



Permutationstest i linjär regression

*Examensarbete för kandidatexamen i Matematik vid Göteborgs Universitet
Kandidatarbete inom civilingenjörsutbildningen vid Chalmers*

Christopher Backström

Adina Berg

Marcus Forsberg

Daniel Skorczynski

Samuel Vrede

Permutationstest i linjär regression

Examensarbete för kandidatexamen i Matematisk statistik inom Matematikprogrammet vid Göteborgs Universitet

Christopher Backström Daniel Skorczynski

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers

Adina Berg Marcus Forsberg Samuel Vrede

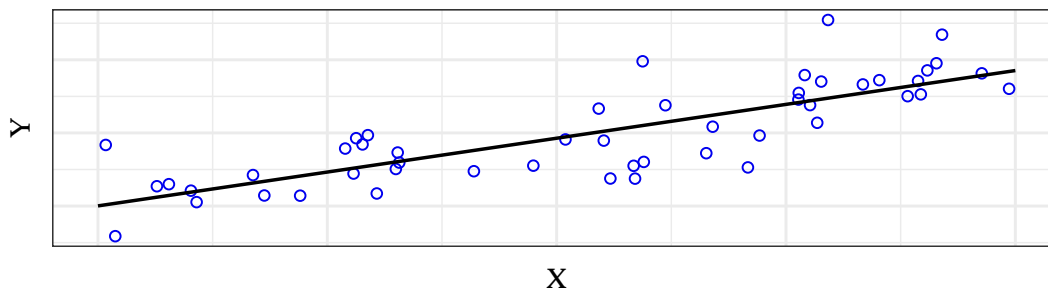
Handledare: Henrik Imberg

Examinator: Maria Roginskaya Marina Axelson-Fisk

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2018

Populärvetenskaplig presentation

Linjär regression är en metod som handlar om att hitta den linje som bäst förklarar ett antal mätpunkter. I praktiken används ofta linjär regression som ett hjälpmedel för att hitta trender i data. Trender kan vara alltifrån tidsberoendet hos väder till hur blodtryck varierar med ålder i en grupp. Att känna till om mätpunkter följer en linje kan vara bra för att statistiskt kunna förutse situationer och anpassa val därefter. Linjen hittas vanligtvis, och även i denna rapport, med den så kallade minsta kvadrat-metoden som går ut på att hitta en linje som är sådan att summan av kvadraten av avvikelserna mellan linjen och varje mätpunkt är minimal. I Figur 1 visas ett exempel på en linje som tagits fram med hjälp av minsta kvadrat-metoden.



Figur 1: En linjeanpassning till datapunkter med minsta kvadrat-metoden.

I Figur 1 kallas X den *förklarande variabeln* och kan exempelvis motsvara ålder, medan Y kallas *responsvariabel* och kan svara mot blodtryck. Linjen skulle därav kunna beskriva sambandet mellan ålder och blodtryck.

I praktiken är man ofta intresserad av att ta reda på om det finns ett samband mellan faktorer. Ett samband skulle kunna ses i en linjär modell om lutningen är positiv eller om lutningen är negativ. Om det inte finns ett samband förväntas mätpunkterna vara centrerade kring en linje med lutning noll.

Inom linjär regression används så kallade *statistiska test* som kan ge mått på hur troligt det är att mätpunkterna följer en viss linje. Beroende på vilket test som används beräknas ett testvärde utifrån en ekvation som är unik för det testet, och som beror på avvikelserna mellan mätpunkterna och den undersökta linjen. Testvärdet jämförs sedan mot testets så kallade *referensfördelning*, varpå man får ett procentuellt mått på hur bra den undersökta linjen förklarar mätpunkterna. Om det procentuella måttet är mindre än den förutbestämda felnivån¹ visar testet på att den tänkta linjen inte förklarar mätpunkterna tillräckligt bra.

Referensfördelningen beskriver vad som kan förväntas givet en viss hypotes, exempelvis vilken lutning som förväntas observeras i en viss situation givet att det inte finns något samband mellan variablerna. För att konstruera en referensfördelning kan t -test eller permutationstest användas. Referensfördelningen för t -test är en känd matematisk funktion, den så kallade t -fördelningen. För permutationstest genereras referensfördelningen istället för det specifika fallet, ofta med hjälp av simuleringar.

Permutationstest har ökat i användning sedan det utvecklades av Fisher och Pitman under 1930-talet, men användningen av permutationstest inom linjär regression är något som blivit vanligare först på senare år. För att utvärdera om det finns ett samband mellan X och Y i Figur 1 med hjälp av permutationstest, det vill säga undersöka om lutningen på en linje som förklarar mätpunkterna är skild från noll, konstrueras en referensfördelning genom att slumpmässigt låta mätpunkterna byta plats med varandra längs med X -axeln. Med minsta kvadrat-metoden anpassas sedan en linje till de omkastade mätpunkterna varpå ett nytt värde för lutningen beräknas. Detta slumpmässiga omkastande av mätpunkterna utförs sedan upprepade gånger och på så sätt erhålls en samling lutningar som utgör en referensfördelning. För att undersöka om en linje med lutning noll kan förklara mätpunkterna jämförs den ursprungliga linjens lutning mot denna referensfördelning varpå man, som med t -testet, erhåller ett mått på hur troligt det är att ett samband existerar. Här indikerar ett litet mått på att finns ett samband.

¹Inom statistik kallas detta för signifikansnivå och sätts vanligtvis till 5%.

Ett t -test kräver att avvikelserna mellan mätpunkterna och den undersökta linjen har en viss struktur, medan permutationstest baseras på färre sådana antaganden. Permutationstest är därav inte lika känsligt för extrema utfall hos avvikelserna. De specifika antaganden som krävs för de båda testen presenteras utförligare i huvuddelen av rapporten.

I detta arbete har data simulerats på ett sådant sätt att avvikelserna antar olika typer av extrema utfall, som bryter mot villkoren för det konventionella t -testet, men som uppfyller villkoren för permutationstest. Med hjälp av simuleringar med kända egenskaper hos den data testen utförts på, har prestandan hos de båda statistiska testen kunnat bedömas. I rapporten jämförs t -test och permutationstest för att avgöra i vilka situationer det ena är att föredra framför det andra inom linjär regression. Resultaten motiverar viss vidare användning av t -testet, även för data som bryter mot vissa grundläggande antaganden. Det presenteras även fall där permutationstest kan vara att föredra, speciellt i situationer där data har varierande spridning.

Sammanfattning

Inom statistisk dataanalys är linjär regression en metod som används för att anpassa en linje till datapunkter. Metoden kan användas för att studera och dra slutsatser kring samband mellan olika faktorer, vilket exempelvis kan vara relevant inom medicinska studier för att undersöka effekten av en viss behandling. När en linjär modell anpassats till en uppsättning datapunkter används ofta parametriska statistiska test för att dra slutsatser kring samband mellan olika variabler i modellen, där t -test är en typ av test som vanligen används. Som ett alternativ till konventionella parametriska test har permutationstestet börjat användas allt mer under de senaste decennierna. I permutationstest konstrueras en referensfördelning genom permutationer under nollhypotes, istället för att utgå från en redan existerande referensfördelning som t -testet gör. Permutationstest är beräkningstunga och kräver andra antaganden än t -test, som förutsätter normalfördelade feltermen med väntevärde 0 och konstant varians. I denna rapport presenteras en del av teorin för permutationstest inklusive kravet på utbytbarhet, samt testets tillämpning inom linjär regression. Med hjälp av simuleringar studeras data med feltermen från olika fördelningar för att jämföra hur permutationstest presterar jämfört med t -testet. Datan som undersöks har feltermen med tunga svansar, skev fördelning, utstickare och icke-konstant varians. Resultaten visar att t -testet verkar vara robust för alla typer av avvikelser förutom icke-konstant varians. Permutationstest bevarar signifikansnivån för samtliga datasimuleringar och är giltigt även för icke-konstant varians. Permutationstest presterar därmed ibland bättre, men aldrig sämre, än t -test för de undersökta datatyperna.

Nyckelord: *Permutationstest, enkel linjär regression, t -test, utbytbarhet, Freedman Lane, multipel linjär regression.*

Abstract

Within statistics, linear regression is a method for fitting a linear model to a set of data points. The method may be used to study and make conclusions about correlation between data, and may be used within medical studies to determine the effect of a treatment. To draw conclusions about relationships between variables in linear regression models, parametric tests have been used for a long time. As an alternative to the commonly used parametric tests, the permutation test has become more common during the last decades. Instead of relying on a known reference distribution, as the parametric tests do, the permutation test constructs its own distribution under the null hypothesis, based on permutations of the data. Permutation tests usually have a higher computational complexity and require other assumptions about the data as compared to the t -test which assumes normally distributed error terms with mean 0 and constant variance. This report covers a part of the theory regarding permutation tests, including the theoretical assumptions of exchangeability, and its applications in linear regression. Using simulations, data with different error term distributions have been generated to compare the permutation test to the t -test. The investigated data have error terms with heavy tails, a skewed distribution, outliers and non-constant variance. Data with such error terms are not uncommon in practice. This report covers simple as well as multivariate linear models. Results show that the t -test is robust to all kinds of error terms analyzed except for non-constant variance. Permutation tests preserve the significance level for all data simulations and is also valid for non-constant variance. In conclusion, permutation tests perform better than the t -test in some of the studied cases, though never worse.

