3000, 5000\}$ sind es 2000 Replikationen.

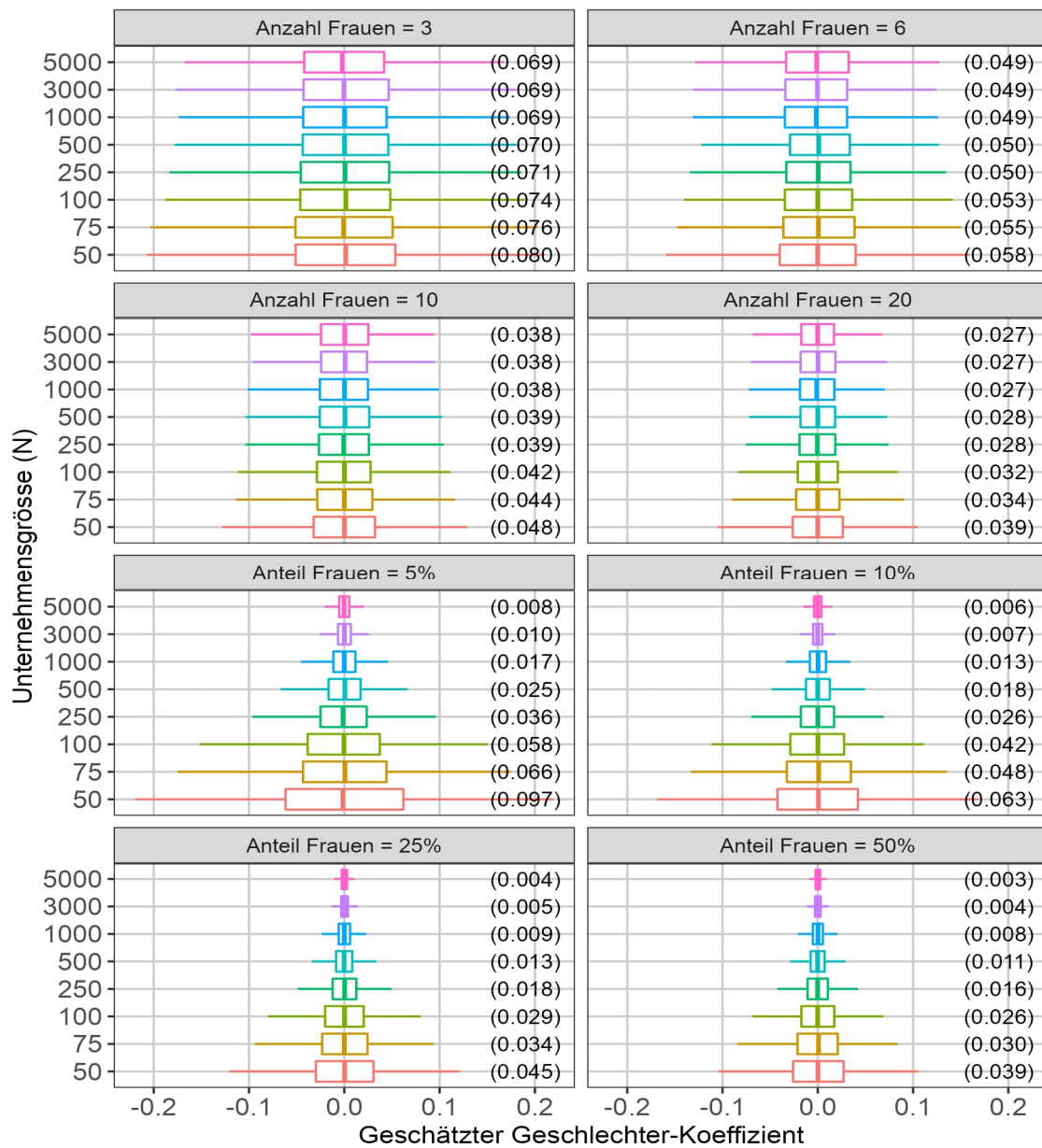
¹⁶ Die drei vertikalen Linien des Boxplot-Kastens beziehen sich auf das 1. Quartil, den Median und das 3. Quartil, während die horizontale Linie das 1,5-fache des Interquartilabstands (IQR) angibt.

¹⁷ Die Grafik sieht genau gleich aus, wenn die Spezifikation mit einem anderen wahren Geschlechter-Koeffizienten (z.B. -0.08) gewählt wird; einzig die x-Achse ist anders skaliert, siehe Abbildung 14 im Anhang.

mensgrösse kaum abnimmt. Zum Beispiel sinkt der mittlere Standardfehler bei 6 Frauen im Unternehmen nur von 0.06 (N=50) auf 0.05 (N=5000). Das heisst: Ist die absolute Anzahl Frauen sehr tief, wird die Schätzung mit zunehmender Unternehmensgrösse kaum präziser. Drittens ist in der unteren Hälfte der Abbildung zu sehen, dass die Präzision der Schätzung bei einem fixen *relativen* Frauenanteil (5% bis 50%) viel stärker von der Unternehmensgrösse abhängt. Beispielsweise reduziert sich der mittlere Standardfehler bei einem Frauenanteil von 50% von 0.039 (N=50) auf 0.003 (N=5000) Dies widerspiegelt relativ exakt die asymptotischen Eigenschaften des OLS-Schätzers.¹⁸ Im Weiteren ist die Streuung umso kleiner, je ausgeglichener das Geschlechterverhältnis im Unternehmen ist.

¹⁸ Die asymptotische Theorie besagt, dass der Standardfehler sich mit einer Rate von $\sqrt{N}$ verkleinert. Das heisst: Wenn die Stichprobengrösse um das 100-fache ansteigt, reduziert sich der Standardfehler um das 10-fache. Dieses Gesetz gilt dann, wenn die *Verteilung* der Variablen mit zunehmender Stichprobengrösse gleichbleibt. Bei einer absoluten Anzahl von Frauen hingegen, verändert sich die Verteilung des Geschlechter-Indikators mit zunehmender Stichprobengrösse.

Abbildung 2: Boxplot Geschlechter-Koeffizient und durchschnittliche Standardfehler



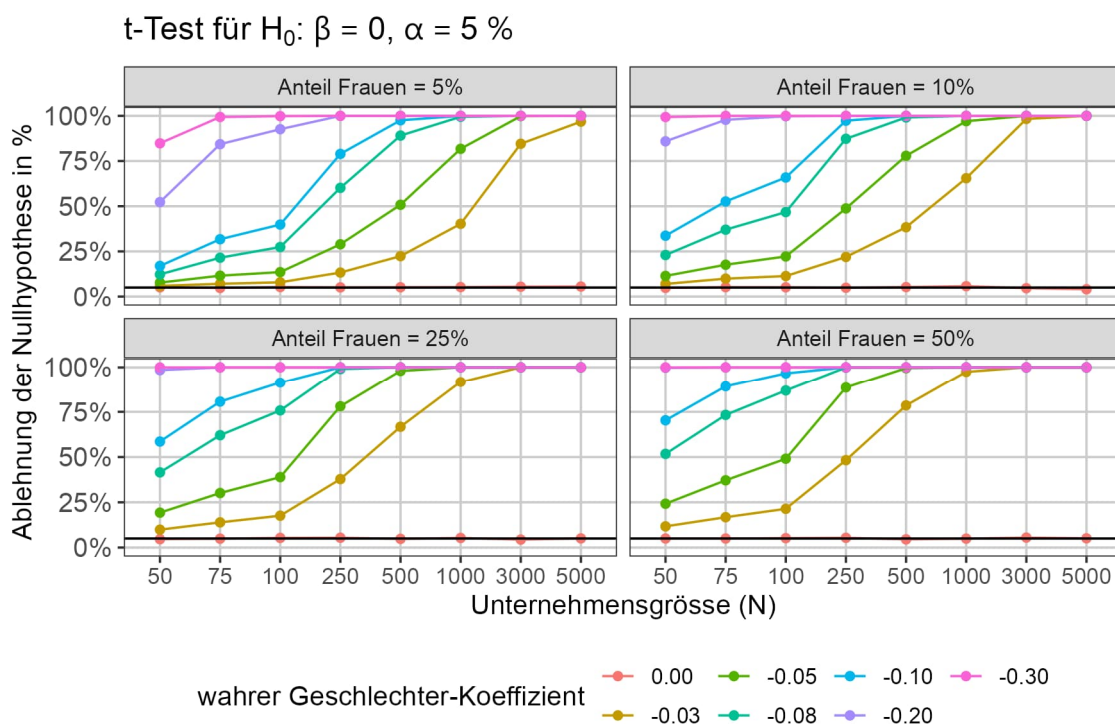
Anmerkungen: Die Grafik zeigt die Streuung der geschätzten Geschlechter-Koeffizienten als Boxplots für verschiedene Unternehmensgrössen und Frauenanteile. Der wahre Geschlechter-Koeffizient ist null. Der durchschnittliche Standardfehler ist in Klammern angegeben. Die Ergebnisse basieren auf einer Monte-Carlo-Simulation mit 2000 bis 8000 Replikationen.

4.2 Mächtigkeit nach Grösse und Frauenanteil

In diesem Abschnitt wird die Mächtigkeit *einzelner* t-Tests bei einem standardmässigen Signifikanzniveau von 5% untersucht, so wie sie zurzeit in Logib implementiert sind. Abbildung 3 zeigt die Mächtigkeit des ersten, zweiseitigen t-Tests für die Nullhypothese, dass der wahre Geschlechter-Koeffizient null ist. Die Mächtigkeit ist für verschiedene Unternehmensgrössen, relative Frauenanteile und unterschiedliche Werte des wahren Geschlechter-Koeffizienten darstellt. Folgende

Resultate sind festzuhalten: Erstens steigt die Güte des Tests mit zunehmender Unternehmensgrösse deutlich an, was an den steigenden Kurven zu erkennen ist. Beträgt der wahre Koeffizient -0.05 (d.h. -5%) und der Frauenanteil 25%, steigt die Mächtigkeit von rund 39% auf 98%, wenn sich die Beschäftigtenzahl von 100 auf 500 erhöht. Ab 500 Beschäftigten wird ein wahrer Lohngefekt von -10% nahezu immer detektiert, auch wenn der Frauenanteil sehr klein ist. Umgekehrt ist die Mächtigkeit bei kleinen Unternehmen (N < 100) vergleichsweise gering. Zweitens zeigt sich wiederum, dass der Frauenanteil eine massgebliche Rolle spielt: Je ausgeglichener das Geschlechterverhältnis, desto grösser die Mächtigkeit. Drittens, und wenig überraschend, liefert der t-Test häufiger das korrekte Ergebnis, je grösser der wahre Geschlechter-Koeffizient ist. Geringe Lohnunterschiede von -0.03 (d.h. -3%) führen in kleinen und mittleren Unternehmen nur selten zu einem signifikanten Ergebnis. Umgekehrt wird ein grosser Koeffizient von -0.2 (d.h. -20%) und grösser immer korrekt detektiert, ausser in sehr kleinen Unternehmen mit geringem Frauenanteil. Schliesslich zeigt Abbildung 3 auch die empirische Häufigkeit des Alpha-Fehlers (rote Linie). Bei einem wahren Geschlechterkoeffizienten von null ist die Nullhypothese wahr und die Häufigkeit eines Alpha-Fehlers sollte theoretisch dem Signifikanzniveau von 5% entsprechen. Wir sehen, dass diese Eigenschaft empirisch gut erfüllt ist. Weitere Ergebnisse für die Szenarien mit einer kleinen, *absoluten* Anzahl Frauen (3 bis 20) sind in Abbildung 15 in Anhang A.2.2 zu finden.

Abbildung 3: Mächtigkeit des zweiseitigen t-Tests, relativer Frauenanteil



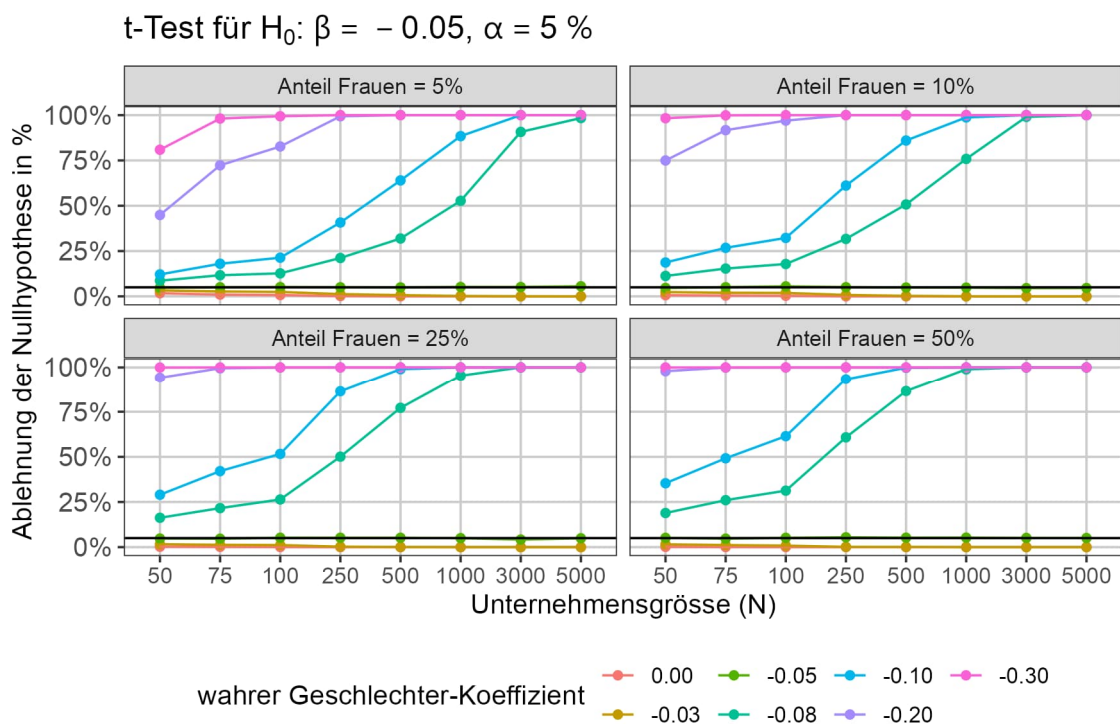
Anmerkungen: Die Grafik zeigt, wie häufig die Nullhypothese $H_0: \beta = 0$ in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten abgelehnt wird. Wenn der wahre Geschlechter-Koeffizient kleiner als null ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich null ist, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers (rote Linie).

Die Ergebnisse des zweiten, einseitigen t-Tests bezogen auf die Nullhypothese, dass der wahre Geschlechter-Koeffizient -0.05 beträgt, sind in Abbildung 4 dargestellt. Der Test weist erst bei relativ grossen Lohnunterschieden eine angemessene Mächtigkeit auf. Bei einem Frauenanteil

von 25% und einem wahren Geschlechter-Koeffizienten von -0.10 wird die Nullhypothese bei 100 Mitarbeitenden nur in rund 52% der Fälle verworfen. Erst in mittelgrossen Unternehmen (N=500) beträgt die Mächtigkeit im selben Szenario rund 100%. Beträgt der wahre Geschlechter-Koeffizient -0.20, liefert der t-Test nahezu immer das korrekte Ergebnis, wenn das Geschlechter-verhältnis nicht zu ungleich ist (d.h. bei mindestens 25% Frauen). Weitere Ergebnisse zu den absoluten Frauenanteilen finden sich in Abbildung 17 im Anhang.

Zweitens ist ersichtlich, wie häufig der t-Test je nach Szenario ein falsch-positives Ergebnis generiert (Alpha-Fehler). Bei einer Toleranzschwelle von -5% beträgt die Fehlerhäufigkeit gemäss dem gewählten Signifikanzniveau 5%, wenn der wahre Geschlechter-Koeffizient exakt dem Toleranzwert entspricht. Liegt kein geschlechterspezifischer Lohneffekt vor (d.h. der wahre Koeffizient ist null), dann beträgt die Fehlerhäufigkeit in einem kleinen Betrieb mit 50 Mitarbeitenden und 25% Frauen gerade einmal 0.2%. Mit anderen Worten führt die Toleranzschwelle dazu, dass Unternehmen ohne unerklärbare Lohnunterschiede praktisch nie ein falsch-positives Ergebnis erhalten.