Keywords: *Permutation tests, simple linear regression, t -test, exchangeability, Freedman Lane, multivariate linear regression.*

Förord

Detta kandidatarbete är skrivet av Christopher Backström och Daniel Skorczynski, Matematikprogrammet vid Göteborgs universitet, och Adina Berg, Marcus Forsberg och Samuel Vrede, Teknisk matematik vid Chalmers Tekniska Högskola.

I början av arbetet deltog alla gruppmedlemmarna i inläsning, diskussioner kring teoretisk bakgrund och inledande simuleringar. Christopher, Adina, Marcus och Samuel har stått för kodimplementeringar, och Marcus och Samuel har undersökt och jämfört olika implementationsvals effekt på beräkningstid. Vidare har Christopher och Daniel bidragit till arbetet genom att undersöka alternativa statistikor.

För de slutgiltiga implementationerna och simuleringarna har Samuel ansvarat, där Christopher och Adina har hjälpt till att verifiera resultaten och Marcus har skrivit kod för att generera samtliga grafer som används i rapporten. Gruppen i helhet har deltagit i tolkning av resultat, diskussion och slutgiltig korrekturläsning. Nedan ges en kort sammanfattning av gruppmedlemmarnas personliga bidrag till rapporten:

- Christopher är huvudförfattare till *avsnitt 2.1 Statistisk slutledning*, *avsnitt 2.2 Linjär regression* samt *B Exempel på utbytbara beroende slumpvariabler*. Christopher har även bidragit till *avsnitt 5 Resultat* och *avsnitt 6 Diskussion*.
- Adina är huvudförfattare till *avsnitt 3 Permutationstest i Linjär Regression* och *avsnitt 6 Diskussion*. Adina har även bidragit till diskussion kring *avsnitt 2.3 Permutationstest* och *avsnitt 5 Resultat*.
- Marcus är huvudförfattare till *Sammanfattning*, *Abstract*, *Populärvetenskaplig presentation*, *avsnitt 2.2.1 Hypotestest i linjär regression*, *avsnitt 2.4 Praktiska aspekter* samt *avsnitt A Övriga Resultat*, och medförfattare till *avsnitt 5 Resultat*.
- Daniel är huvudförfattare till *avsnitt 1.1 Bakgrund* och *avsnitt 2.3 Permutationstest*, och medförfattare till *avsnitt 5 Resultat* och *avsnitt 6.1 Resultatdiskussion*. Daniel har även skrivit ett första utkast till den populärvetenskapliga presentationen.
- Samuel är huvudförfattare till *avsnitt 3.4 Permutationstest och p-värden* och *avsnitt 4 Metod*, och medförfattare till *avsnitt 6 Diskussion*.

Vid sidan av projektet har vi fört en loggbok samt en veckovis dagbok.

Vi vill avslutningsvis tacka vår handledare Henrik Imberg för det stöd och råd vi fått genom arbetets gång.

Innehåll

1	Inledning	1
1.1	Bakgrund	1
1.2	Syfte	1
1.3	Metod	1
1.4	Avgränsningar	2
1.5	Konventioner	2
2	Teoretisk bakgrund	2
2.1	Statistisk slutledning	2
2.2	Linjär regression	3
2.2.1	Hypotestest i linjär regression	4
2.3	Permutationstest	5
2.3.1	Permutationstest - beskrivning och motivering	5
2.3.2	Utbytbarhet - ett nödvändigt villkor	6
2.3.3	Permutationstest med teckenbyte	7
2.4	Praktiska aspekter	7
3	Permutationstest i linjär regression	7
3.1	Permutationstest i simpel linjär regression	7
3.2	Freedman-Lanes permutationstest för multipel linjär regression	8
3.3	Permutationstest med teckenbyte	9
3.4	Permutationstest och p -värden	9
4	Metod för simuleringsstudie av permutationstest	10
4.1	Modellavvikelser	10
4.1.1	Tunga svansar	10
4.1.2	Skevfördelade feltermmer	10
4.1.3	Utstickare	11
4.1.4	Icke-konstant varians	11
4.2	Simpel linjär regression	11
4.3	Multipel linjär regression	13
5	Resultat	13
5.1	Simpel linjär regression	13
5.1.1	Tunga svansar	13
5.1.2	Skevfördelade feltermmer	13
5.1.3	Utstickare	15
5.1.4	Icke-konstant varians	15
5.2	Multipel linjär regression	16
5.2.1	Tunga svansar	16
5.2.2	Skevfördelade feltermmer	17
5.2.3	Utstickare	17
5.2.4	Icke-konstant varians	17
6	Diskussion	19
6.1	Resultatdiskussion	19
6.2	Allmän diskussion	19
6.2.1	R^2 -statistika i permutationstest	19
6.2.2	Konfidensintervall via permutationstest	20
6.2.3	Befintliga implementationer av permutationstest	20
6.3	Etiska aspekter	20

A	Övriga resultat	i
A.1	Simpel linjär regression	i
A.1.1	Tunga svansar	i
A.1.2	Skevfördelade feltermmer	ii
A.1.3	Utstickare	iv
A.1.4	Icke-konstant varians	iv
A.2	Multipel linjär regression	v
A.2.1	Tunga svansar	v
A.2.2	Skevfördelade feltermmer	vi
A.2.3	Utstickare	vi
A.2.4	Icke-konstant varians	vii
B	Exempel på utbytbara beroende slumpvariabler	viii
C	Kod	viii

1 Inledning

1.1 Bakgrund

För att beskriva samband mellan en eller flera förklarande variabler och en responsvariabel används ofta linjär regression. I en sådan modell inkluderas en felterm som beskriver individuella slumpmässiga avvikelser mellan modell och observation. Ett antagande som ofta görs är att anta att dessa feltermerna är oberoende och normalfördelade med konstant varians. Därpå är det vanligt att ett parametriskt test, som förutsätter att alla tre antaganden är uppfyllda, följer för vidare inferens. I praktiken uppfyller dock inte alltid feltermerna dessa antaganden, varför andra metoder utvecklats.

Ett alternativ till de fördelningsbaserade testen är *permutationstest*, vilket är en typ av test som i samband med linjär regression är ett test som utförs utan specifika antaganden om feltermernas fördelning. För att ett permutationstest ska kunna vara giltigt krävs dock ett antagande om *utbytbarhet*. Detta är exempelvis uppfyllt då feltermerna är oberoende och likafördelade. Det finns även vissa fall då vi har utbytbarhet trots att båda dessa krav inte är uppfyllda. Detta innebär att permutationstest kan användas i vidare utsträckning än många av de klassiska fördelningsbaserade testen, och de kan även användas som jämförelse för att utvärdera hur väl andra metoder presterar. En naturlig följdfråga i sammanhanget är således varför testet inte används i större utsträckning.

Permutationstest utvecklades redan under 1930-talet av Fisher och Pitman, och har med utvecklingen av allt effektivare datorer ökat i popularitet. På senare tid har det även börjats att användas i linjär regression. Eftersom antalet möjliga permutationer snabbt växer med antalet element är permutationstest vanligen beräkningsintensiva, vilket ställer krav på effektiv implementation². Med fortsatt ökad tillgång på beräkningskraft är det rimligt att tro att tillämpningsmöjligheterna samt bruket av permutationstest kommer att fortsätta att öka framöver.

Ett exempel på verklig data där ett permutationstest hade kunnat användas återfinns i Andelin m. fl. (2016) där precision av en glukosmätare, DexCom, studeras i relation till kapillärt respektive venöst blodsocker. Från Figur 2 i Andelin m. fl. (2016) framgår det att avvikelserna från den presenterade modellen både är skevfördelade och har icke-konstant varians³. Ytterligare ett exempel där ett permutationstest hade kunnat användas återfinns i Tancredi m. fl. (2017) där det undersöktes hur två hormoner, aldosteron och renin, relaterar till diabetesrelaterade komplikationer. Från Figur 2 i Tancredi m. fl. (2017) kan närvaron av utstickare bland residualerna observeras.

De två föregående exemplen visar att det förekommer fall med verklig data där residualerna inte är normalfördelade, vilket inte är ovanligt förekommande, och denna uppsats kommer via en simuleringsstudie att behandla fyra typer av brott mot detta fördelningsantagande. Utöver residualer med utstickare, icke-konstant varians och skevhet i fördelningen studerades residualer med tunga svansar⁴.

1.2 Syfte

Syftet med detta projekt är att studera permutationstest i linjära regressionsmodeller med icke-normalfördelade eller icke-likafördelade feltermerna, och jämföra dessa med klassiska normalfördelningsbaserade test med avseende på robusthet och effektivitet.

1.3 Metod

Denna rapport är resultatet av en kombinerad litteratur- och simuleringsstudie. Litteraturdelen har bestått i att ta fram förslag till permutationstest i linjära modeller, som korrigerats med hjälp av befintlig litteratur kring ämnet, parallellt med teoretisk inläsning. I simuleringsdelen har permutationstest implementerats varpå simulering av olika typer av data som skulle kunna vara problematiska för normalfördelningsbaserade test har genomförts. För dessa olika datatyper

²För $n=20$ element finns $n! = 20! \approx 2.43 \cdot 10^{18}$ möjliga permutationer.

³Figur 2 i Andelin m. fl. (2016) illustrerar MARD (mean absolute relative difference, det vill säga relativ absolut skillnad mellan glukosvärden mätt med DexCom jämfört med kapillära respektive venösa blodsockervärden) som en funktion av faktiskt sockervärde.

⁴Residualer med tunga svansar innebär att fördelningen har fler kraftiga avvikelser än en normalfördelning.

jämfördes permutationstest med normalfördelningsbaserade test. De situationer som studerats är feltermar med tunga svansar, skev fördelning, utstickare samt icke-konstant varians.

Simuleringarna som presenteras i denna rapport gjordes i *R* (R Core Team, 2017), version 3.4.1, och utnyttjade utöver grundpaket även *moments* (Komsta och Novomestky, 2015), version 0.14, och *MASS* (Venables och Ripley, 2002), version 7.3-49, för beräkningar och simuleringar. För grafisk visualisering användes *ggplot2* (Wickham, 2009), version 2.2.1.

1.4 Avgränsningar

Permutationstesten vi studerar gäller endast linjära modeller, eller modeller som efter transformation och inkludering av högre ordningens termer kan skrivas på sådan form. Vi fokuserar i första hand på Freedman-Lanes metod för permutationstest i multipel linjär regression, vilken introduceras i avsnitt 3.2.

I simuleringsstudien fokuserar vi på enkla regressionsmodeller samt regressionsmodeller med två förklarande variabler. Parametrar skattas med minsta kvadrat-metoden. Notera att permutationstesten dock kan användas i allmänna linjära modeller med godtyckligt antal förklarande variabler, och även i kombination med andra skattningsmetoder än minsta kvadrat-metoden.

1.5 Konventioner

I denna uppsats använder vi ordet *utstickare* när vi syftar på kraftigt avvikande värden från den skattade regressionen.⁵ Med en standardiserad fördelning menar vi en fördelning med varians 1 och väntevärde 0.

2 Teoretisk bakgrund

Teoriavsnittet i denna rapport är uppdelat i tre delar: två inledande avsnitt med nödvändiga förkunskaper inom statistisk slutledning och linjär regression, följt av ett avsnitt om permutationstest i generella termer. I avsnitt 3 presenteras teori kring permutationstest specifikt i linjära modeller.

2.1 Statistisk slutledning

I det här avsnittet presenteras grundläggande begrepp inom statistisk slutledning: nollhypotes, alternativhypotes, p -värde, signifikansnivå, Typ 1- och Typ 2-fel, konservativa test, exakta test samt icke-giltiga test. Innehållet baseras i huvudsak på kapitel 2 i Good (2000).

Exempel 2.1. Låt $\{172, 172, 173, 174, 175, 177, 179, 179, 184, 185\}$ vara ett stickprov av längder för en population. Antag att populationen är normalfördelad med känd varians $\sigma^2 = 4$, men med ett okänt medelvärde μ . Låt H_0 vara antagandet att $\mu = 180$ och H_1 antagandet att $\mu \neq 180$. Medelvärdet, $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 177$, för stickprovet är en realisation av medelvärdet av slumpvariabeln X , $\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i$. Där väntevärdet är $E[\bar{X}] = \mu$, och variansen är $\text{var}[\bar{X}] = \frac{\text{var}[X_i]}{10} = \frac{\sigma^2}{10}$. Således kommer under H_0 antagandet $\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{10}}} \stackrel{H_0}{\sim} N(0, 1)$. Teststatistikan $T_{obs} = \frac{\bar{x} - 180}{\sqrt{\frac{4}{10}}} = -4.7$ är då under nollhypotesen en realisation av T som följer en standardiserad normalfördelning. Låt $\phi(\cdot)$ vara fördelningsfunktionen för den standardiserade normalfördelningen. Då är $P(|T_{obs}| \leq T) = 2 - 2\phi(4.7) < 0.0001$. Eftersom sannolikheten för att få värdet T_{obs} är mycket låg förkastas H_0 .

Inom statistisk slutledning vill vi dra slutsaster om egenskaper för en eller flera variabler. För att hitta sådana egenskaper används en *statistika* som beskriver en egenskap hos ett stickprov. Exempel på denna egenskap kan vara stickprovets medelvärde eller varians. Givet ett slumpmässigt stickprov från en population kan en statistika beräknas som fångar den eftersökta egenskapen. Genom att anta komplementet till egenskapen vi vill påvisa, kan en referensfördelning för statistikan, som är en slumpvariabel, tas fram. Om vi exempelvis vill visa att medelvärdet av en population är skiljt

⁵Utstickare brukar kallas *outliers* på engelska.

ifrån 180, antar vi att det är 180 som i exemplet. Detta antagande kallas för *nollhypotes*. Vi har även en *alternativhypotes*, som vi kan sätta till att populationens medelvärde antingen är större än, mindre än eller skiljt ifrån 180.

Eftersom statistikan beräknas från ett slumpmässigt stickprov av en population, kommer vi observera olika värden av statistikan för olika stickprov. Statistikan har under antagandet av nollhypotesen en fördelning, och om detta antagande är sant kommer den observerade statistikan vara en realisation av den fördelningen. Vi kan då jämföra den observerade statistikan med fördelningen, för att se hur sannolik den är. Om en observerad statistika är osannolik att observera i statistikans fördelning under nollhypotesen, förkastas nollhypotesen, och accepteras annars.

För att kvantifiera hur extrem observationen T_{obs} är i T -statistikans fördelning under nollhypotesen, används dess fördelningsfunktion. Detta uttrycks som sannolikheten att en realisation av T är lika eller mer extrem än T_{obs} . Mer exakt kan det skrivas som, $P(|T_{obs}| \leq T)$ och kallas för *p-värdet*.

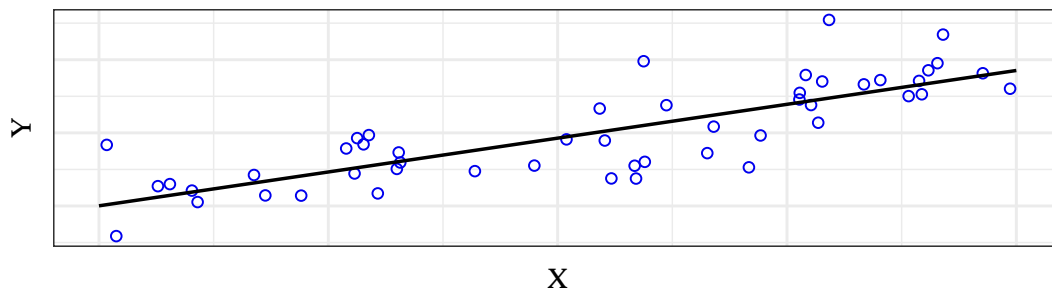
Vid statistisk slutledning finns det alltid en risk att fel slutsatser dras. Exempelvis kan nollhypotesen förkastas när den är sann eller accepteras när den är falsk. Benämningen för de två olika fallen är *Typ 1-fel*, respektive *Typ 2-fel*. Sannolikheten för ett test att göra ett Typ 1-fel kallas för *signifikansnivån*, och betecknas ofta α .

Statistiska test ska även identifiera samband när sådana existerar, vilket innebär att nollhypotesen förkastar när denna är falsk. Styrkan för ett test är sannolikheten att förkasta nollhypotesen givet att den är falsk. Detta är även ekvivalent med komplementet till Typ 2-felet. Om flera test jämförs säger vi att det test med högst styrka är *effektivast*. Det optimala testet kan därför tolkas som ett test med signifikansnivå och styrka lika med 0 respektive 1. Detta är dock inte alltid möjligt och signifikansnivån bestäms då på förhand, ofta till 5%. Därmed är ett test ett *giltigt test* om sannolikheten för att göra ett Typ 1 fel inte överskrider den valda signifikansnivån.

Olika avvikelser i datan från de antaganden ett visst test bygger på kan få statistiska test att göra Typ 1-fel i högre grad än signifikansnivån. I den här rapporten kallar vi statistiska test som gör Typ 1-fel med lägre sannolikhet än signifikansnivån för *konservativa test*. Test där sannolikheten för Typ 1-fel överskrider signifikansnivån benämns som *icke-giltiga test* då dessa förkastar nollhypotesen oftare än signifikansnivån anger. De test där sannolikheten för ett Typ 1-fel är lika med signifikansnivån benämns som *exakta test*. Test som bevarar signifikansnivån trots avvikelser från testets antaganden kallas för *robusta test*.