Abbildung 4: Mächtigkeit des einseitigen t-Tests (-0.05), relativer Frauenanteil



Anmerkungen: Die Grafik zeigt, wie häufig die Nullhypothese $H_0: \beta = -0.05$ in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten abgelehnt wird. Wenn der wahre Geschlechter-Koeffizient kleiner als -0.05 ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich -0.05 oder kleiner ist, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers.

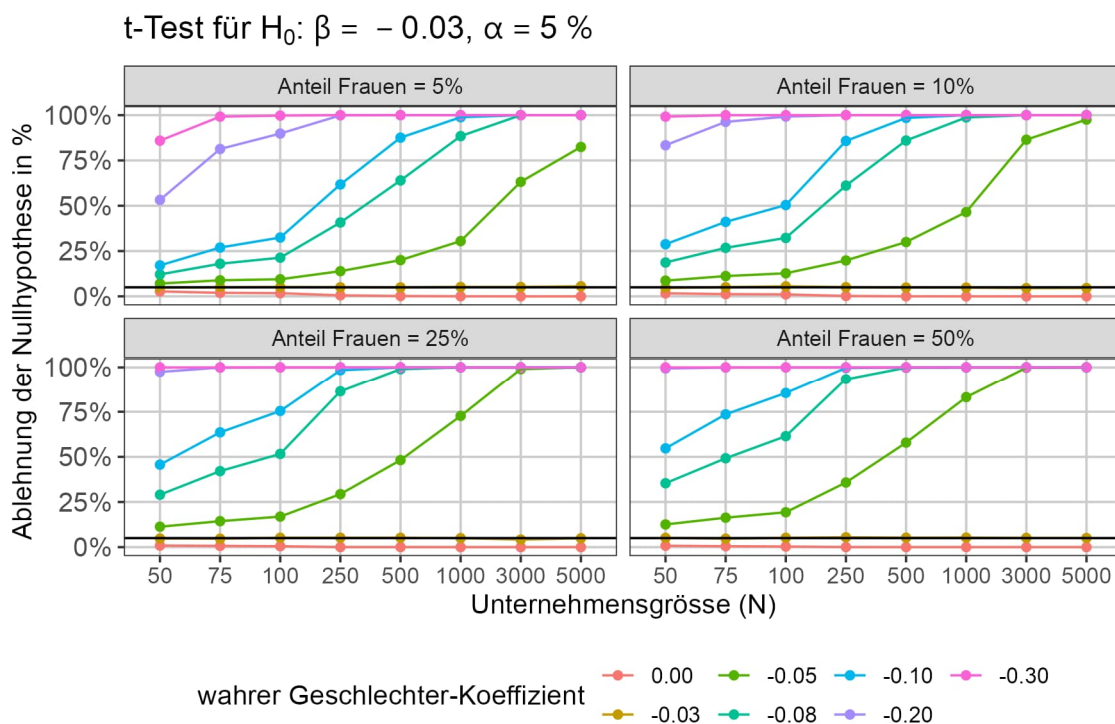
4.3 Tiefere Toleranzschwelle

Die bisherigen Ergebnisse beziehen sich auf die t-Tests, so wie sie momentan in Logib Modul 1 implementiert sind. Nun wird untersucht, wie gross die Mächtigkeit des zweiten Tests ausfällt,

wenn anstatt -5% eine Toleranzschwelle von -3% zur Anwendung kommt. Abbildung 5 zeigt im Vergleich zu Abbildung 4, dass die Mächtigkeit, wie erwartet, ansteigt: Im Szenario mit 100 Mitarbeitenden, 25% Frauen und einem Koeffizienten von -0.08 steigt die Mächtigkeit von rund 26% auf 52%, wenn die Toleranzschwelle von -5% auf -3% abgesenkt wird. Der Unterschied zwischen den beiden Toleranzschwellen fällt vor allem bei kleinen Unternehmen sowie bei moderaten Lohnunterschieden ins Gewicht, wo die Mächtigkeit deutlich unter 100% liegt. Weitere Ergebnisse mit absoluten Frauenanteilen finden sich in Abbildung 16 im Anhang.

Die Wahrscheinlichkeit, dass ein falsch-positives Ergebnis resultiert (Alpha-Fehler), wenn in Wahrheit kein geschlechterspezifischer Effekt im Unternehmen vorliegt, beträgt bei einem Unternehmen mit 50 Beschäftigten und 50% Frauenanteil 0.8%. Dieser Wert ist leicht höher als bei einer Toleranzschwelle von -5%, ist jedoch immer noch sehr tief.

Abbildung 5: Mächtigkeit des einseitigen t-Tests (-0.03), relativer Frauenanteil



Anmerkungen: Die Grafik zeigt, wie häufig die Nullhypothese $H_0: \beta = -0.03$ in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten abgelehnt wird. Wenn der wahre Geschlechter-Koeffizient kleiner als -0.03 ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich -0.03 oder kleiner ist, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers.

4.4 Modifizierter Test mit numerischer Toleranzschwelle

Aus den obigen Analysen zur Mächtigkeit des t-Tests resultiert die Erkenntnis, dass die Tests einen Geschlechtereffekt umso besser detektieren können, je grösser das Unternehmen ist. Der Grund dafür ist, dass der Geschlechtereffekt mit zunehmender Anzahl Beobachtungen präziser

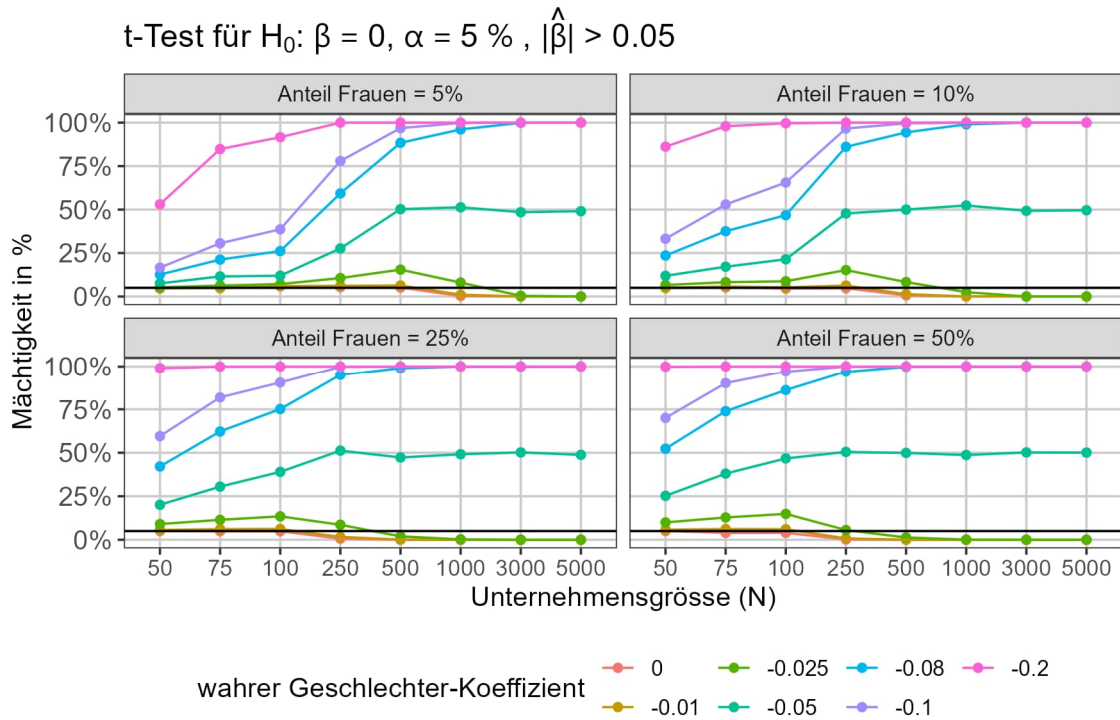
geschätzt wird und der zweite t-Test deshalb zunehmend häufiger ein signifikantes Ergebnis produziert. Eine mögliche Modifikation des Testverfahrens wäre, den ersten t-Test beizubehalten und im Falle eines signifikanten Ergebnisses lediglich zu prüfen, ob der geschätzte Koeffizient im Betrag *numerisch* grösser ist als 0.05. Somit wird anstatt des zweiten statistischen t-Tests eine numerische Toleranzschwelle verwendet. Überschreitet der geschätzte Koeffizient im Betrag 0.05, gilt der Lohngleichheitstest als nicht bestanden. Das Ziel eines solchen Tests ist es, sämtliche Abweichungen des Geschlechter-Koeffizienten von null korrekt zu detektieren. Die Häufigkeit der korrekten Ablehnung der Nullhypothese bei einem wahren Geschlechter-Koeffizienten ungleich null bezeichnen wir hier der Vergleichbarkeit halber ebenfalls als «Mächtigkeit», obwohl es sich nicht um einen reinen statistischen Signifikanztest handelt. Gleichermassen definieren wir verworfene Tests bei einem wahren Geschlechter-Koeffizienten von null als falsch-positives Ergebnis bzw. Alpha-Fehler.

Simulationsergebnisse für dieses modifizierte, zweistufige Testverfahren mit numerischer Toleranzschwelle werden nachfolgend präsentiert. Zu beachten ist, dass in diesem Abschnitt für die Simulation andere Werte des wahren Geschlechter-Koeffizienten gewählt wurden als in den übrigen Teilen der Studie.¹⁹ Abbildung 6 zeigt die Mächtigkeit des Tests für verschiedene Unternehmensgrössen, relative Frauenanteile und Grössen des wahren Geschlechter-Koeffizienten. Beispielsweise beträgt die Mächtigkeit bei 250 Beschäftigten und ausgeglichenem Geschlechterverhältnis beinahe 100%, wenn der wahre Geschlechter-Koeffizient -8% (-0.08) beträgt. Im Vergleich dazu beträgt die Mächtigkeit im zweistufigen t-Test rund 60% im selben Szenario. Allgemein lässt sich sagen, dass die Mächtigkeit des modifizierten Tests bei Geschlechter-Koeffizienten ab 0.05 (im Betrag) immer grösser ist als im zweistufigen t-Test. Der Grund liegt im unterschiedlichen Testdesign: Der modifizierte Test «schlägt schneller an», weil die Toleranzschwelle von 5% nur numerisch, nicht aber statistisch signifikant, überschritten werden muss. Der Unterschied zwischen den beiden Testverfahren ist vor allem für Unternehmen mit kleiner und mittlerer Grösse bedeutsam; bei grossen Unternehmen liefern die beiden Testverfahren in aller Regel das selbe Ergebnis.

In Abbildung 6 fällt zudem auf, dass die Mächtigkeit des modifizierten Tests bei einem wahren Geschlechter-Koeffizienten von -0.05 gegen 50% konvergiert. Dies ist theoretisch so zu erwarten. Denn: In diesem Fall streut der OLS-Schätzer des Geschlechter-Koeffizienten mit einer symmetrischen Normalverteilung um den wahren Wert von -0.05. Zudem wird ab rund 500 Beschäftigten die Nullhypothese eines Nulleffekts nahezu immer abgelehnt, sodass in diesem Fall der numerische Schätzwert allein für das Testergebnis ausschlaggebend ist. In rund 50% der Fälle ist dieser kleiner als -0.05 und in rund 50% der Fälle grösser als -0.05. Entsprechend resultiert eine Mächtigkeit von rund 50%.

¹⁹ Der Unterschied ergibt sich aus dem Umstand, dass die Analysen in Abschnitt 4.4 zu einem späteren Zeitpunkt durchgeführt wurden als die übrigen Teile der Studie. Auf Basis der Überlegungen des EBG zum modifizierten Testverfahren waren hier leicht andere Werte von Interesse.

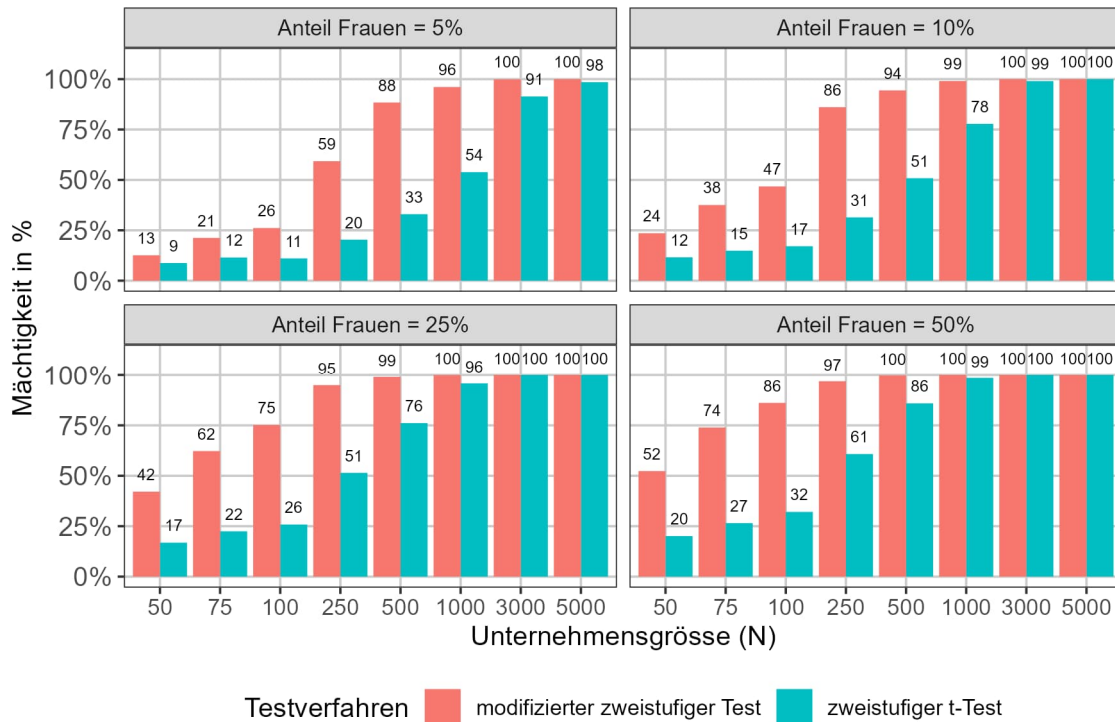
Abbildung 6: Mächtigkeit des modifizierten Tests, Schranke 0.05, relativer Frauenanteil



Anmerkungen: Die Grafik zeigt, wie häufig der modifizierte Lohngleichheitstest in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten verworfen wird. Dieser Test besteht aus zwei Schritten: einem zweiseitigen t-Test der Nullhypothese eines Nulleffekts sowie dem Überschreiten des geschätzten Koeffizienten von 0.05 im Betrag. Wenn der wahre Geschlechter-Koeffizient gleich null ist, entspricht die Ablehnungsrate der Häufigkeit eines falsch-positiven Ergebnisses.

Um den Vergleich zwischen den Ergebnissen des zweistufigen t-Tests und des modifizierten Tests mit numerischer Toleranzschwelle transparenter darzustellen, zeigt Abbildung 7 die Mächtigkeit für beide Tests für einen wahren Geschlechter-Koeffizienten von -0.08 sowie für verschiedene Unternehmensmerkmale. Der Unterschied in der Mächtigkeit der Tests ist für kleinere und mittlere Unternehmen bedeutsam. Ein Beispiel: Bei 100 Beschäftigten und ausgeglichenem Geschlechterverhältnis detektiert der modifizierte Test einen bestehenden Lohnunterschied in 86% der Fälle, der zweistufige t-Test hingegen nur in 32% der Fälle. In kleinen Unternehmen mit geringem Frauenanteil sind beide Tests schwach, weil die Nullhypothese des ersten t-Tests häufig nicht verworfen werden kann.