2.2 Linjär regression

I det här avsnittet introduceras linjär regression och skattning av parametrar i sådana modeller. Innehållet baseras i huvudsak på kapitel 1 och 3 i Rawlings, Pantula och Dickey (1998), och kapitel 14 i Rice (2007).



Figur 2: En linjeanpassning till datapunkter med minsta kvadrat-metoden.

Givet en mängd punkter $\{(x_i, y_i)\}_{i=0}^n$, vill vi beskriva ett samband mellan y och x , där vi betraktar y som *responsvariabel* och x som en *förklarande variabel*. Ett samband mellan en förklarande variabel och en responsvariabel kan i många fall beskrivas med, eller approximeras till, ett linjärt samband på formen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

där β_0 , β_1 är konstanter och ε_i en felterm, med andra ord en slumpmässig avvikelse som inte kan förklaras av x . Det finns oändligt många linjer av ovanstående form som kan tänkas beskriva

sambandet, men vi söker den som bäst beskriver y_i . Ett tillvägagångssätt för att definiera bäst är att minimera kvadratsumman av feltermerna, det vill säga, minimera $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. De optimala skattningarna $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ ges då av

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=0}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (1)$$

Den här metoden kallas för minsta kvadrat-metoden. För att hitta de optimala $\hat{\beta}$ -argumenten kan summan i (1) minimeras genom att beräkna de stationära punkterna för β_0 och β_1 . Detta leder till dessa explicita formler för skattningarna $\hat{\beta}_0$ och $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

där $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ och $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. För det här beviset hänvisas läsaren till kapitel 14 i Rice (2007). I Figur 2 ges ett exempel på en mängd datapunkter och en linje som anpassats med minsta kvadrat-metoden.

Under antaganden om icke-konstant varians, symmetrisk fördelad och oberoende felterm, presterar minsta kvadrat-metoden optimalt, eftersom den är ekvivalent med Maximum likelihood-metoden.

Modellen kan också generaliseras för ett godtyckligt antal förklarande variabler. Låt då y vara en linjär funktion av m förklarande variabler på formen,

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_m x_{mi} + \varepsilon_i. \quad (2)$$

För att komprimera notationen kan modellen skrivas i matrisform som

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

där

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix} \quad \text{och} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Koefficienterna $\boldsymbol{\beta}$ kan uppskattas genom att minimera den euklidiska normen av $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ för $\boldsymbol{\beta}$. Det motsvarar att hitta $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\|_2^2$. Det optimala $\boldsymbol{\beta}$ argumentet kan återigen beräknas genom att hitta en stationär punkt till ekvationssystemet. En explicit lösning till $\hat{\boldsymbol{\beta}}$ ges då av $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ och detta har fått namnet "normalekvationerna". För komplett bevis hänvisar vi till Demmel (1997).

2.2.1 Hypotestest i linjär regression

Låt oss säga att vi vill testa en hypotes på formen $H_0 : \beta_i = m$ mot alternativet $H_1 : \beta_i \neq m$, där β_i är någon parameter i den linjära modellen (2). Om feltermerna ε_i i regressionsmodellen antas normalfördelade kommer $\hat{\beta}_i$ vara normalfördelad⁶. Vidare, under antagande om oberoende och likafördelade feltermer med väntevärde 0 och konstant varians, leder detta till att statistikan

$$T = \frac{\hat{\beta}_i - m}{s(\hat{\beta}_i)} \quad (3)$$

följer en t -fördelning med lika många frihetsgrader som $s(\hat{\beta}_i)$. Här är $\hat{\beta}_i$ vår skattning av β_i , $s(\hat{\beta}_i)$ standardfelet⁷ för $\hat{\beta}_i$ och m värdet på β_1 under nollhypotesen som vi vill testa $\hat{\beta}_i$ mot.

För att sedan utföra ett t -test på det uppskattade $\hat{\beta}_i$ beräknar vi T enligt ekvationen ovan och jämför med motsvarande kritiska värde ur t -fördelningen för en given signifikansnivå och frihetsgrad

⁶Se sidorna 16-17 i Rawlings, Pantula och Dickey (1998).

⁷Se sida 15 i Rawlings, Pantula och Dickey (1998).

$\nu = n - p$, där n är antalet observationer och p antalet skattade parametrar i modellen (2). Vi säger att vi förkastar $H_0 : \beta_i = m$ mot $H_1 : \beta_i \neq m$ på signifikansnivå α om $|T| > t_\nu(\frac{\alpha}{2})$, där $t_\nu(\frac{\alpha}{2})$ är $\frac{\alpha}{2}$ kvantilen för t -fördelningen med ν frihetsgrader. På liknande sätt kan vi ställa upp modeller för att undersöka nollhypotesen $H_0 : \beta_i = m$ mot $H_1 : \beta_i < m$ eller $H_1 : \beta_i > m$.

2.3 Permutationstest

I denna del av teorin följer först en introduktion till permutationstest, som vidare konkretiseras med ett exempel. Därefter kommer vi att gå in på villkoren för att testet ska vara giltigt, där begreppet *utbytbarhet* kommer att studeras, och två tärningsexempel presenteras. Det första exemplet illustrerar ett fall då villkoren för utbytbarhet är uppfyllda, och det andra ett fall då så inte är fallet. Avslutningsvis kommer vi att se på ett fall där feltermerna inte är utbytbara, men där ett anpassat permutationstest ändå är giltigt.

2.3.1 Permutationstest - beskrivning och motivering

Permutationstest är ett bland flera typer av icke-parametriska statistiska test. Den studerade teststatistikan antas inte följa en känd parametrisk fördelning, så som för parametriska test, utan istället tas en möjlig referensfördelning för denna fram implicit.⁸

Grundidén är att varje permutation av den ursprungliga datan, under ett antagande om utbytbarhet⁹, kommer att generera ett nytt och lika troligt värde för teststatistikan. På så sätt erhålls en samling värden som bildar en fördelning som approximerar teststatistikans verkliga, dock okända, fördelning. Denna fördelning används sedan som referensfördelning för den ursprungliga teststatistikans värde, T_{obs} , ett värde som under nollhypotesen är lika sannolik att observera för teststatistikan som var och ett av de omräknade värdena för var permutation. Genom att jämföra T_{obs} mot denna referensfördelning kan därför ett p -värde för T_{obs} beräknas. Nedan presenteras ett exempel på hur ett permutationstest kan genomföras.

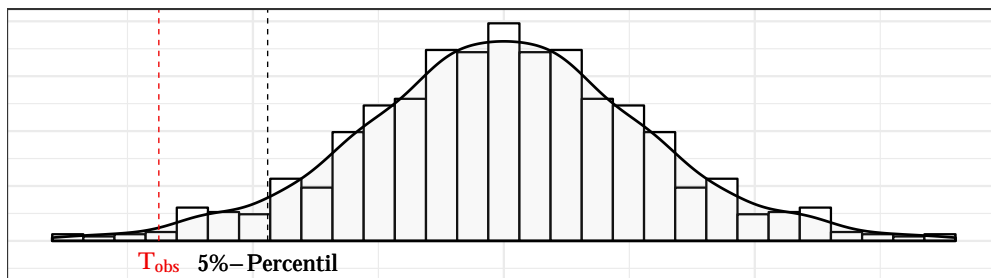
Exempel 2.2. Antag att det ska avgöras om en ny medicin är blodtryckssänkande. Antag även att det finns data att tillgå från blodtrycksmätningar för patienter i en grupp A om tio patienter, som behandlats med medicinen, och för en grupp B om tio patienter som behandlats med placebo. Låt signifikansnivån vara 5%. Vår nollhypotes är att det inte föreligger någon skillnad mellan medelvärdet för mätningarna i de båda grupperna. Den alternativa hypotesen är att medelvärdet för blodtrycket hos patienterna i grupp A är lägre än det för patienterna i grupp B.

Låt T_{obs} beteckna differensen mellan medelvärdena i grupp A och grupp B. Under ett antagande om utbytbarhet och om att nollhypotesen gäller, skulle en observation från grupp B lika gärna kunnat vara från grupp A. Grupptillhörighet A respektive B är då endast etiketter som kan bytas ut utan att differensens fördelning ändras. Således kan vi genom att byta etikett på observationer, mellan grupperna, erhålla två nya grupper A^* och B^* , som under nollhypotesen är lika sannolika som den ursprungliga observationen av medelvärdet för grupp A och för grupp B. Genom att beräkna teststatistikan, i detta fall differensen för gruppernas medelvärde, för denna permuterade data erhålls ett nytt värde för testvariabeln. Upprepas förfarandet många gånger resulterar detta i att en samling möjliga statistikavärden erhålls, som givet nollhypotesen utgör ett stickprov från fördelningen för testvariabeln.