Abbildung 7: Mächtigkeit des modifizierten und herkömmlichen Tests, ($b=-0.08$)

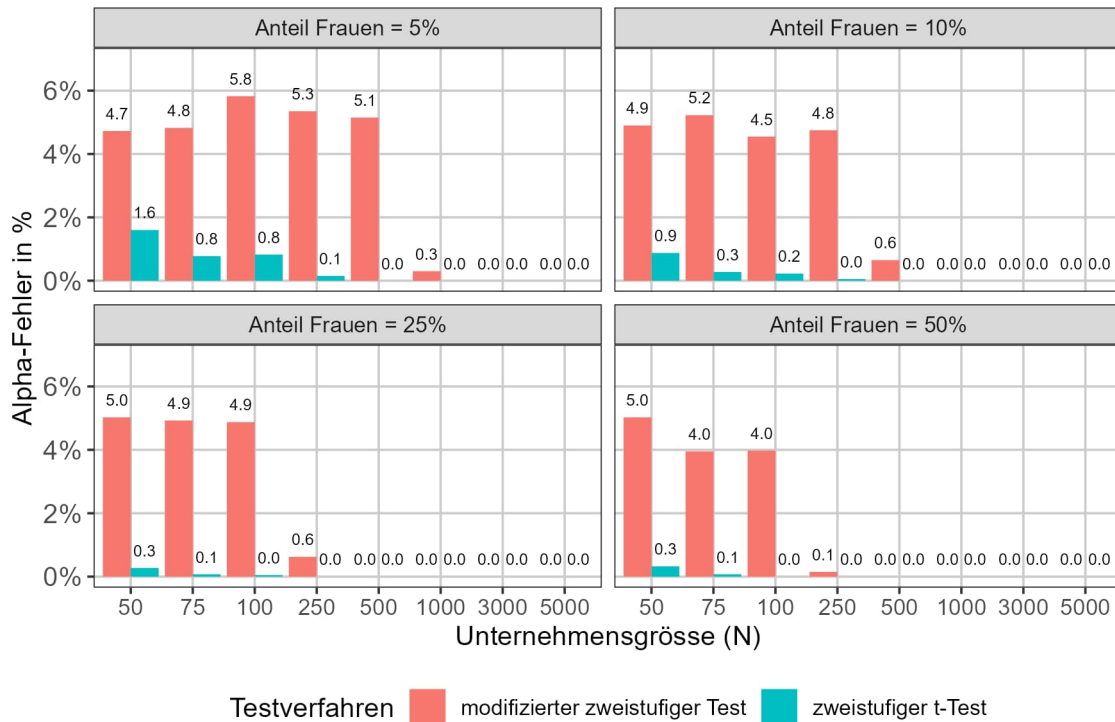


Anmerkungen: Die Grafik zeigt einen Vergleich der Mächtigkeit zwischen dem modifizierten zweistufigen Test mit numerischer Toleranzschwelle von 0.05 (im Betrag) und dem herkömmlichen zweistufigen t-Test, wobei der zweite Test die Nullhypothese $H_0: \beta = -0.05$ auf einem Signifikanzniveau von 5% testet. Die Werte beziehen sich auf einen wahren Geschlechter-Koeffizienten von -0.08.

Die grössere Mächtigkeit des modifizierten zweistufigen Tests mit numerischer Toleranzschwelle geht nicht überraschend mit einem grösseren Alpha-Fehler (falsch-positive Ergebnisse) einher. Das heisst, im Vergleich zum herkömmlichen zweistufigen t-Test wird häufiger eine Lohnungleichheit festgestellt, wenn in Wahrheit gar keine solche vorliegt. In [Abbildung 8](#) ist die Häufigkeit von Alpha-Fehlern für beide Tests differenziert nach Unternehmensmerkmalen dargestellt. Bemerkenswert ist, dass falsch-positive Ergebnisse mit dem zweistufigen t-Test selbst in kleinen Unternehmen praktisch nie auftreten. Bei 50 Beschäftigten und einem ausgeglichenen Geschlechterverhältnis beträgt die Wahrscheinlichkeit, dass bei Unternehmen mit einem wahren Geschlechter-Koeffizienten von null *fälschlicherweise* eine Lohnungleichheit detektiert wird, gerade einmal 0.3%. Bei 100 Beschäftigten fällt diese Wahrscheinlichkeit auf 0%. Der Grund liegt darin, dass der zweite t-Test ($H_0: \beta = -0.05$) mit der Toleranzschwelle einen sehr hohen «Schutz» gegen falsch-positive Ergebnisse bietet.

Demgegenüber ist die Häufigkeit von Alpha-Fehlern beim modifizierten Test mit numerischer Toleranzschwelle bedeutend grösser und liegt bei kleinen Unternehmen nahe bei 5%. Das heisst, bei jedem zwanzigsten Unternehmen mit einem wahren Geschlechter-Koeffizienten von null wird mit diesem Test *fälschlicherweise* eine Lohnungleichheit festgestellt. Der Grund dafür ist einleuchtend: Der empirische Alpha-Fehler des ersten zweiseitigen t-Tests ($H_0: \beta = 0$) beträgt aufgrund des Signifikanzniveaus rund 5%. Diese 5% falsch-positive Ergebnisse sind in aller Regel gleichzeitig auch die «extremsten» Schätzwerte des Geschlechter-Koeffizienten und sind in kleinen Unternehmen im Betrag grösser als 0.05 (vgl. dazu auch die Standardfehler in [Abbildung 2](#)).

Abbildung 8: Alpha-Fehler des modifizierten und herkömmlichen Tests, ($b=0$)



Anmerkungen: Die Grafik zeigt einen Vergleich des Alpha-Fehlers (falsch-positive Ergebnisse) zwischen dem modifizierten zweistufigen Test mit numerischer Toleranzschwelle von 0.05 (im Betrag) und dem herkömmlichen zweistufigen t-Test, wobei der zweite t-Test die Nullhypothese $H_0: \beta = -0.05$ auf einem Signifikanzniveau von 5% testet. Die Werte beziehen sich auf einen wahren Geschlechter-Koeffizienten von null.

In Abbildung 18 und Abbildung 19 in Anhang A.2.3 ist ersichtlich, wie sich die Ergebnisse verändern, wenn die numerische Schranke des modifizierten Tests auf 0.025 gesenkt wird. In diesem Fall steigt die Mächtigkeit bei einem wahren Geschlechter-Koeffizienten von -0.05 in mittelgrossen Unternehmen auf beinahe 100%. Wie zu erwarten ist, konvergiert die Mächtigkeit bei einem wahren Geschlechter-Koeffizienten von -0.025 gegen 50%. Die höhere Mächtigkeit wird wiederum mit einem grösseren Alpha-Fehler erkauft (siehe Abbildung 20): Die Häufigkeit von falsch-positiven Ergebnissen liegt hier bei mittelgrossen Unternehmen mit 250 Beschäftigten noch bei rund 5%. Bei geringem Frauenanteil (10%) liegt der Alpha-Fehler auch bei 1000 Beschäftigten noch bei rund 5%.

Zusammenfassend lässt sich sagen, dass der modifizierte zweistufige Test mit numerischer Toleranzschwelle eine Lohnungleichheit häufiger korrekt detektiert als der herkömmliche zweistufige t-Test. Diese höhere Mächtigkeit wird jedoch mit einer grösseren Häufigkeit von falsch-positiven Ergebnissen von bis zu rund 5% «erkauft», welche vor allem bei kleineren Unternehmen auftreten.

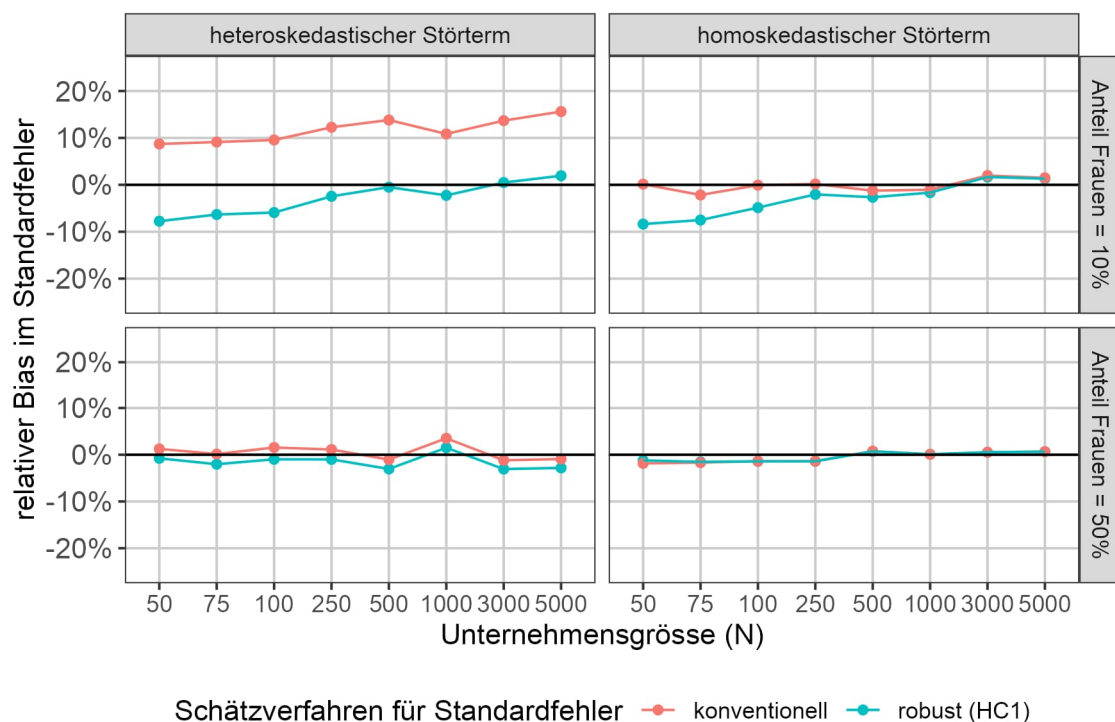
4.5 Heteroskedastizität und robuste Standardfehler

In den bisherigen Abschnitten wurden simulierte Löhne untersucht, die einen homoskedastischen Störterm aufweisen. Diese Annahme ist unrealistisch, da sich die Streuung der Löhne je

nach Geschlecht und Ausprägung der Kontrollvariablen unterscheidet. In diesem Abschnitt untersuchen wir, welchen Einfluss die *Heteroskedastizität* auf die Ergebnisse ausübt. Dabei wird auch geprüft, wie gut heteroskedastizitätsrobuste Standardfehler mögliche Verzerrungen der konventionellen Standardfehler (wie zurzeit in Logib implementiert) korrigieren können.

In Abbildung 9 ist der durchschnittliche, prozentuale Bias im Standardfehler für verschiedene Szenarien dargestellt. Im homoskedastischen Fall produzieren die konventionellen Standardfehler, wie erwartet, unverzerrte Schätzwerte. Die robusten Standardfehler sind in kleinen Stichproben ($N \leq 100$) leicht nach unten verzerrt, wenn das Geschlechterverhältnis sehr einseitig ist. Dies ist darauf zurückzuführen, dass die robusten Standardfehler nur asymptotisch (d.h., in grossen Stichproben) gültig sind und in kleinen Stichproben gegen null verzerrt sind. Im heteroskedastischen Fall sind die konventionellen Standardfehler hingegen deutlich nach oben verzerrt, wenn der Frauenanteil im Unternehmen gering ist (Kasten oben rechts). Das Problem der Heteroskedastizität scheint bei ausgeglichenem Geschlechterverhältnis deutlich weniger ins Gewicht zu fallen.

Abbildung 9: Bias im Standardfehler nach Varianzfunktion und Schätzverfahren

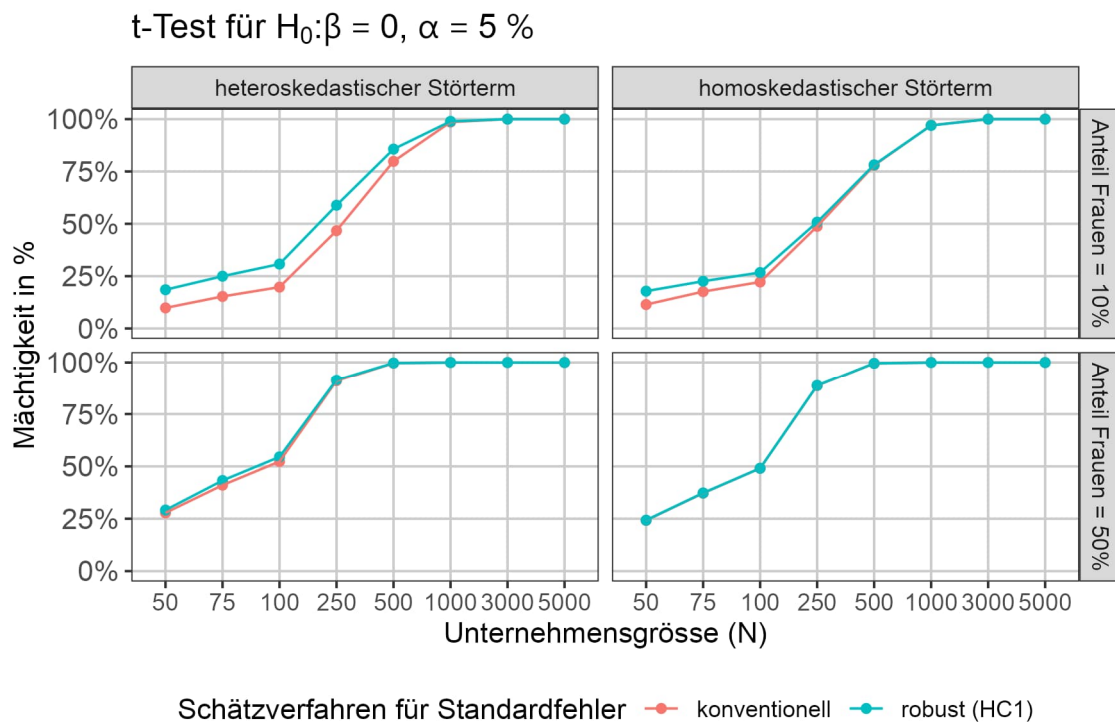


Anmerkungen: Die Grafik zeigt den prozentualen Bias im durchschnittlichen Standardfehler über alle Monte-Carlo-Replikationen. Der Bias ist für unterschiedliche Störterme (homo- und heteroskedastisch), Frauenanteile und Unternehmensgrößen dargestellt. Der wahre Geschlechter-Koeffizient beträgt -0.05. Logib verwendet zurzeit konventionelle Standardfehler.

Abbildung 10 zeigt die Mächtigkeit des zweiseitigen t-Tests ($H_0: \beta = 0$) für dieselben Szenarien bei einem wahren Geschlechter-Koeffizienten von -0.05. Die Ergebnisse zeigen, dass die Mächtigkeit des t-Tests grösser ausfällt, wenn anstatt der konventionellen Standardfehler die robusten Standardfehler verwendet werden. Dies ist insbesondere bei einer Kombination von Heteroskedastizität und geringem Frauenanteil der Fall (Kasten oben links). Bei grossen Unternehmen fällt

das Schätzverfahren für die Standardfehler nicht mehr ins Gewicht, da die Mächtigkeit ohnehin nahezu 100% beträgt. Zu beachten ist aber auch, dass robuste Standardfehler in kleinen Unternehmen mit wenigen Frauen mit einem grösseren Alpha-Fehler einhergehen (vgl. Abbildung 21 im Anhang).

Abbildung 10: Mächtigkeit nach Varianzfunktion und Schätzverfahren für Standardfehler

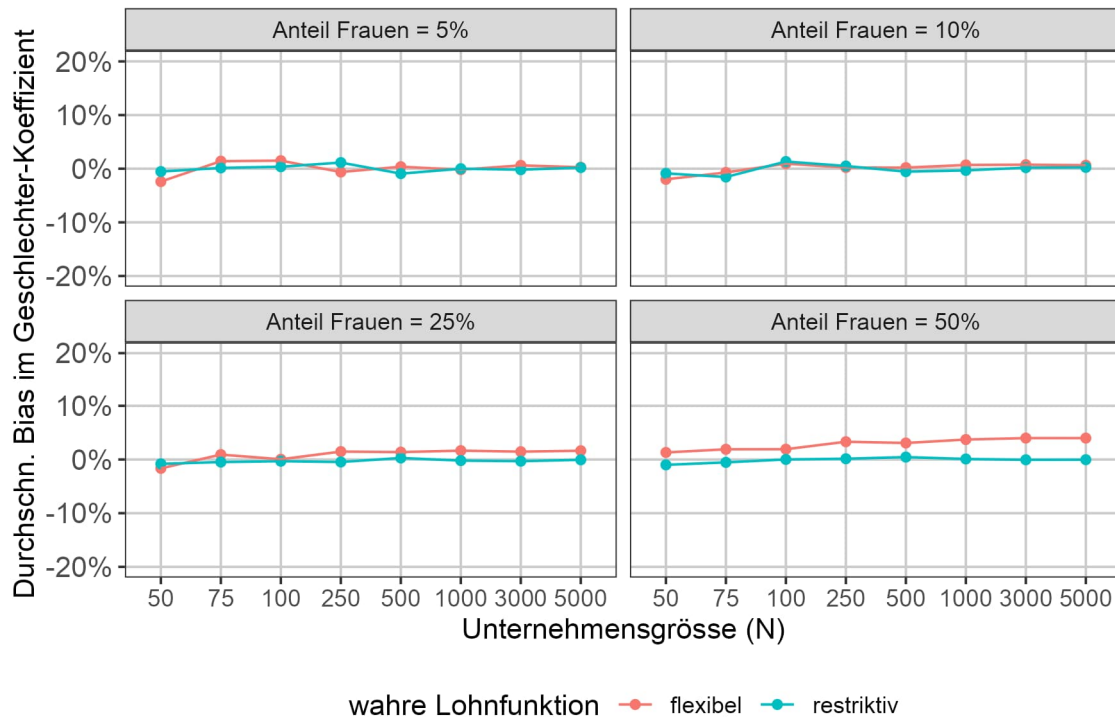


Anmerkungen: Die Grafik zeigt die Mächtigkeit des zweiseitigen t-Tests ($H_0: \beta = 0$) bei unterschiedlichen Störtermen (homo- und heteroskedastisch), Frauenanteilen, Unternehmensgrössen und Schätzverfahren für die Berechnung der Standardfehler. Der wahre Geschlechter-Koeffizient beträgt -0.05.

4.6 Spezifikation der Lohnfunktion

In den bisherigen Analysen wurden simulierte Lohnvariablen verwendet, die auf der Logib-Modellspezifikation basieren. Somit wurde die Annahme unterstellt, dass das Modell die Zusammenhänge zwischen den Löhnen, dem Geschlecht und den Kontrollvariablen korrekt abbildet. In diesem Abschnitt wird diese Annahme etwas gelockert, indem simulierte Lohnvariablen basierend auf einer leicht flexibleren Spezifikation mit geschlechterspezifischen Koeffizienten verwendet werden. Zuerst betrachten wir in Abbildung 11 den durchschnittlichen Bias im geschätzten Geschlechter-Koeffizienten für verschiedene Szenarien. Bei einem geringen Frauenanteil fällt der Unterschied zwischen den beiden Lohnfunktionen nicht ins Gewicht. Einzig bei einem ausgeglichenen Frauenanteil ist eine leichte, positive Verzerrung mit der flexibleren Lohnfunktion ersichtlich. Der prozentuale Bias liegt aber unter 5% und ist damit quantitativ relativ klein. Entsprechend ist der Einfluss auf die Mächtigkeit des t-Tests meistens nahe bei null bzw. sehr begrenzt (vgl. Abbildung 22 im Anhang).

Abbildung 11: Bias im Geschlechter-Koeffizienten nach Lohnfunktion



Anmerkungen: Die Grafik zeigt den prozentualen Bias im durchschnittlichen Koeffizienten über alle Monte-Carlo-Replikationen. Der Bias ist für unterschiedliche Lohnfunktionen (flexibel und restriktiv), Frauenanteile und Unternehmensgrößen dargestellt. Der wahre Geschlechter-Koeffizient beträgt -0.05. In der flexiblen Lohnfunktion sind alle Einflüsse der erklärenden Variablen geschlechterspezifisch; in der restriktiven Lohnfunktion (Logib) ist der geschlechtsspezifische Einfluss additiv.

Zu erwähnen ist allerdings, dass die simulierten Lohnvariablen basierend auf dem «flexibleren» Modell je nach Perspektive immer noch als eher restriktiv bezeichnet werden können. Diese basieren nach wie vor auf einem parametrischen log-linearen Modell. Weitaus flexibler und damit realitätsnäher wäre ein sogenanntes *nichtparametrisches* Modell, welches sich zum Beispiel mit modernen Machine-Learning-Methoden schätzen liesse. Ein solches Modell beinhaltet die *gleichen* Variablen; lediglich der Zusammenhang zwischen den Löhnen und den erklärenden Variablen wird realitätsnäher ermittelt. Die OLS-Schätzung des Logib-Modells könnte in einem solchen Szenario zu grösseren Verzerrungen führen. Eine entsprechende Erweiterung der Analyse geht aber über den Umfang dieser Studie hinaus.

4.7 Zusätzliche Analysen

4.7.1 Alternative Signifikanzniveaus

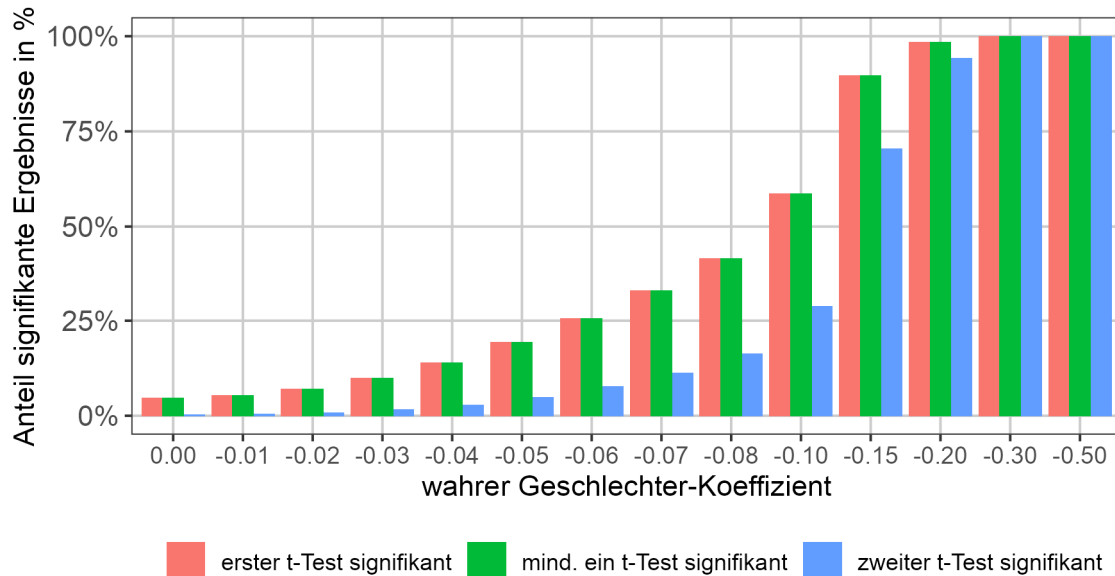
In diesem Abschnitt wird für verschiedene Szenarien untersucht, wie sich die Mächtigkeit des zweiseitigen t-Tests ändert, wenn anstelle von 5% ein von der Unternehmensgrösse abhängiges Signifikanzniveau gewählt wird. Die zwei geprüften alternativen Signifikanzniveaus sind $5\%/\sqrt{N/200}$ und $5\%/\sqrt{N/400}$. Abbildung 23 im Anhang zeigt die Mächtigkeit des Tests, wenn der wahre Geschlechter-Koeffizient -0.05 (d.h. 5%) beträgt. Die alternativen Signifikanzniveaus führen dazu, dass unerklärte Lohnunterschiede in kleinen Unternehmen tatsächlich etwas

häufiger detektiert werden. Die Steigerung der Mächtigkeit ist aber relativ moderat: Bei 50 Beschäftigten und 20 Frauen beträgt die Mächtigkeit rund 25% ($\alpha = 5\%$), 36% ($\alpha = 5\%/\sqrt{50/200} = 10\%$) bzw. 43% ($\alpha = 5\%/\sqrt{50/400} = 14.1\%$). Umgekehrt ist die Mächtigkeit der alternativen Signifikanzniveaus bei grösseren Unternehmen tiefer, wenn nur eine sehr geringe Anzahl Frauen im Unternehmen arbeiten. Insgesamt können die alternativen Signifikanzniveaus die Güte des t-Tests bei kleinen Unternehmen leicht steigern. Dabei wird aber ein häufigerer Alpha-Fehler in Kauf genommen (vgl. Abbildung 24 im Anhang). Liegt in Wahrheit kein geschlechtsspezifischer Lohneffekt vor, wird ein solcher in kleinen Unternehmen ($N=50$) in 10% bzw. 14% der Fälle fälschlicherweise detektiert.

4.7.2 Exkurs zum sequenzielles Testen

Bis anhin wurden stets die Ergebnisse *einzelner* t-Tests betrachtet. In Logib Modul 1 werden jedoch zwei sequenzielle t-Tests durchgeführt. Es stellt sich daher die Frage, ob das mehrstufige Testen ein Problem darstellt. Aus der statistischen Theorie ist bekannt, dass multiples Testen zu einer höheren Wahrscheinlichkeit für mindestens ein falsch-positives Ergebnis führt (siehe z.B. Hansen 2022, Kapitel 9.2). Zwischen dem sequenziellen Test in Logib und dem Multiple-Testing-Problem gibt es jedoch zwei entscheidende Unterschiede. Zum einen betrifft dieses Problem das Testen mehrerer, unterschiedlicher Parameter oder Koeffizienten, während im vorliegenden Fall zwei Nullhypothesen mit Bezug auf *denselben* Koeffizienten getestet werden. Zum anderen ist der zweite Test in Logib strikt strenger als der erste Test. Das heisst: Ein signifikantes Ergebnis im zweiten Test impliziert immer auch ein signifikantes Ergebnis im ersten Test. Die Wahrscheinlichkeit, dass mindestens einer der beiden Tests die Nullhypothese ablehnt, entspricht somit der Wahrscheinlichkeit, dass der erste Test die Nullhypothese ablehnt. Dieser Umstand ist in Abbildung 12 klar ersichtlich. Dem gegenüber ist bei einem Multiple-Testing-Problem die Wahrscheinlichkeit für mindestens ein signifikantes Ergebnis über alle Tests hinweg grösser als in den einzelnen Tests. Aus den beiden genannten Gründen handelt es sich im vorliegenden Fall nicht um ein Multiple-Testing-Problem im eigentlichen Sinn.

Abbildung 12: Signifikante Ergebnisse bei mehrfachem Testen



Anmerkungen: Die Grafik zeigt für verschiedene Geschlechter-Koeffizienten, wie häufig die t-Tests (einzeln oder gemeinsam betrachtet) signifikante Ergebnisse liefern. Das betrachtete Szenario ist ein Unternehmen mit 50 Beschäftigten und 25% Frauen.

In der Literatur werden verschiedene Korrekturverfahren vorgeschlagen, um ein Multiple-Testing-Problem zu adressieren. Zu nennen ist zum Beispiel die Korrektur von Benjamini & Hochberg (1995): Diese begrenzt die Fehlerhäufigkeit unter allen abgelehnten Nullhypothesen (engl. *false discovery rate*). Da es sich im vorliegenden Fall aber nicht um ein klassisches Multiple-Testing-Problem handelt, sind solche Korrekturverfahren in Logib weder notwendig noch zielführend.

Wichtig ist allerdings, dass der zweite t-Test unabhängig vom ersten t-Test *immer* durchgeführt wird, wenn es darum geht, seine statistischen Eigenschaften zu untersuchen. Wird auf ein signifikantes Ergebnis des ersten t-Tests *konditioniert*, sind die Eigenschaften des zweiten t-Tests verzerrt. Es resultiert ein häufigerer Alpha-Fehler in kleinen Unternehmen und/oder in Unternehmen mit geringem Frauenanteil (vgl. Abbildung 25 im Anhang). Dies stellt im vorliegenden Fall kein methodisches Problem dar, kann aber eine interessante Erkenntnis darstellen.

5 Fazit

Im Rahmen dieser Studie wurden die statistischen Eigenschaften der t-Tests in Logib Modul 1 mittels einer Monte-Carlo-Simulation untersucht. Im Zentrum stand die Frage, wie häufig die t-Tests einen geschlechterspezifischen Lohneffekt detektieren, wenn tatsächlich ein solcher existiert. Diese Eigenschaft wird als Mächtigkeit oder Trennschärfe (engl. *power*) bezeichnet. Durch die empirische Kalibrierung der Simulation anhand der LSE wurde versucht, eine möglichst grosse Realitätsnähe zu erzeugen, sodass die Ergebnisse eine möglichst hohe Belastbarkeit aufweisen.

Die Resultate der Monte-Carlo-Simulation zeigen, dass die Mächtigkeit des t-Tests massgeblich von drei Faktoren abhängt. Der erste und naheliegende ist die Unternehmensgrösse: Der erste t-Test erkennt einen Lohneffekt von -0.05 (-5%) bei Unternehmen mit 50 Beschäftigten und 50% Frauen nur in 25% der Fälle. Die Mächtigkeit steigt jedoch auf 50% bei 100 Beschäftigten und schliesslich auf 100% bei 500 Beschäftigten. In mittleren und grösseren Unternehmen werden unerklärbare Lohnunterschiede zwischen den Geschlechtern mit Logib somit weitaus häufiger aufgedeckt als in kleinen Unternehmen. Zweitens spielt auch das Geschlechterverhältnis eine entscheidende Rolle. Sind in einem kleinen Unternehmen mit 100 Mitarbeitenden 50% Frauen beschäftigt, erkennt der erste t-Test einen Lohneffekt von -0.1 (-10%) nahezu immer (97%). Die Mächtigkeit sinkt hingegen auf rund 40%, wenn nur 5% Frauen im Unternehmen arbeiten. Ein sehr einseitiges Geschlechterverhältnis geht somit mit einer deutlich tieferen Güte des Tests einher. Drittens, und wenig überraschend, ist entscheidend, wie gross das Ausmass der Lohnunterschiede im Unternehmen ist. Bei sehr grossen Geschlechter-Koeffizienten ab -0.2 (-20%) liefert der zweite t-Test bei einer Toleranzschwelle von -5% praktisch immer das korrekte Ergebnis, wenn das Geschlechterverhältnis nicht allzu ungleich ist (mindestens 25% Frauen). Bei Geschlechter-Koeffizienten von -0.1 (10%) ist das hingegen nicht mehr der Fall: In kleineren und mittleren Unternehmen mit weniger als 250 Mitarbeitenden erkennt der Test den geschlechter-spezifischen Lohneffekt in vielen Fällen nicht.