Således kan T_{obs} jämföras mot denna samling av värden som i sin tur utgör en fördelning, vilket illustreras i Figur 3. Om värdet T_{obs} är osannolikt, det vill säga har ett p -värde lägre än signifikansnivån på 5%, förkastas nollhypotesen och vi drar slutsatsen att medicinen har en sänkande effekt på blodtryck.

⁸Här bör ordet *möjlig* betonas då det kan existera flera giltiga referensfördelningar för samma statistika samtidigt. För utförlig redogörelse kring detta se Aickin (2010).

⁹Tillsvidare kan vi betrakta detta som ett villkor som garanterar att vi genom att byta ut olika värden mellan två grupper genererar nya observationer av testvariabeln, exempelvis gruppernas medelvärden, som är lika troliga att observera som den ursprungliga observationen. *Utbytbarhet* definieras och behandlas närmare under nästa underrubrik.



Figur 3: Permuterad referensfördelning för skillnaden i medelvärde mellan två grupper.

Låt oss nu gå vidare till det centrala villkoret för permutationstest, *utbytbarhet*, som hittills endast beskrivits övergripigt.

2.3.2 Utbytbarhet - ett nödvändigt villkor

Utbytbarhet är en förutsättning för att ett permutationstest ska vara giltigt. Vi definierar det som följer

Definition 2.1. Låt $\pi: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ vara en godtycklig injektiv funktion. En samling stokastiska variabler, X_1, X_2, \dots, X_n , säges vara utbytbara om den simultana fördelningen är permutationsinvariant, det vill säga $F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = F_{X_1, X_2, \dots, X_n}(x_{\pi(1)}, x_{\pi(2)}, \dots, x_{\pi(n)})$.

Definitionen ovan innebär att en uppsättning slumpvariabler sägs vara utbytbara om den simultana fördelningen inte påverkas av att variabelernas ordning kastas om. Låt oss nu se på ett enkelt exempel då definitionen av utbytbarhet är uppfylld.

Exempel 2.3. Vi har två vanliga sexsidiga tärningar, X_1 och X_2 . De kastas och visar värdena x_1 och x_2 . Den simultana fördelningen kan då skrivas $P(X_1 = x_1, X_2 = x_2)$. I detta scenario är sannolikheten att den första tärningen visar en trea lika stor som den för att den andra ska visa en fyra, och utfallet av den ena tärningen påverkar inte utfallet av den andra. Därför kommer sannolikheten att få en trea på den första tärningen och en fyra på den andra vara samma som den för att observera en fyra på den första tärningen och en trea på den andra, det vill säga $P(X_1 = 3, X_2 = 4) = P(X_1 = 4, X_2 = 3)$. Den simultana fördelningen påverkas därför inte av att permutera utfallen och den är *permutationsinvariant*, varför vi har utbytbarhet.

I Exempel 2.3 gäller att tärningarnas utfall är oberoende och likafördelade, vilket är ett tillräckligt villkor för att uppnå utbytbarhet.

Proposition 2.1. Låt X_1, X_2, \dots, X_n vara oberoende och likafördelade slumpvariabler. Då gäller utbytbarhet.

Bevis.

$$\begin{aligned}
 F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \\
 &= P(X_1 \leq x_1)P(X_2 \leq x_2) \dots P(X_n \leq x_n) = \\
 &= P(X_{i_1} \leq x_1)P(X_{i_2} \leq x_2) \dots P(X_{i_n} \leq x_n) = \\
 &= P(X_1 \leq x_{i_1})P(X_2 \leq x_{i_2}) \dots P(X_n \leq x_{i_n}) = \\
 &= P(X_1 \leq x_{i_1}, X_2 \leq x_{i_2}, \dots, X_n \leq x_{i_n}) = \\
 &= F_{X_1, X_2, \dots, X_n}(x_{i_1}, x_{i_2}, \dots, x_{i_n}).
 \end{aligned}$$

Där i_1, i_2, \dots, i_n är godtyckligt permuterade index. Den första likheten följer från definitionen av den simultana fördelningsfunktionen, medan den andra likheten följer från oberoendet av slumpvariablerna. Den tredje likheten är en följd av att variablerna är likafördelade, ty exempelvis är $P(X_{i_1} \leq x_1) = P(X_1 \leq x_1)$ då vi har likafördelade slumpvariabler. Resten följer av definitionerna av oberoende och fördelningsfunktionen. \square

En konsekvens av detta är att permutationstest kan användas i de flesta situationer där t -test används. Vi presenterar nu ett exempel på ett fall där utbytbarhet mellan observationer inte gäller.

Exempel 2.4. Ett exempel då utbytbarhet inte är uppfyllt är då vi har två oliksidiga tärningar, säg en fyrsidig, och en andra sexsidig. Sannolikheten att slå en etta med den första tärningen och en femma med den andra tärningen är då $\frac{1}{4} \cdot \frac{1}{6}$, medan utfallet $\{5\}$ inte kan observeras på den första tärningen och den sannolikheten därför är noll, varför $P(X_1 = x_1, X_2 = 5) \neq P(X_1 = 5, X_2 = x_1)$. Per definition gäller då inte utbytbarhet.

I Exempel 2.4 gäller oberoendet men kravet på likafördelning bryts eftersom tärningarna har olika många utfall, varför vi inte har utbytbarhet i det fallet.

Värt att notera är att utbytbarhet inte medför oberoende och likafördelning. Exempel på utbytbara stokastiska variabler, är de från en multivariat normalfördelning, där variablerna har samma varians och där den parvisa korrelationen är lika för alla variabler. Bevis för detta presenteras i appendix B.

Kravet om utbytbarhet är ett nödvändigt villkor för att ett permutationstest ska vara giltigt, men under vissa omständigheter kan andra typer av permutationstest härledas, som är giltiga trots att utbytbarheten bryts, vilket vi nu ska se ett exempel på.

2.3.3 Permutationstest med teckenbyte

Permutationstest med teckenbyte kan användas då datan vi önskar permutera inte kan antas vara oberoende och likafördelad, exempelvis när variablerna har icke-konstant varians. Det som krävs för att detta test ska vara giltigt är att datans simultana fördelning ska vara invariant under teckenbyte (Winkler m. fl., 2014), vilket är fallet för symmetriskt fördelade variabler.¹⁰

I det fall då observationerna mellan variabler inte är utbytbara, är det tidigare beskrivna permutationstestet inte längre giltigt. Då kan ett annat permutationstest användas, givet att observationerna i varje variabel är symmetriskt fördelade kring 0. Sannolikheten att tecknet för en observation är negativt är lika med den att det är positivt. Därför kan tecknet betraktas som en realisation av en likafördelad oberoende slumpvariabel, och tecknet är därmed utbytbart. Istället för att slumpmässigt permutera datan permuteras de index för vilka ett teckenbyte ska ske.

2.4 Praktiska aspekter

Efter att ha presenterat teori och exempel på när permutationstest kan användas, är det nu möjligt att gå in på användningsområdet linjär regression. För att göra det behöver vi först nämna hur permutationstest vanligen utförs. Som tidigare beskrivits växer antalet möjliga permutationer snabbt med stickprovsstorleken, och i praktiken är det sällan möjligt att utföra ett komplett permutationstest. Detta för oss in på så kallade Monte-Carlo-simuleringar. Dessa går ut på att en mängd slumpmässigt valda permutationer används för att skapa en referensfördelning. Ofta räcker 1000 slumpgenererade permutationer för att åstadkomma tillräcklig hög styrka. (Nichols och Holmes, 2001).

3 Permutationstest i linjär regression

I detta avsnitt beskrivs hur ett permutationstest kan användas för att studera samband i linjära regressionsmodeller. Beroende på hur datan ser ut är det fördelaktigt att använda olika typer av permutationstest, varför avsnittet är uppdelat i tre områden: permutationstest i simpel linjär regression, multipel regression med Freedman-Lanes permutationstest samt permutationstest med teckenbyte.

3.1 Permutationstest i simpel linjär regression

Hur ett permutationstest kan användas för att testa en skattad regressionskoefficient i en simpel linjär regressionsmodell kommer att behandlas nu. Vi antar att datan följer en modell på formen

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \quad (4)$$

¹⁰Jämför med utbytbarhet under permutationsinvarians, avsnitt 2.3.2.

Vi vill här testa $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$, det vill säga om värdet på Y påverkas av värdet på X eller ej. Det går även att på samma sätt göra motsvarande ensidiga test med $H_1 : \beta_1 > 0$ eller $H_1 : \beta_1 < 0$. För detta test kan regressionskoefficientens T -statistika användas, men om ε_i inte kan antas normalfördelade kan vi inte motivera att den följer en t -fördelning. Förfarandet beskrivet i avsnitt 2.2.1 är därför inte giltigt, utan vi behöver istället generera T -statistikans referensfördelning som beskrivet i avsnitt 2.3.1.