Die Monte-Carlo-Simulation liefert auch Aufschluss darüber, in welchen Situationen der t-Test ein falsch-positives Ergebnis generiert (Alpha-Fehler). Bei einer Toleranzschwelle von -5% beträgt die Fehlerhäufigkeit gemäss dem gewählten Signifikanzniveau 5%, wenn der wahre Geschlechter-Koeffizient exakt dem Toleranzwert entspricht. Liegt kein geschlechterspezifischer Lohneffekt vor (d.h. der Koeffizient ist null), dann beträgt die Fehlerhäufigkeit in einem kleinen Betrieb mit 50 Mitarbeitenden und 25% Frauen gerade einmal 0.2%. Mit anderen Worten führt die Toleranzschwelle dazu, dass Unternehmen ohne unerklärbare Lohnunterschiede praktisch nie ein falsch-positives Ergebnis erhalten.

Weitere Resultate zeigen unter anderem, wie sich die Performance der Tests bei bestimmten Modifikationen verändern würde. Wird zum Beispiel die Toleranzschwelle des zweiten Tests auf -3% reduziert, steigt die Wahrscheinlichkeit, dass geschlechterspezifische Lohneffekte korrekt detektiert werden, während falsch-positive Ergebnisse bei einem wahren Lohneffekt von null nur leicht zunehmen. Dies gilt vor allem für kleinere und mittlere Unternehmen. Zweitens wurde ein modifizierter, zweistufiger Test mit einer *numerischen* Toleranzschwelle von 0.05 geprüft. Das heisst, der Lohnungleichheitstest gilt als nicht bestanden, wenn der Geschlechter-Koeffizient signifikant von null verschieden ist und der Schätzwert im Betrag grösser ist als 0.05. Der modifizierte Test detektiert Lohnungleichheit vor allem bei kleinen und mittleren Unternehmen deutlich häufiger als der zweistufige t-Test. Die grössere Mächtigkeit des Tests geht jedoch mit einer markant höheren Rate an falsch-positiven Ergebnissen einher. Drittens kommt es in bestimmten Szenarien zu einem Performanceverlust aufgrund von Heteroskedastizität, weil die konventionellen Standardfehler, die in Logib verwendet werden, nach oben verzerrt sind. Durch die Verwendung robuster Standardfehler könnte die Mächtigkeit etwas gesteigert werden. Viertens könnte eine Erhöhung des Signifikanzniveaus bei kleinen Unternehmen (z.B. von 5% auf 10% bei 50 Mitarbeitenden) die Performance des ersten t-Tests – unter Inkaufnahme eines häufigeren Alpha-Fehlers – moderat steigern. So würde die Mächtigkeit des ersten t-Tests beispielsweise von etwa 24% auf 36% zunehmen (Szenario: wahrer Lohneffekt von -5% und Frauenanteil 50%).

Zusammengenommen liefert die Monte-Carlo-Simulation wertvolle Erkenntnisse, die in die Beurteilung von Logib Modul 1 sowie dessen potenzielle Weiterentwicklung einfließen können. Zum einen können die Ergebnisse als Informationsgrundlage hinzugezogen werden zur Frage, ab welcher Unternehmensgrösse bzw. ab welchem Mindest-Frauenanteil das Modul 1 empfohlen wird. Zum anderen zeigen die Ergebnisse auf, wie sich allfällige Modifikationen in Bezug auf die Wahl der Toleranzschwelle, das gewählte Signifikanzniveau oder die Schätzmethode für die Standardfehler auf die Performance der t-Tests auswirken würden.

Abschliessend ist auch auf Limitationen dieser Simulationsstudie hinzuweisen. Zum einen kamen bei der Erzeugung künstlicher Lohnvariablen nur lineare Modelle zur Anwendung. Auf die Verwendung von nichtparametrischen Methoden, die die Zusammenhänge zwischen Löhnen und erklärenden Variablen flexibler abbilden können, wurde verzichtet. Zum anderen besteht die wichtigste Limitation wohl aber darin, dass im Rahmen der Simulation die Annahme getroffen wurde, dass das Lohnmodell alle relevanten Kontrollvariablen enthält. Die Implikationen eines potenziellen «Omitted Variable Bias» waren somit nicht Gegenstand dieser Studie.

6 Quellenangaben

- Benjamini, Yoav, und Yosef Hochberg. 1995. „Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing“. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.
- Blinder, Alan S. 1973. „Wage Discrimination: Reduced Form and Structural Estimates“. *The Journal of Human Resources* 8(4):436–55. doi: 10.2307/144855.
- EBG. 2021. *Standard-Analysemodell des Bundes: Methodenbeschrieb zu Logib Modul 1 und 2*. Bern: Eidgenössisches Büro für die Gleichstellung von Frau und Mann. https://www.ebg.admin.ch/dam/ebg/de/dokumente/lohngleichheit/logib/methodenbeschrieb_logib_m1_m2.pdf.download.pdf/Methodenbeschrieb%20Logib%20Modul%201%20und%202_DE.pdf.
- Eicker, Friedhelm. 1963. „Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions“. *The Annals of Mathematical Statistics* 34(2):447–56. doi: 10.1214/aoms/1177704156.
- Hansen, Bruce. 2022. *Econometrics*. Princeton University Press.
- Kaiser, Boris, und Thomas Möhr. 2021. *Analyse der Lohnunterschiede zwischen Frauen und Männern anhand der Schweizerischen Lohnstrukturerhebung (LSE) 2018*. BSS Volkswirtschaftliche Beratung. Studie im Auftrag des Bundesamts für Statistik.
- Oaxaca, Ronald. 1973. „Male-Female Wage Differentials in Urban Labor Markets“. *International Economic Review* 14(3):693–709. doi: 10.2307/2525981.
- White, Halbert. 1980. „A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity“. *Econometrica* 48(4):817–38. doi: 10.2307/1912934.
- Wooldridge, Jeffrey M. 2015. *Introductory Econometrics: A Modern Approach*. Cengage Learning.

A. Anhang

A.1 Populationsdaten

A.1.1 Filterung

Tabelle 4 zeigt, wie viele Beobachtungen aus dem LSE-Datensatz aufgrund der verschiedenen Filter ausgeschlossen werden. Durch die gemeinsame Anwendung aller Filter (A bis E) werden 25% der Beobachtungen ausgeschlossen.

Tabelle 4: Ausschlüsse aus den LSE-Daten nach Filter

Filter	Ausschlüsse, N	Ausschlüsse, Anteil
filter_Alle	500'415	25.3%
filter_A	340'324	17.2%
filter_B	103'913	5.3%
filter_E	95'075	4.8%
filter_C	4'393	0.2%
filter_D	248	0.0%

Anmerkungen: Der LSE-Datensatz beinhaltet 1'978'116 Beobachtungen. Quelle: LSE 2018, BFS.

A.1.2 Verteilung der erklärenden Variablen

In Tabelle 5 sind die relativen Häufigkeiten der einzelnen Ausprägungen der Kontrollvariablen differenziert für Frauen und Männer ersichtlich. Grundlage ist der aus der LSE konstruierte Populationsdatensatz. Die Tabelle zeigt zudem die Durchschnitte der logarithmierten, standardisierten Löhne nach Geschlecht und Merkmalsausprägung.

Tabelle 5: Verteilung der Kontrollvariablen nach Geschlecht

Variable	Ausprägung	log(Lohn), Männer	log(Lohn), Frauen	Anteil, Männer	Anteil, Frauen
years_of_training	7	8.59	8.43	6.2%	6.2%
	11	8.63	8.48	3.3%	2.2%
	12	8.76	8.64	25.7%	21.5%
	13	8.87	8.75	1.9%	1.9%
	14	9.07	8.90	7.0%	5.0%

	15	9.15	8.96	4.3%	3.8%
	17	9.30	9.07	6.5%	4.5%
years_of_service	0	8.69	8.59	8.7%	7.4%
	1	8.75	8.63	5.4%	5.1%
	2	8.81	8.66	4.4%	4.1%
	3	8.81	8.68	3.9%	3.4%
	4	8.84	8.69	3.5%	2.9%
	5	8.85	8.71	2.8%	2.5%
	6	8.88	8.73	2.4%	2.2%
	7	8.90	8.75	2.3%	1.9%
	8	8.96	8.83	2.0%	1.9%
	9	8.94	8.77	1.6%	1.3%
	10	8.95	8.77	1.8%	1.4%
	11	8.97	8.75	1.6%	1.2%
	12	8.96	8.77	1.2%	0.9%
	13	8.99	8.80	1.0%	0.8%
	14	8.98	8.79	0.9%	0.8%
	15	9.00	8.80	0.8%	0.7%
	16	9.01	8.81	1.0%	0.8%
	17	9.04	8.83	1.3%	0.9%
	18	9.02	8.78	0.9%	0.7%
	19	9.03	8.80	0.6%	0.5%
	20	9.04	8.81	0.6%	0.4%
	21	9.04	8.82	0.5%	0.3%
	22	9.06	8.82	0.4%	0.3%
	23	9.03	8.77	0.4%	0.3%
	24	9.07	8.80	0.3%	0.2%
	25	9.07	8.81	0.3%	0.2%
	26	9.07	8.82	0.4%	0.2%
	27	9.03	8.83	0.4%	0.3%
	28	9.04	8.84	0.5%	0.3%
	29	9.03	8.83	0.4%	0.2%

	30	9.08	8.87	0.4%	0.2%
	31	9.04	8.83	0.3%	0.1%
	32	9.04	8.85	0.3%	0.1%
	33	9.06	8.84	0.2%	0.1%
	34	9.04	8.83	0.2%	0.1%
	35	9.07	8.87	0.2%	0.1%
	36	9.03	8.86	0.2%	0.1%
	37	9.03	8.87	0.2%	0.1%
	38	9.02	8.85	0.1%	0.1%
	39	9.05	8.89	0.1%	0.0%
	40	9.05	8.91	0.1%	0.0%
	41	9.02	8.83	0.1%	0.0%
	42	8.98	8.65	0.0%	0.0%
	43	8.99	8.81	0.1%	0.0%
	44	8.97	8.82	0.0%	0.0%
	45	8.96	8.63	0.0%	0.0%
	46	8.95	8.61	0.0%	0.0%
	47	8.95	8.75	0.0%	0.0%
	48	9.16	8.15	0.0%	0.0%
	49	9.06	8.80	0.0%	0.0%
	50	8.97	8.22	0.0%	0.0%
	51	9.30	8.45	0.0%	0.0%
	52	9.36	8.85	0.0%	0.0%
	53	9.04	8.96	0.0%	0.0%
	54	9.54	9.23	0.0%	0.0%
	55	9.01	7.93	0.0%	0.0%
	56	8.82		0.0%	
	57	8.59	9.78	0.0%	0.0%
	58	8.76	8.60	0.0%	0.0%
years_of_earning	0	8.46	8.40	0.1%	0.1%
	1	8.45	8.42	0.2%	0.3%
	2	8.48	8.45	0.4%	0.5%

	3	8.51	8.49	0.6%	0.7%
	4	8.56	8.53	0.7%	0.8%
	5	8.61	8.56	0.9%	0.9%
	6	8.62	8.59	1.1%	1.0%
	7	8.67	8.62	1.2%	1.1%
	8	8.69	8.65	1.2%	1.1%
	9	8.71	8.66	1.3%	1.2%
	10	8.74	8.67	1.4%	1.2%
	11	8.77	8.69	1.4%	1.2%
	12	8.79	8.71	1.4%	1.1%
	13	8.80	8.72	1.4%	1.1%
	14	8.82	8.73	1.5%	1.1%
	15	8.84	8.74	1.4%	1.1%
	16	8.85	8.76	1.4%	1.0%
	17	8.87	8.75	1.4%	1.1%
	18	8.87	8.75	1.4%	1.1%
	19	8.89	8.75	1.5%	1.0%
	20	8.91	8.77	1.4%	1.0%
	21	8.91	8.77	1.4%	1.0%
	22	8.93	8.78	1.3%	1.0%
	23	8.93	8.76	1.3%	1.0%
	24	8.95	8.76	1.3%	1.0%
	25	8.96	8.77	1.3%	1.0%
	26	8.96	8.75	1.3%	1.0%
	27	8.97	8.77	1.3%	1.1%
	28	8.97	8.77	1.4%	1.1%
	29	8.97	8.76	1.4%	1.1%
	30	8.99	8.77	1.4%	1.2%
	31	8.98	8.76	1.4%	1.2%
	32	8.98	8.74	1.4%	1.2%
	33	8.98	8.75	1.4%	1.2%
	34	8.98	8.74	1.4%	1.2%

	35	8.98	8.73	1.3%	1.2%
	36	8.97	8.74	1.4%	1.1%
	37	8.96	8.72	1.3%	1.1%
	38	8.97	8.72	1.3%	1.0%
	39	8.96	8.72	1.2%	1.0%
	40	8.94	8.72	1.1%	0.9%
	41	8.94	8.69	1.0%	0.8%
	42	8.92	8.68	0.9%	0.7%
	43	8.89	8.68	0.8%	0.6%
	44	8.89	8.65	0.6%	0.5%
	45	8.84	8.64	0.5%	0.4%
	46	8.83	8.59	0.4%	0.2%
	47	8.77	8.58	0.2%	0.2%
	48	8.78	8.58	0.1%	0.1%
	49	8.71	8.53	0.1%	0.1%
	50	8.75	8.57	0.1%	0.1%
	51	8.77	8.57	0.1%	0.0%
	52	8.86	8.53	0.0%	0.0%
	53	8.79	8.54	0.0%	0.0%
	54	8.78	8.54	0.0%	0.0%
	55	8.79	8.64	0.0%	0.0%
	56	8.62	8.71	0.0%	0.0%
	57	8.67	8.59	0.0%	0.0%
	58	8.43	8.50	0.0%	0.0%
	59	8.79	8.30	0.0%	0.0%
	60	8.76	8.40	0.0%	0.0%
	61	8.62	8.57	0.0%	0.0%
	62	8.44	8.67	0.0%	0.0%
	63	8.40	8.39	0.0%	0.0%
	64	8.76	8.51	0.0%	0.0%
	65	8.25	8.42	0.0%	0.0%
	66	8.04	8.22	0.0%	0.0%

	67	8.23	8.35	0.0%	0.0%
skill_level	1	9.17	8.99	16.7%	10.6%
	2	8.93	8.79	10.9%	11.7%
	3	8.67	8.55	22.6%	18.8%
	4	8.64	8.45	4.7%	4.1%
professional_position	1	9.30	9.00	4.8%	1.5%
	2	9.28	9.08	4.0%	2.5%
	3	9.07	8.91	4.8%	3.1%
	4	8.91	8.79	4.2%	2.8%
	5	8.74	8.64	37.0%	35.4%

Anmerkungen: Die Anteile beziehen sich auf die relative Häufigkeit einer Merkmalsausprägung im Populationsdatensatz nach Geschlecht. Die log-Löhne sind Durchschnitte der logarithmierten, standardisierten Löhne nach Merkmalsausprägung und Geschlecht. Der Populationsdatensatz wurde auf Basis der LSE 2018 gebildet.

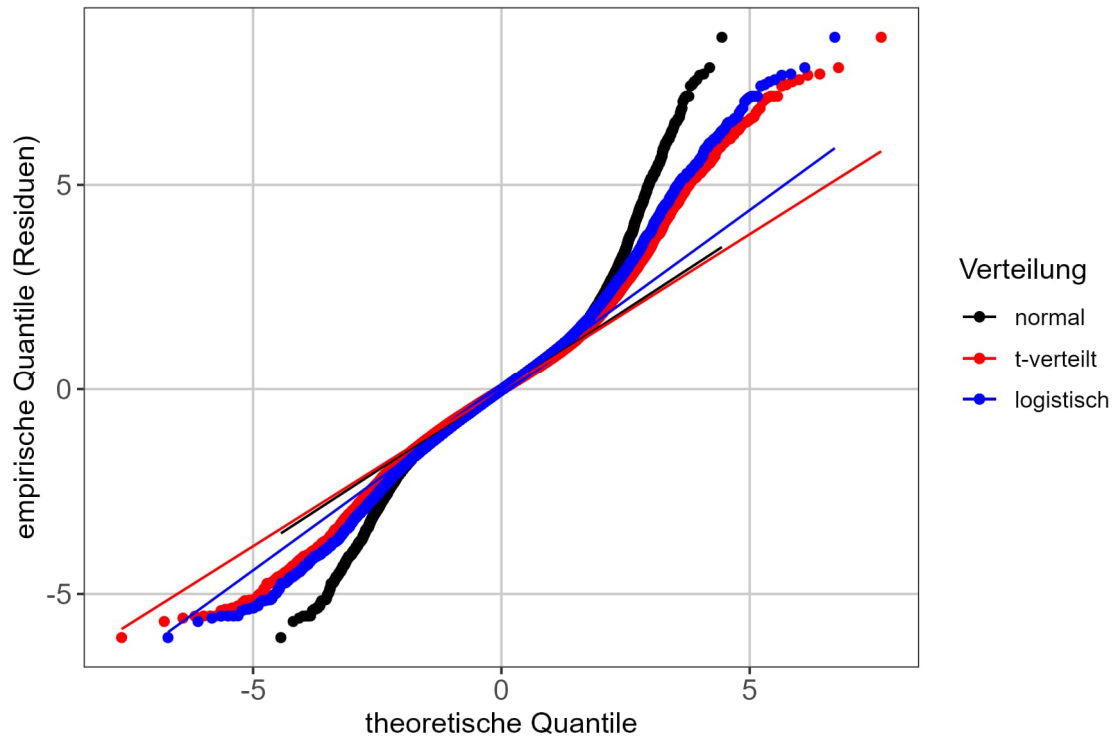
A.1.3 Störterm

Wir evaluieren drei mögliche Verteilungen auf ihre Eignung hin: die Normalverteilung, die Student-t-Verteilung und die logistische Verteilung. Das Vorgehen ist mehrstufig:

1. Mit dem Populationsdatensatz wird zuerst das Logib-Modell (Populationsregression) geschätzt und daraus die standardisierten OLS-Residuen gebildet, die den Mittelwert null und die Standardabweichung eins haben.
2. Die Parameter der drei genannten Verteilungen werden anhand der OLS-Residuen geschätzt. Bei der Normalverteilung müssen im Grunde keine Parameter geschätzt werden, da die standardisierten OLS-Residuen Mittelwert 0 und Standardabweichung 1 haben. Bei der Student-t-Verteilung werden die Freiheitsgrade mit Maximum-Likelihood geschätzt. Bei der logistischen Verteilung ergeben sich der «Ortsparameter» und der «Skalenparameter» direkt aus den Momenten der Verteilung.
3. Schliesslich werden die Informationskriterien AIC und BIC berechnet, die angeben, wie gut die theoretische und empirische Verteilung jeweils übereinstimmen.

Im Ergebnis zeigt sich, dass die *logistische Verteilung* den OLS-Residuen am nächsten kommt, da AIC und BIC kleiner ausfallen als bei den anderen beiden Kandidaten. Abbildung 13 stellt eine Quantil-Quantil-Grafik dar, wobei die empirischen Quantile der OLS-Residuen auf der y-Achse und die theoretische Quantile dreier Wahrscheinlichkeitsverteilungen auf der x-Achse abgetragen sind. Je näher sich die Punkte bei der entsprechenden Linie befinden, desto besser werden die OLS-Residuen durch die jeweilige Wahrscheinlichkeitsverteilung beschrieben. Die Grafik zeigt, dass die Normalverteilung die OLS-Residuen weniger gut approximiert als die logistische Verteilung und die t-Verteilung.

Abbildung 13: Quantil-Quantil-Grafik der OLS-Residuen mit drei Verteilungen



Anmerkungen: Die Grafik zeigt die Quantile der theoretischen Wahrscheinlichkeitsverteilungen auf der x-Achse und die empirischen Quantile der standardisierten OLS-Residuen aus der Populationsregression auf der y-Achse.

In Tabelle 6 sind die numerischen Ergebnisse der Approximation der OLS-Residuen durch die drei theoretischen Wahrscheinlichkeitsverteilungen (Normalverteilung, Student-t-Verteilung, logistische Verteilung) dargestellt. Je *kleiner* die Informationskriterien (AIC, BIC), desto besser der Fit. Die Zahlen zeigen, dass die logistische Verteilung die OLS-Residuen am besten beschreibt.

Tabelle 6: Informationskriterien für den Fit der drei Verteilungen

Verteilung	AIC	BIC
t-verteilt	278'032	278'042
logistisch	273'081	273'100
normal	283'791	283'810

A.1.4 Populationsregression

Tabelle 7: Populationsregression mit echten und simulierten Lohndaten, vollständig

	real	simuliert (b=0)	simuliert (b=-0.05)	simuliert (b=-0.10)
	(1)	(2)	(3)	(4)

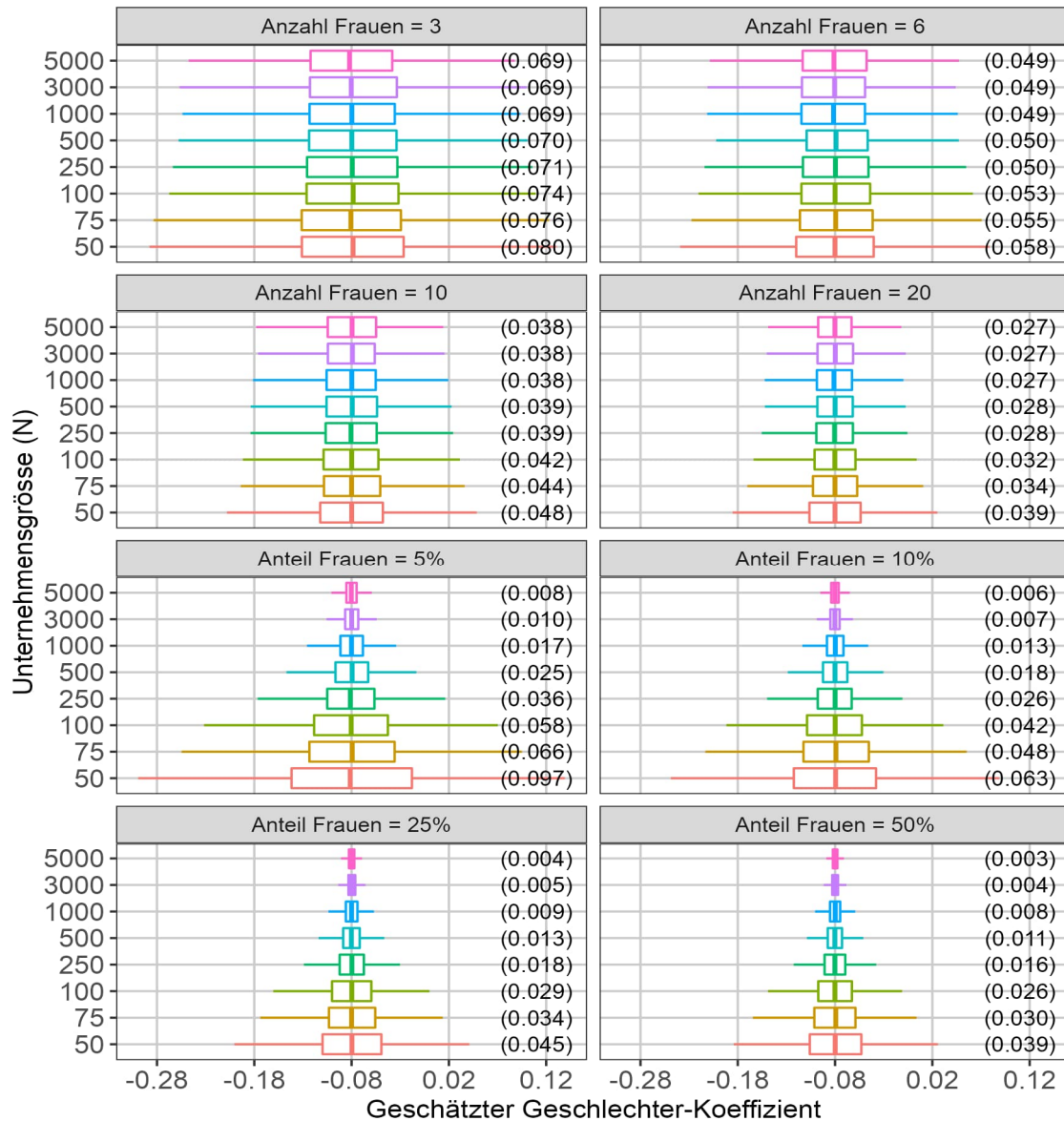
years_of_training	0.049***	0.049***	0.049***	0.049***
	(0.0001)	(0.00003)	(0.00003)	(0.00003)
years_of_service	0.005***	0.005***	0.005***	0.005***
	(0.00002)	(0.00001)	(0.00001)	(0.00001)
years_of_earning	0.023***	0.023***	0.023***	0.023***
	(0.0001)	(0.00002)	(0.00002)	(0.00002)
years_of_earning2	-0.0003***	-0.0003***	-0.0003***	-0.0003***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
skill_level2	-0.100***	-0.100***	-0.100***	-0.100***
	(0.0005)	(0.0002)	(0.0002)	(0.0002)
skill_level3	-0.259***	-0.259***	-0.259***	-0.259***
	(0.0005)	(0.0002)	(0.0002)	(0.0002)
skill_level4	-0.270***	-0.270***	-0.270***	-0.270***
	(0.001)	(0.0003)	(0.0003)	(0.0003)
professional_position2	0.018***	0.018***	0.018***	0.018***
	(0.001)	(0.0004)	(0.0004)	(0.0004)
professional_position3	-0.099***	-0.100***	-0.100***	-0.100***
	(0.001)	(0.0004)	(0.0004)	(0.0004)
professional_position4	-0.170***	-0.170***	-0.170***	-0.170***
	(0.001)	(0.0004)	(0.0004)	(0.0004)
professional_position5	-0.212***	-0.212***	-0.212***	-0.212***
	(0.001)	(0.0003)	(0.0003)	(0.0003)
sexF	-0.114***	0.00001	-0.050***	-0.100***
	(0.0003)	(0.0001)	(0.0001)	(0.0001)
Konstante	8.239***	8.239***	8.239***	8.239***
	(0.001)	(0.001)	(0.001)	(0.001)
Beobachtungen	2,970,032	2,970,032	2,970,032	2,970,032
R-Quadrat	0.524	0.839	0.842	0.848
Residual Std. Error (df = 2970019)	0.270	0.120	0.120	0.120

Anmerkungen: Die Signifikanzniveaus sind *p<0.05; **p<0.01; ***p<0.001. Standardfehler werden in Klammern ausgewiesen.

A.2 Weitere Simulationsergebnisse

A.2.1 Koeffizienten und Standardfehler

Abbildung 14: Boxplot Geschlechter-Koeffizient und Standardfehler, $\beta = -0.08$

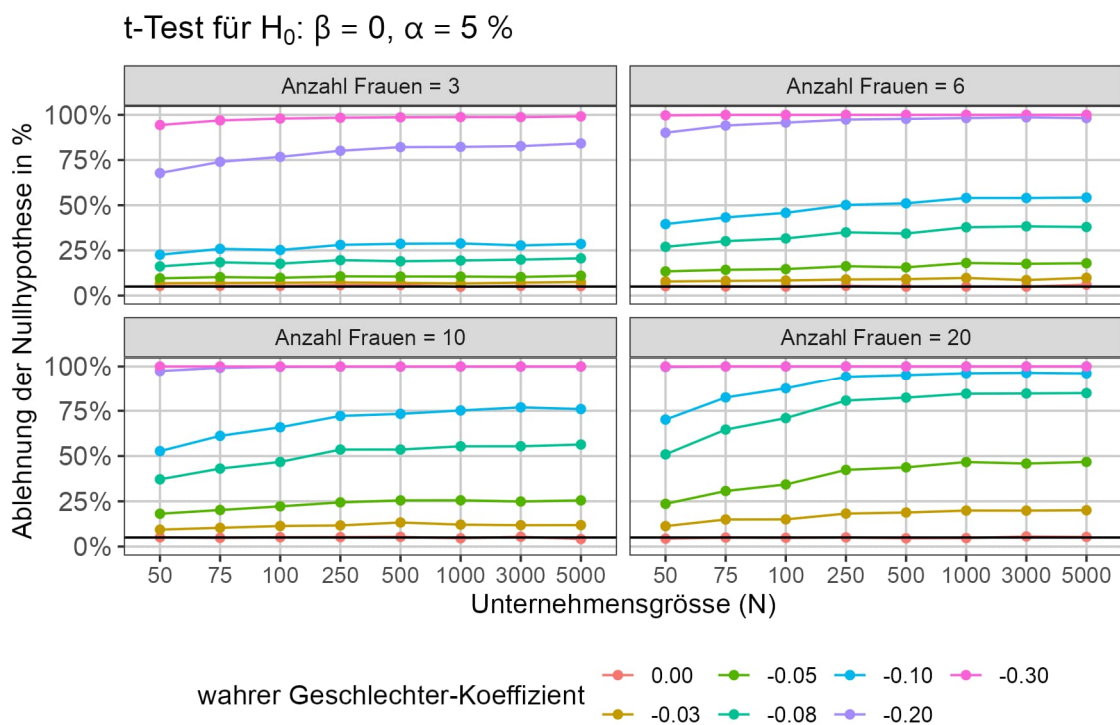


Anmerkungen: Die Grafik zeigt die Streuung der geschätzten Geschlechter-Koeffizienten als Boxplots für verschiedene Unternehmensgrößen und Frauenanteile. Der wahre Geschlechter-Koeffizient ist -0.08. Der durchschnittliche Standardfehler ist in Klammern angegeben. Die Ergebnisse basieren auf einer Monte-Carlo-Simulation mit 2000 bis 8000 Replikationen.

A.2.2 Weitere Ergebnisse mit absolutem Frauenanteil

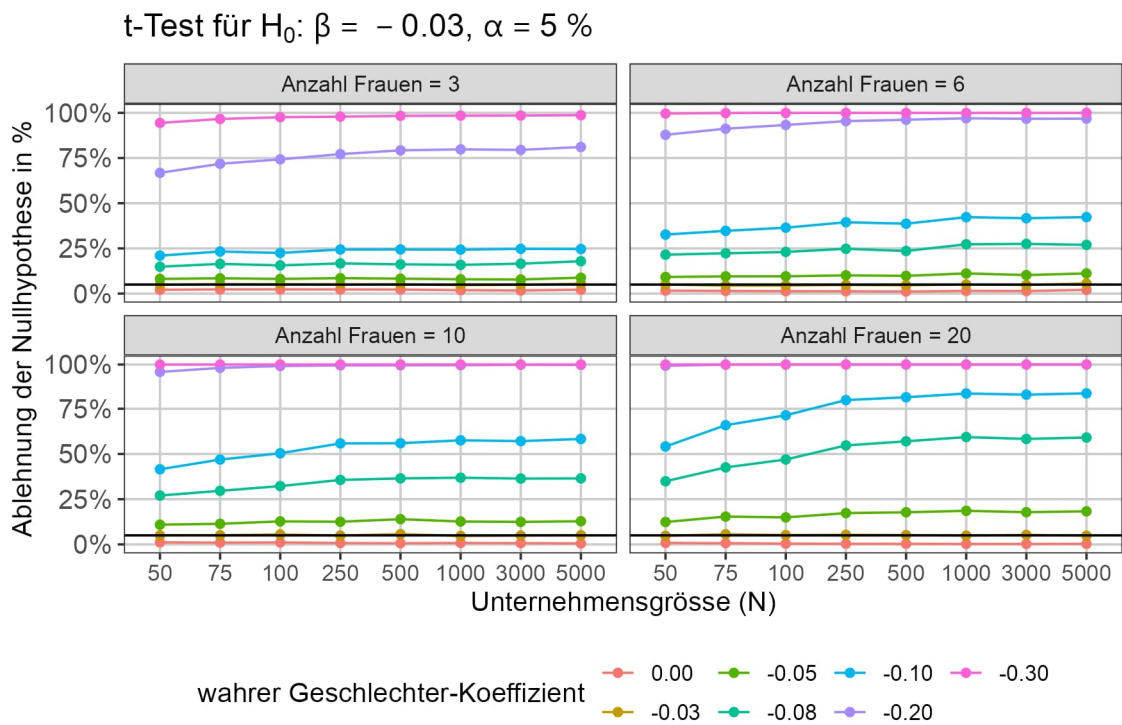
In Abbildung 15 sind dieselben Ergebnisse für die Szenarien mit einer kleinen, absoluten Anzahl Frauen (3 bis 20) dargestellt. Die Ergebnisse verdeutlichen, dass hier der zweiseitige t-Test aufgrund der grossen Standardfehler in Szenarien mit moderaten Lohnunterschieden selten in der Lage ist, unerklärbare Lohnunterschiede zwischen Frauen und Männern korrekt zu erkennen. Bei 10 Frauen im Unternehmen liegt die Mächtigkeit des t-Tests höchstens bei 25%, wenn der wahre Koeffizient -0.05 (d.h. 5%) beträgt, wobei der Wert mit der Unternehmensgrösse nur geringfügig ansteigt. Mit anderen Worten: Auch in grossen Unternehmen ist die Performance des t-Tests vergleichsweise schwach, wenn nur eine sehr geringe Anzahl Frauen beschäftigt sind.