Vi börjar med att ta fram ett värde på vår teststatistika för den fulla modellen (4), vilken vi kallar T_{obs} . Därefter studerar vi modellen under H_0 , som blir

$$Y_i = \beta_0 + \varepsilon_i,$$

vilket brukar kallas för den reducerade modellen. Då β_0 inte är känd vet vi inte de faktiska feltermerna $\varepsilon_i = Y_i - \beta_0$. Istället kan vi använda residualerna $e_i = Y_i - \hat{\beta}_0$, där $\hat{\beta}_0 = \bar{Y}$. Om feltermerna ε_i antas utbytbara gäller även att residualerna e_i är utbytbara under H_0 . Vi kan därför generera nya observationer

$$Y_i^* = \hat{\beta}_0 + e_i^*$$

av responsvariabeln som under H_0 är lika troliga som vår första observation. Vårt att notera är att vi i detta simpla fall med endast en förklarande variabel i vår modell lika gärna kan permutera responsvariablerna Y_i direkt, istället för att permutera e_i och addera dessa till $\hat{\beta}_0$.

De nya värdena Y_i^* på responsvariabeln används nu för att skatta nya värden på T -statistikan under H_0 . Genom att utföra ett stort antal slumpmässiga permutationer och påföljande skattningar ger detta en referensfördelning för teststatistikan, som kan användas för att testa hypotesen $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$.

3.2 Freedman-Lanes permutationstest för multipel linjär regression

Förfarandet för att genomföra ett permutationstest blir något mer komplicerat när vi antar en modell med fler än en förklarande variabel. För detta har ett flertal tillvägagångssätt föreslagits - se exempelvis Anderson och Legendre (1999). I detta avsnitt presenteras metoden föreslagen av Freedman och Lane (1983), sammanfattad av Anderson och Robinson (2001). Teorin presenteras för fallet med två förklarande variabler, men den är enkelt generaliserad till en allmän situation med ett godtyckligt antal förklarande variabler.

Den fulla modell med två förklarande variabler X och Z är

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i. \quad (5)$$

Vi vill nu testa $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$. Vi kan se det som att X är den variabel vi primärt är intresserade av, medan vi samtidigt vill kontrollera för effekt från Z . Om $H_0 : \beta_1 = 0$ förkastas, tolkas det som att det finns variation i responsvariabeln Y som inte kan förklaras av Z och därmed måste bero på variation i X .

Återigen kan vanliga parametriska test inte användas om ε inte kan antas normalfördelade. En första skattning av vår teststatistika görs under den fulla modellen, ekvation (5), och vi får ett observerat värde T_{obs} .

Under H_0 ser den reducerade modellen ut som följer:

$$Y_i = \beta_0 + \beta_2 Z_i + \varepsilon_i. \quad (6)$$

Skattningar $\hat{\beta}_0$ och $\hat{\beta}_2$ fås genom minsta kvadrat-metoden, beskriven i avsnitt 2.2.

Analogt till det univariata fallet kan vi inte känna till de faktiska feltermerna $\varepsilon_i = Y_i - (\beta_0 + \beta_2 Z_i)$ då β_0 och β_2 inte är kända. Residualerna

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_2 Z_i)$$

kan då användas istället. På grund av att residualerna i det multivariata fallet påverkas av en förklarande variabel Z är de beroende och har icke-konstant varians.¹¹ Residualerna e kan därför

¹¹Se sida 342 i Rawlings, Pantula och Dickey (1998).

inte antas utbytbara även om detta gäller för feltermerna ε . Det har dock visat sig att Freedman-Lanes metod ger ett, med stickprovsstorleken, asymptotiskt exakt permutationstest (Anderson och Robinson, 2001), även om residualerna inte strikt är utbytbara.

Vi kan nu generera nya värden på responsvektorn Y genom att lägga till en permuterad residualvektor e^* av e till den skattade regressionslinjen:

$$Y_i^* = \hat{\beta}_0 + \hat{\beta}_2 Z_i + e_i^*.$$

Den permuterade responsvektorn Y^* kan då ses som ett nytt, under H_0 lika troligt, värde på responsvektorn Y . Vi kan därför byta ut Y mot Y^* i den ursprungliga fulla modellen (5) och en ny regression görs för varje permutation e^* av e . Varje permutation ger då ett nytt värde på våra valda teststatistika, vilket möjliggör skapande av en referensfördelning för denna under H_0 .

3.3 Permutationstest med teckenbyte

Permutationstesten beskrivna i föregående avsnitt utgår båda från ett antagande om utbytbarhet för att permutera residualtermerna i vår linjära regressionsmodell. I det fall då residualerna inte kan antas utbytbara kan ett annat permutationstest motiveras, där vi istället använder oss av teckenbyte - se avsnitt 2.3.3. Detta test kan användas om feltermerna har icke-konstant varians, men antas då följa en symmetrisk fördelning.

Förfarandet i ett permutationstest med teckenbyte är likt Freedman-Lanes metod beskriven i föregående avsnitt. För att utföra detta test antas de faktiska feltermerna $\varepsilon_i = Y_i - (\beta_0 + \beta_2 Z_i)$ från den reducerade modellen (6) vara symmetriska och oberoende. Som tidigare kan vi inte känna till ε , för vilka antagandet gäller, utan vi använder istället $e_i = Y_i - (\hat{\beta}_0 - \hat{\beta}_2 Z_i)$. Vi permuterar nu inte e_i som innan, istället permuterar vi dess tecken. Återigen ger detta ett test som endast är asymptotiskt exakt.

3.4 Permutationstest och p -värden

Vi kan beräkna T -statistikan från avsnitt 2.2.1 för varje permutation för att på så sätt bygga upp en referensfördelning för att undersöka vår ursprungsobservation. Med hjälp av denna referensfördelning kan sedan ett p -värde beräknas på olika sätt.

För att nollhypotesen H_0 ska förkastas med lika stor felmarginal i både positiv och negativ riktning kan ett tvåsidigt p -värde beräknas på följande sätt:

$$p_{2min} = \frac{2\min(\#T_{obs} \leq T_{perm}, \#T_{obs} \geq T_{perm}) + 1}{\#perm + 1}. \quad (7)$$

Här betecknar T_{obs} värdet på T -statistikan från vår ursprungsobservation, T_{perm} motsvarande värde för en permutation och $\#perm$ är antalet utförda permutationer. Anledningen till att $+1$ adderas i både täljaren och nämnaren är att T_{obs} inkluderas i referensfördelningen och vilket vi behöver ta hänsyn till i beräkningen av p -värdet. Detta sätt att beräkna p -värdet ger, som tidigare konstaterat, en lika stor felmarginal i båda riktningar, även om referensfördelningen är skev.

Ett alternativt tvåsidigt p -värde är

$$p_{abs} = \frac{(\#|T_{obs}| \geq |T_{perm}|) + 1}{\#perm + 1}.$$

Detta vänsterjusterade p -värde är inte garanterat att förkasta i lika stor grad i båda riktningar, och bör därför inte användas i fall med skeva referensfördelningar. De två beskrivna sätten att beräkna p -värden förväntas dock ge liknande resultat i fall med symmetriska referensfördelningar.

Det är viktigt att notera att p -värdet för ett permutationstest alltid är strikt positivt, eftersom vi inkluderar vår ursprungsobservation i referensfördelningen. Således kan vi aldrig få ett p -värde mindre än $\frac{1}{\#perm+1}$.

4 Metod för simuleringsstudie av permutationstest

För att studera robustheten och effektiviteten hos permutationstest i linjär regression simulerades stora mängder data med och utan linjära samband och med olika sorters feltermar med en stokastisk fördelning. Resultaten från permutationstesten jämfördes därefter med motsvarande resultat från t -testet från avsnitt 2.2.1. Till att börja med studerades simpel linjär regression för att sedan generaliseras till regression med två förklarande variabler. Oavsett antalet förklarande variabler genererades de olika typerna av feltermar på samma sätt. Koden för simuleringarna finns tillgänglig online, se appendix C.

4.1 Modellavvikelser

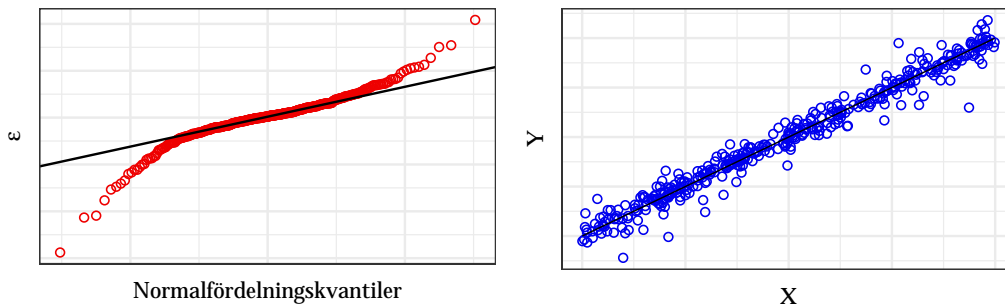
När feltermerna ε_i , där $i = 1, \dots, n$ och n är stickprovsstorleken, inte är normalfördelade faller den teoretiska motiveringen för användandet av t -test som omnämndes i avsnitt 2.2.1. För att generera data med dessa avvikelser simulerades feltermerna med följande fördelningar.