Abbildung 15: Mächtigkeit des zweiseitigen t-Tests, absoluter Frauenanteil



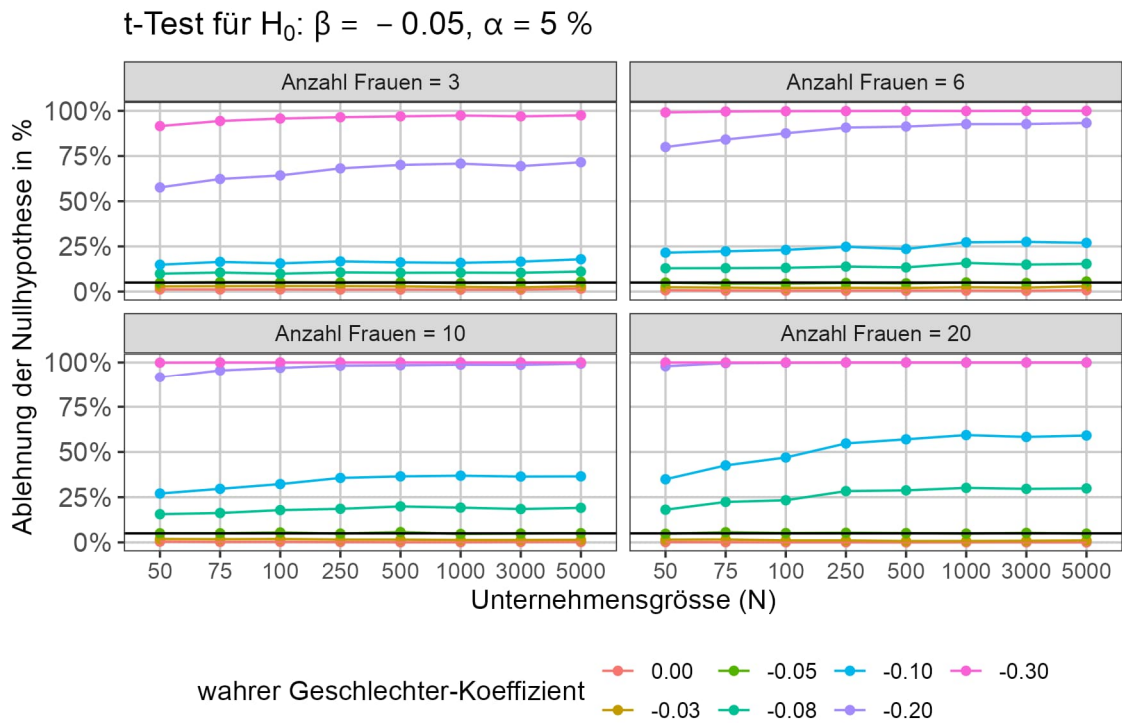
Anmerkungen: Die Grafik zeigt, wie häufig die Nullhypothese $H_0: \beta = 0$ in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten abgelehnt wird. Wenn der wahre Geschlechter-Koeffizient kleiner als null ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich null ist, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers (rote Linie).

Abbildung 16: Mächtigkeit des einseitigen t-Tests (-0.03), absoluter Frauenanteil



Anmerkungen: Die Grafik zeigt, wie häufig die Nullhypothese $H_0: \beta = -0.03$ in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten abgelehnt wird. Wenn der wahre Geschlechter-Koeffizient kleiner als -0.03 ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich -0.03, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers (rote Linie).

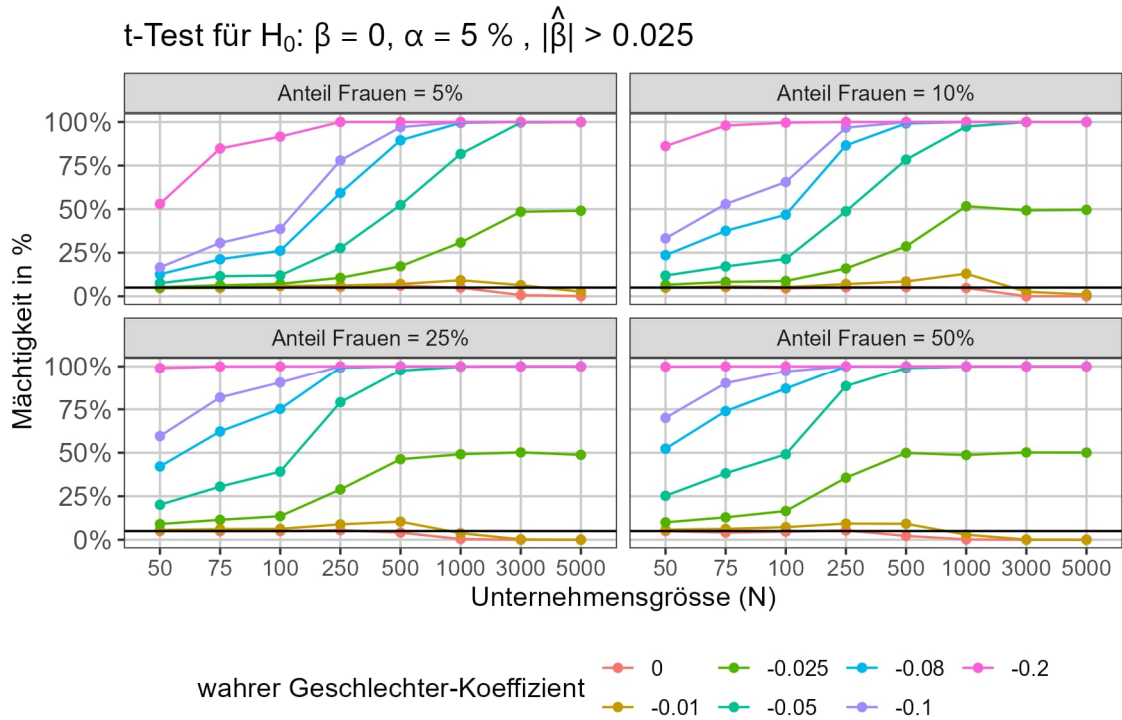
Abbildung 17: Mächtigkeit des einseitigen t-Tests (-0.05), absoluter Frauenanteil



Anmerkungen: Die Grafik zeigt, wie häufig die Nullhypothese $H_0: \beta = -0.05$ in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten abgelehnt wird. Wenn der wahre Geschlechter-Koeffizient kleiner als -0.05 ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich -0.05, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers (rote Linie).

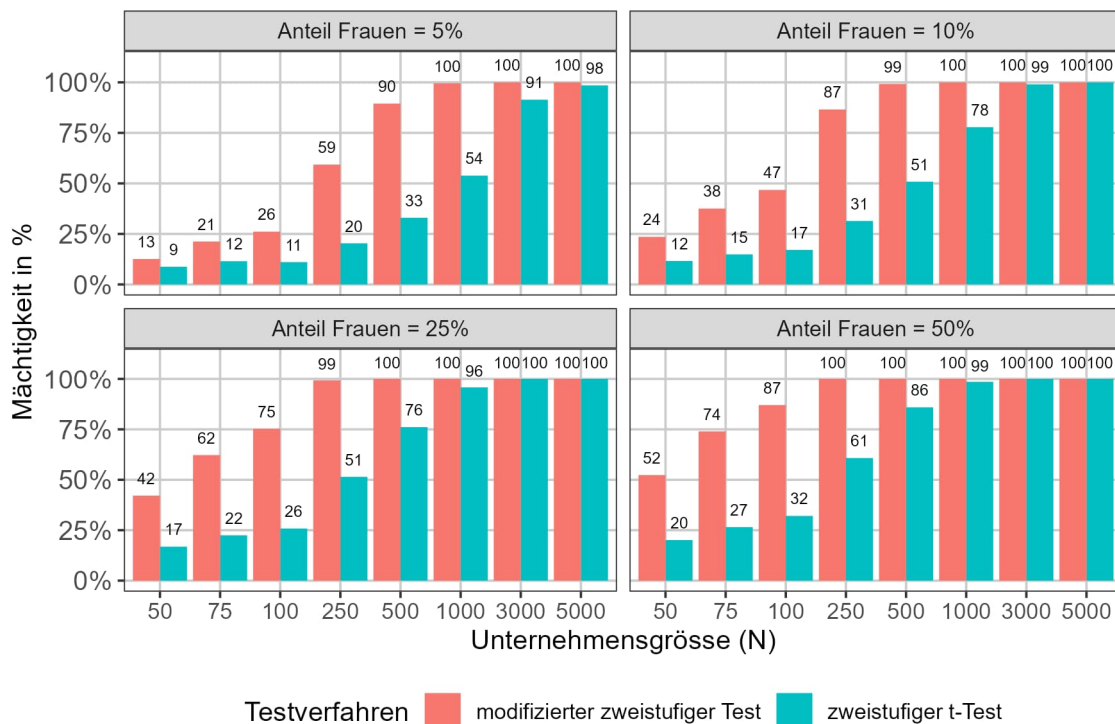
A.2.3 Weitere Ergebnisse zum modifizierten Test

Abbildung 18: Mächtigkeit des modifizierten Tests, Schranke 0.025, relativer Frauenanteil



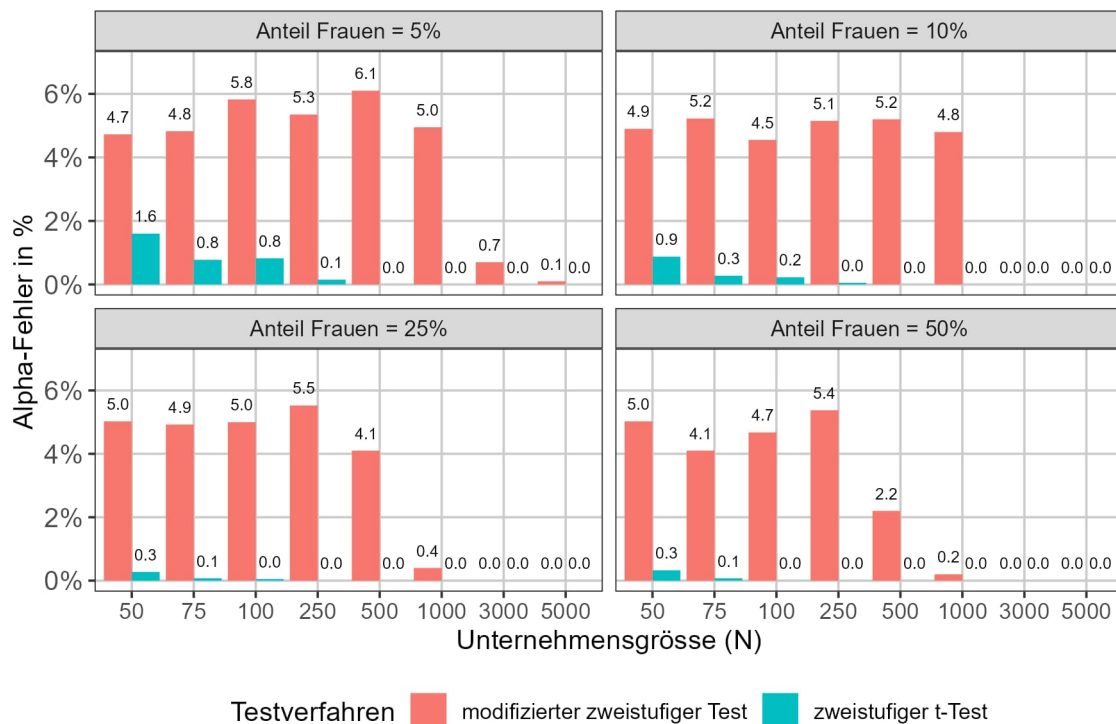
Anmerkungen: Die Grafik zeigt, wie häufig der modifizierte Lohngleichheitstest in verschiedenen Szenarien nach Unternehmensgrösse, Frauenanteil und Grösse des wahren Geschlechter-Koeffizienten verworfen wird. Dieser besteht aus Schritten: einem zweiseitigen t-Test der Nullhypothese eines Nulleffekts sowie dem Überschreiten des geschätzten Koeffizienten von 0.025 im Betrag. Wenn der wahre Geschlechter-Koeffizient gleich null ist, entspricht die Ablehnungsrate der Häufigkeit eines falsch-positiven Ergebnisses.

Abbildung 19: Mächtigkeit des modifizierten und herkömmlichen Tests, Toleranzschwelle 0.025, (b=-0.08)



Anmerkungen: Die Grafik zeigt einen Vergleich der Mächtigkeit zwischen dem modifizierten zweistufigen Test mit numerischer Toleranzschwelle von 0.025 (im Betrag) und dem herkömmlichen zweistufigen t-Test, wobei der zweite Test die Nullhypothese $H_0: \beta = -0.05$ auf einem Signifikanzniveau von 5% testet. Die Werte beziehen sich auf einen wahren Geschlechter-Koeffizienten von -0.08.

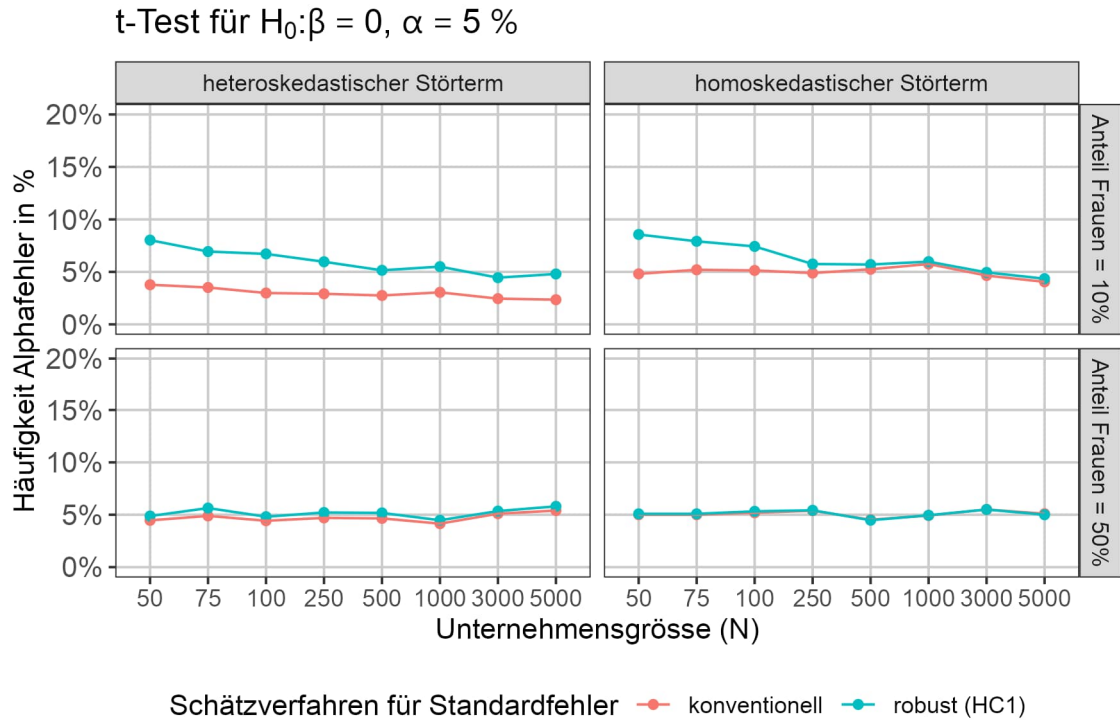
Abbildung 20: Alpha-Fehler des modifizierten und herkömmlichen Tests, Toleranzschwelle 0.025, (b=0)



Anmerkungen: Die Grafik zeigt einen Vergleich des Alpha-Fehlers (falsch-positive Ergebnisse) zwischen dem modifizierten zweistufigen Test mit numerischer Toleranzschwelle von 0.025 (im Betrag) und dem herkömmlichen zweistufigen t-Test, wobei der zweite t-Test die Nullhypothese $H_0: \beta = -0.05$ auf einem Signifikanzniveau von 5% testet. Die Werte beziehen sich auf einen wahren Geschlechter-Koeffizienten von null.