4.1.1 Tunga svansar

För att få en fördelning med tunga svansar (hög kurtosis), det vill säga med fler extrema observationer än vad normalfördelning ger, genererades feltermerna från en standardiserad t -fördelning,

$$\varepsilon_i = \frac{\delta_i}{\sqrt{\frac{\nu}{\nu-2}}},$$

där δ_i är t -fördelad med ν frihetsgrader. Då en t -fördelning med ν frihetsgrader har variansen $\frac{\nu}{\nu-2}$ ger detta att $\text{Var}(\varepsilon_i) = 1$. Genom att ändra $\nu \in \{3, 4, 5, 6, 8, 10, 25, 100\}$ varierades kurtosis i simuleringarna, där ökande ν ger en feltermfördelning som går mot normalfördelning. Ett exempel på feltermar med tunga svansar presenteras i Figur 4. Kurtosis beräknades, för residualerna av den skattade modellen, med funktionen *kurtosis* från R-paketet *moments* (Komsta och Novomestky, 2015) och är en skattning av Pearsons kurtosismått. Som referens har en normalfördelning kurtosis 3 och större värde än detta svarar mot tyngre svansar.



(a) Residualer med tunga svansar mot motsvarande normalfördelingskvantil (QQ-graf). (b) Linjär regression, residualer med tunga svansar.

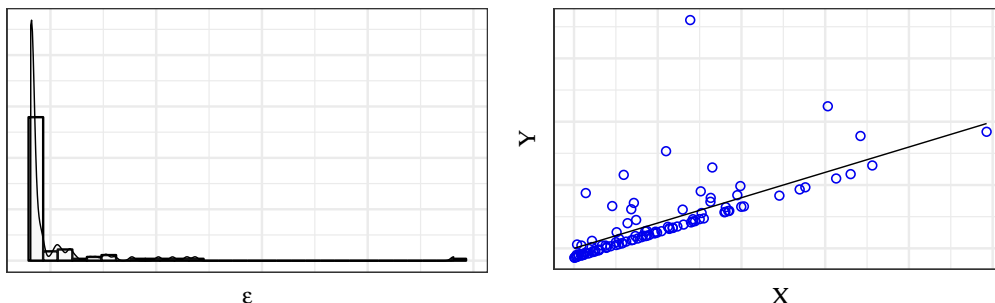
Figur 4: Feltermar med tunga svansar.

4.1.2 Skevfördelade feltermar

För att generera residualer med skev fördelning genererades en variabel från en exponentialfördelning $X_i \sim \text{Exp}(1)$ vilken upphöjdes med en exponent $m \in \{1, 1.5, 2, 2.5, 3, 5, 7, 10\}$ så att $E_i = X_i^m$. Slumpvariabeln E_i standardiserades till medelvärde 0 och varians 1 genom att dra bort medelvärdet av stickprovet, \bar{E} , och dividera med dess standardavvikelse, $SD(E)$,

$$\varepsilon_i = \frac{E_i - \bar{E}}{SD(E)}.$$

Olika skevhet erhöles genom att variera exponenten m , där ett högre värde svarar mot större skevhet. I Figur 5 presenteras ett exempel på data med skevfördelade felstermer. Skevheten, för residualerna av den skattade modellen, beräknades med funktionen *skewness* från R-paketet *moments* (Komsta och Novomestky, 2015) och är en skattning av Pearsons momentkoefficient för skevhet (Pearson's moment coefficient of skewness).

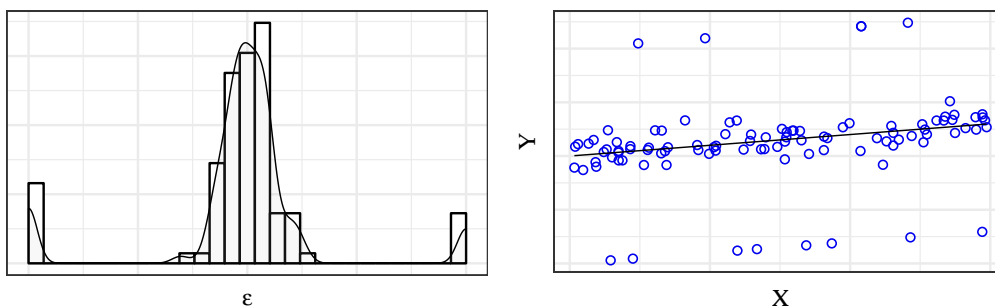


(a) Histogram för skevfördelade residualer. (b) Linjär regression, skevfördelade residualer.

Figur 5: Skevfördelade felstermer.

4.1.3 Utstickare

För att skapa data med utstickare generades observationer med felstermer från en standardiserad normalfördelning. För varje genererad felsterm fanns en viss sannolikhet $\{i/100 : i \in 1, 2, \dots, 10\}$ för ursprungsvärdet att ersättas med ± 10 , där tecknet bestämdes av dess ursprungstecken¹². I Figur 6 presenteras ett exempel på data med utstickare i felstermerna.



(a) Histogram för residualer med utstickare. (b) Linjär regression, residualer med utstickare.

Figur 6: Felstermer med utstickare.

4.1.4 Icke-konstant varians

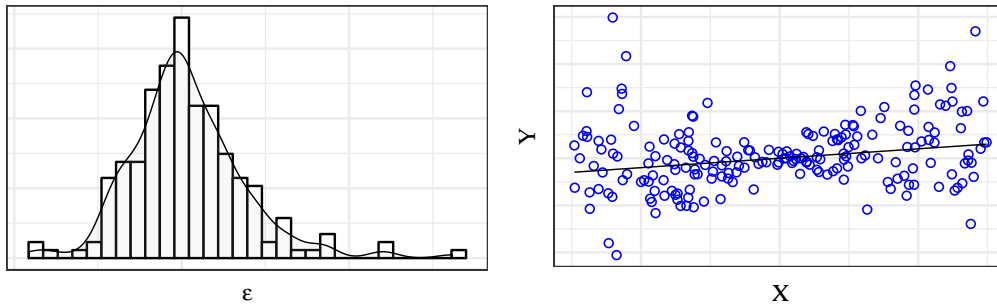
Slutligen, för att skapa felstermer med icke-konstant varians, genererades ε enligt $\varepsilon_i \sim N(0, 1 + v \cdot |x_i|/6)$. Här är $v \in \{1, 2, \dots, 10\}$ en lutningskoefficient som bestämmer hur snabbt variansen ökar med de förklarande variablerna x_i . En visualisering av felstermer med icke-konstant varians presenteras i Figur 7.

4.2 Simpel linjär regression

För simpel linjär regression simulerades den förklarande variabeln X_i från en exponentialfördelning med medelvärde 1 eller en likformig fördelning på intervallet $[0, 3]$, varpå responsvariabeln Y_i genererades enligt

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

¹²Som referens är $P(\varepsilon_i > 10) = 7.619853 \cdot 10^{-24}$ för standardnormalfördelningen.



(a) Histogram för residualer med ickekonstant (b) Linjär regression, residualer med icke-
varians.

Figur 7: Feltermen med icke-konstant varians.

För de olika slagen av feltermen varierades stickprovsstorleken $n \in \{10, 100, 1000\}$, den faktiska lutningen $\beta_1 \in \{0, 0.1, 0.5, 1\}$, samt parametern kopplad till den specifika feltermen. Även $n = 10000$ studerades under projektets gång men togs inte med i slutgiltiga simuleringar då det inte tillförde ytterligare information. Vid datagenerering sattes $\beta_0 = 0$ och den simulerade datan testades under $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Efter att ha gjort detta för enbart det normalfördelningsbaserade t -testet tillämpades även permutationstest för varje simulerad datamängd.

För varje permutation beräknades T -statistikan svarande mot β_1 som i (3) och p -värdet enligt (7). För varje parameterkombination genomfördes 500 repetitioner där en ny datamängd simulerades. Från varje simulerad datamängd anpassades en simpel linjär regressionsmodell och p -värdet beräknades för det normalfördelningsbaserade testet. För varje datamängd genomfördes 999 slumpade permutationer av den skattade residualvektorn från den reducerade modellen som beskrivet i avsnitt 3, med återläggning av varje permutation, varpå 999 permurerade responsvektorer bildades. Med originalobservationen gav detta totalt 1000 permutationer. För varje permurerad responsvektor anpassades en linjär modell för vilken teststatistikan svarande mot β_1 beräknades. Valet av antal permutationer motiveras genom Nichols och Holmes (2001) som nämnt i avsnitt 2.4, samt för att begränsa simuleringstiden. Slutligen undersöktes för alla repetitioner om nollhypotesen förkastades vid 5% signifikansnivå, och medelvärde av detta togs för varje parameterkombination.