A.2.4 Heteroskedastizität und robuste Standardfehler

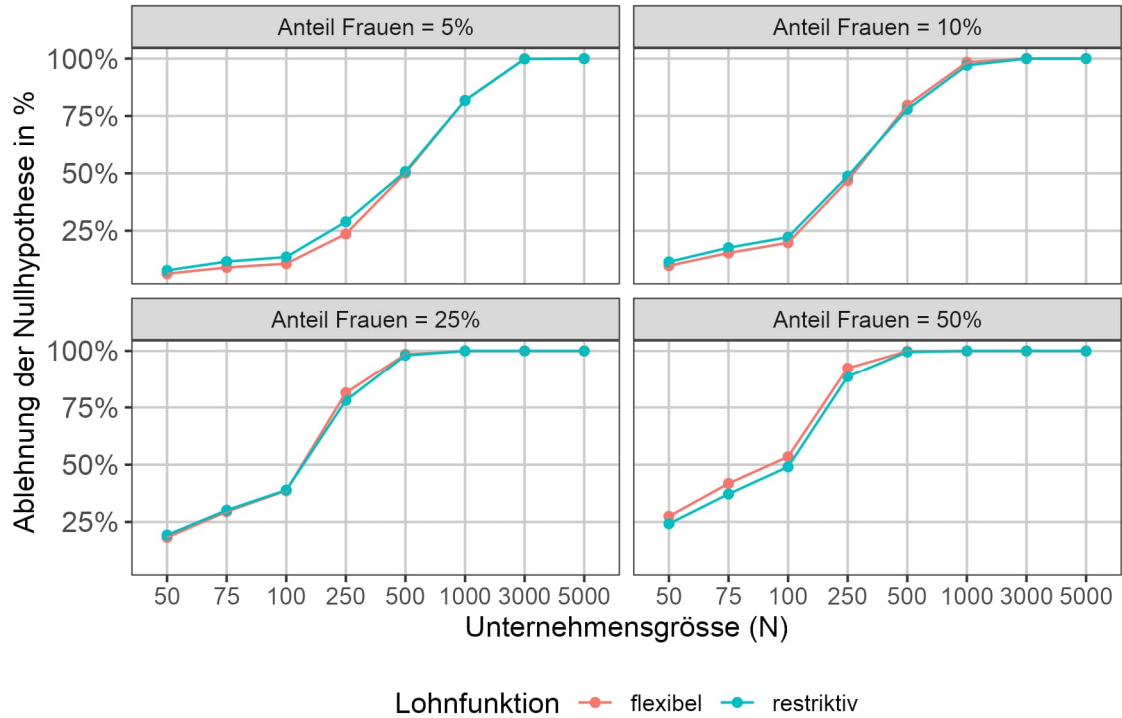
Abbildung 21: Alpha-Fehler nach Varianzfunktion und Schätzverfahren für Standardfehler



Anmerkungen: Die Grafik zeigt die Häufigkeit des Alpha-Fehlers des zweiseitigen t-Tests ($H_0: \beta = 0$) bei unterschiedlichen Störtermen (hetero- und homoskedastisch), Frauenanteilen, Unternehmensgrößen und Schätzverfahren für die Berechnung der Standardfehler. Der wahre Geschlechter-Koeffizient beträgt null, das heisst, die Nullhypothese ist wahr.

A.2.5 Spezifikation der Lohnfunktion

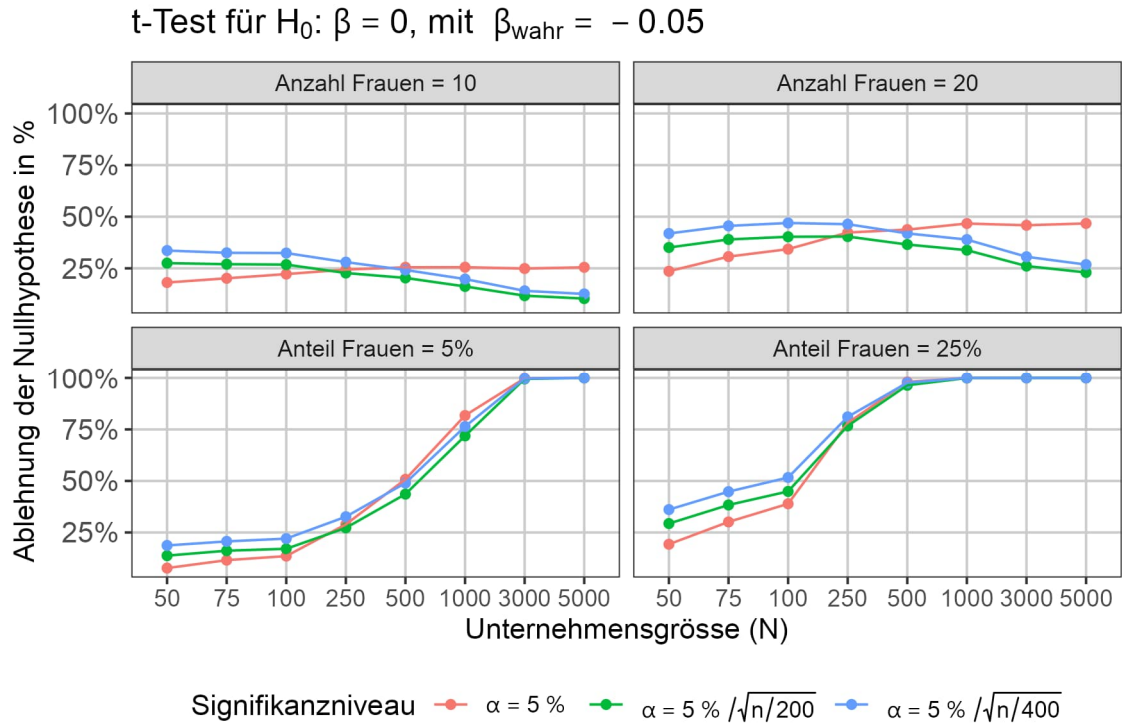
Abbildung 22: Mächtigkeit nach Spezifikation der Lohnfunktion



Anmerkungen: Die Grafik zeigt die Mächtigkeit des zweiseitigen t-Tests ($H_0: \beta = 0$) bei unterschiedlichen Lohnfunktionen (flexibel, restriktiv), Frauenanteilen und Unternehmensgrößen. Der wahre Geschlechter-Koeffizient beträgt -0.05.

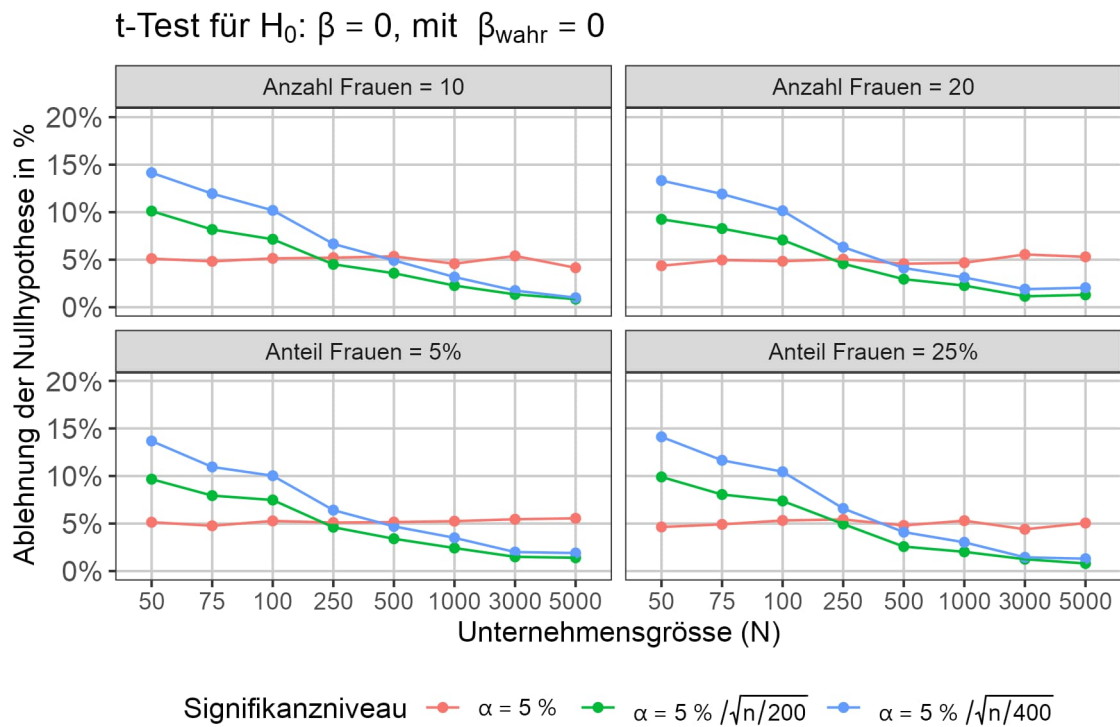
A.2.6 Alternative Signifikanzniveaus

Abbildung 23: Mächtigkeit nach Signifikanzniveau



Anmerkungen: Die Grafik zeigt, die Mächtigkeit des t-Tests in verschiedenen Szenarien nach Unternehmensgrösse und Frauenanteil, wenn das Signifikanzniveau unterschiedlich gewählt wird. Der wahre Geschlechter-Koeffizient beträgt hier -0.05. Die Ergebnisse basieren auf einer Monte-Carlo-Simulation mit 2000 bis 8000 Replikationen.

Abbildung 24: Häufigkeit des Alpha-Fehlers nach Signifikanzniveau



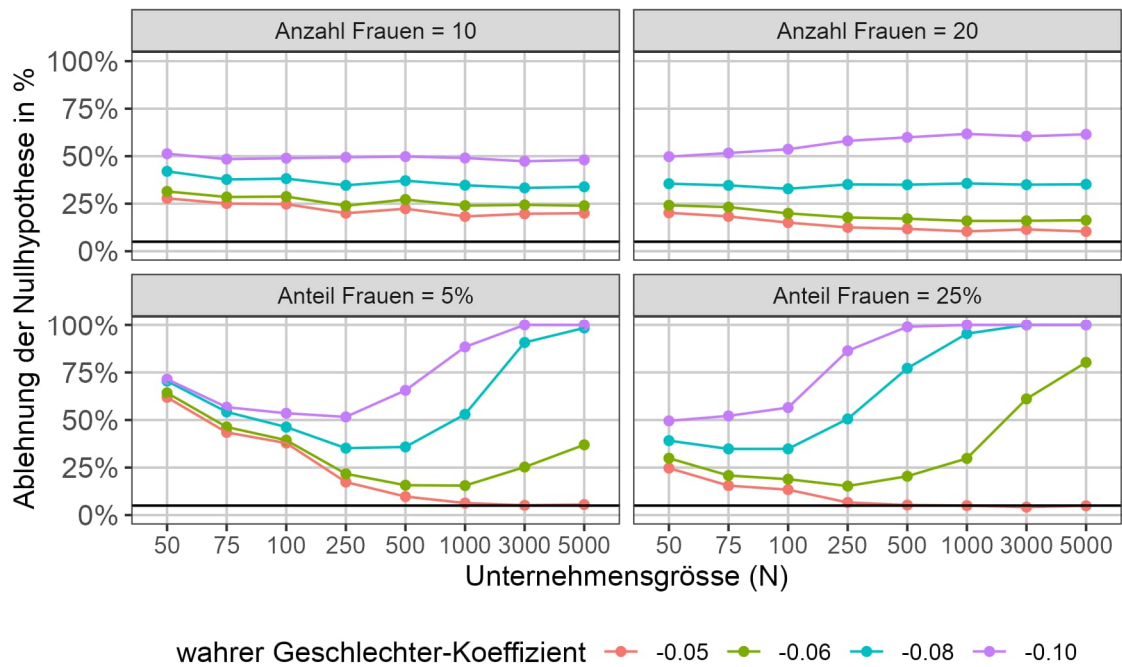
Anmerkungen: Die Grafik zeigt, die Häufigkeit des Alpha-Fehlers des t-Tests in verschiedenen Szenarien nach Unternehmensgrösse und Frauenanteil, wenn das Signifikanzniveau unterschiedlich gewählt wird. Der wahre Geschlechter-Koeffizient beträgt hier null, das heisst, die Nullhypothese ist wahr. Die Ergebnisse basieren auf einer Monte-Carlo-Simulation mit 2000 bis 8000 Replikationen.

A.2.7 Sequenzielles Testen

Hier werden die statistischen Eigenschaften des zweiten t-Tests untersucht, *gegeben*, dass die Nullhypothese des ersten Tests verworfen wurde. Es wird hier also nur ein Teil der Monte-Carlo-Replikationen betrachtet. Abbildung 25 zeigt die Ergebnisse des zweiten t-Tests für verschiedene Szenarien, wenn der erste t-Test die Nullhypothese verworfen hat. Die Kurven sind nicht wie gewohnt steigend, sondern U-förmig. Die Häufigkeit des Alpha-Fehlers (rote Linie) ist bei kleinen Unternehmen und/oder einer geringen Anzahl Frauen deutlich grösser als 5%. Der Alpha-Fehler erreicht nur bei grossen Unternehmen und ausgeglichenem Geschlechterverhältnis das korrekte Niveau von 5%. Dieses Ergebnis ist im Grunde nicht überraschend, da nur hinreichend extreme Schätzwerte zu einer Ablehnung der ersten Nullhypothese führen. Entsprechend ist dann im zweiten Test die Wahrscheinlichkeit höher, dass die Nullhypothese ebenfalls abgelehnt wird.

Abbildung 25: Mächtigkeit zweiter t-Test bei Ablehnung des ersten t-Tests

t-Test für $H_0: \beta = -0.05$, $\alpha = 5\%$, wenn $H_0: \beta = 0$ verworfen wurde



Anmerkungen: Die Grafik zeigt, wie häufig die zweite Nullhypothese $H_0: \beta = -0.05$ in verschiedenen Szenarien abgelehnt wird, gegeben, dass die erste Nullhypothese $H_0: \beta = 0$ bereits abgelehnt wurde. Wenn der wahre Geschlechter-Koeffizient kleiner als -0.05 ist, entspricht die Ablehnungsrate der Mächtigkeit des t-Tests. Wenn der wahre Geschlechter-Koeffizient gleich -0.05 ist, entspricht die Ablehnungsrate der Häufigkeit des Alpha-Fehlers (rote Linie).