Vid icke-konstant varians tillämpades primärt ett slumpmässigt teckenbyte på residualerna, såsom beskrivet i avsnitt 3.3, istället för permutationstestet. Detta motiverades genom att kriterierna för utbytbarhet inte var uppfyllda för det ursprungliga permutationstestet, vilket de var för permutationstestet med teckenbyte.

När testens robusthet studerades, det vill säga hur väl de bevarar signifikansnivån, skapades ett 95% referensintervall kring den valda signifikansnivån, α , enligt

$$\alpha \pm 1.96 \sqrt{\frac{\alpha(1-\alpha)}{N}}. \quad (8)$$

I (8) är N antalet repetitioner per parameterkombination, $\alpha = 0.05$ den valda signifikansnivån och 1.96 kommer från normalapproximation av de 500 binomialfördelade simulering utfallen för varje parameterkombination och 95% sannolikhet. Detta 95%-iga referensintervall beskriver den feltolerans för Typ 1-fel ett test har på grund av simuleringssosäkerhet. Således bör vi ifrågasätta om signifikansnivån bevaras om mer än 5% av våra erhållna förkastningsfrekvenser överstiger detta intervall. Om det underskrids är testet fortfarande giltigt men konservativt.

4.3 Multipel linjär regression

För multipel linjär regression studerades det bivariata fallet där den förklarande variabeln X_i samt den störande variabeln (nuisance variable) Z_i simulerades från en multivariat normalfördelning med medelvärde 0, varians 1 och korrelation 0.5. Responsvariabeln Y_i beräknades enligt

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i, \quad i = 1, \dots, n,$$

där $\beta_2 = 0.5$ och β_1 varierades. Parameterkombinationer samt antalet permutationer och repetitioner hölls till det samma som i det enkla fallet. Permutationsförfarandet var såsom beskrivet i avsnitt 3.2 och ett referensintervall konstruerades enligt (8) på samma sätt som för det enkla fallet.

5 Resultat

I kommande avsnitt presenteras *förkastningsfrekvensen* för olika typer av data, som genererats enligt avsnitt 4.1. I figurerna som presenteras, innebär **t-test** ett normalfördelningsbaserat test med T -statistika, **permTest** ett permutationstest med T -statistika och **permTestSign** ett permutationstest med teckenbyte och T -statistika. Med **R.I** avses ett 95% referensintervall som beskriver den feltolerans för Typ 1-fel ett test har på grund av simuleringssosäkerhet. Förkastningsfrekvensen är andelen förkastade nollhypoteser $H_0 : \beta_1 = 0$ med alternativhypotesen $H_1 : \beta_1 \neq 0$. I de fall där det inte finns något samband, tyder en förkastningsfrekvens inom eller under referensintervallet på att testet är giltigt och där en frekvens under intervallet svarar mot ett konservativt test. I de fall där $\beta_1 \neq 0$ ökar styrkan med förkastningsfrekvensen. Inom samtliga figurer är förkastningsfrekvensen baserad på 500 simuleringar av data som i sin tur permuterats 999 gånger.

5.1 Simpel linjär regression

I kommande avsnitt presenteras resultaten för olika datatyper inom simpel linjär regression, för olika stickprovsstorlekar n och för fallen $\beta_1 = 0$ och $\beta_1 \neq 0$. Notera att endast simuleringarna för $n = 10, 1000$ presenteras för att spara plats då dessa bedömts vara mest relevanta. Samtliga resultat presenteras inte heller i detta avsnitt utan läsaren hänvisas till appendix för ytterligare resultat.

5.1.1 Tunga svansar

För feltermers vars fördelning har tunga svansar, undersöktes skillnaden mellan data när X genererades från en exponentialfördelning respektive en likformig fördelning mellan 0 och 3. För feltermersfördelning med tunga svansar, bevarade t -test och permutationstest signifikansnivån då $\beta_1 = 0$. Med $\beta_1 \neq 0$ presterade de båda testen likartat med avseende på styrka, det vill säga att de korrekt identifierar när det finns ett samband i datan. Dessa resultat styrks av motsvarande simuleringar för det multivariata fallet, se avsnitt 5.2.1.

Samtliga resultat för feltermers med tunga svansar presenteras i appendix A.1.1.

5.1.2 Skevfördelade feltermers

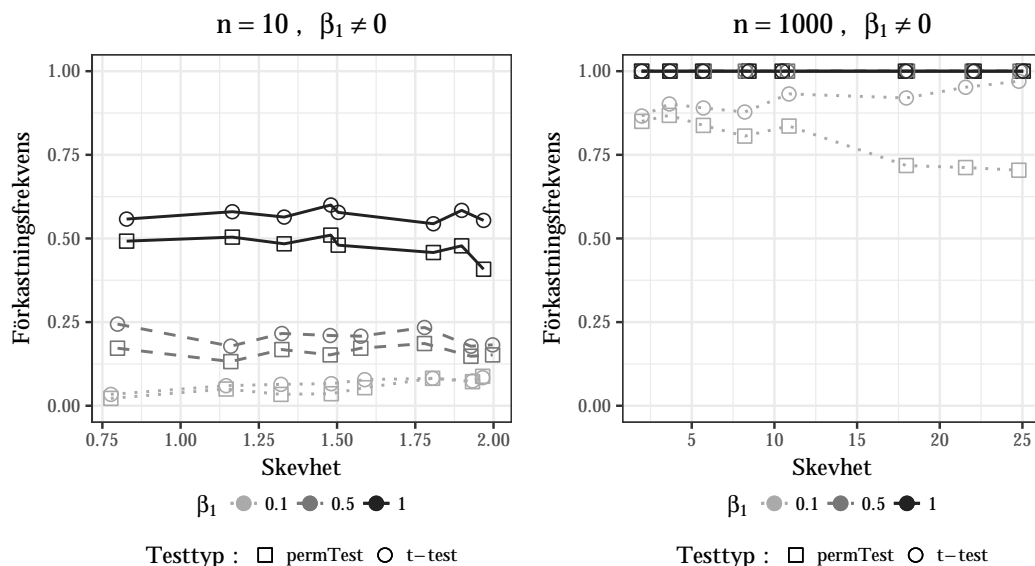
För denna typ av avvikelser undersöktes som komplement till simuleringar med $\beta_1 > 0$ även fallet $\beta_1 < 0$, i kombination med X från en exponentialfördelning med väntevärde 1. Utöver det simulerades förkastningsfrekvensen med skevfördelade residualer och X likformigt fördelad mellan 0 och 3.

5.1.2.1 Med X från en exponentialfördelning

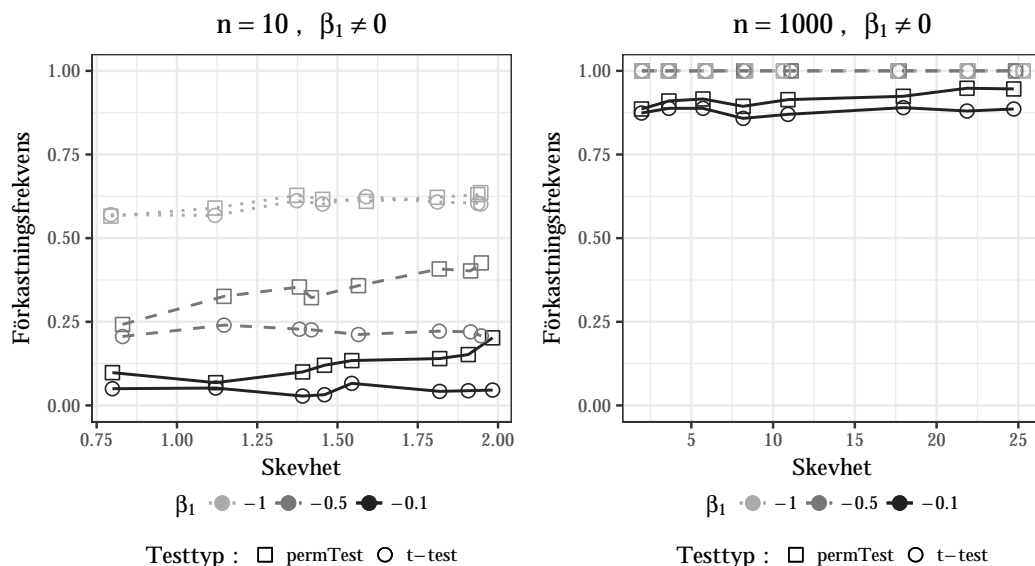
När X genererades från en exponentialfördelning noterades att både det normalfördelningsbaserade testet och permutationstestet presterade inom referensintervallet för $\beta_1 = 0$. Grafen för detta fall presenteras i appendix A.1.2. I Figur 8 visas förkastningsfrekvensen för respektive test som en funktion av ökande skevhet för $\beta_1 > 0$. Vi ser att för $n = 10$ och $n = 1000$ upptäcker t -testet det

faktiska sambandet i större utsträckning än permutationstestet. I fallet $n = 1000$ visar figuren att förkastningsfrekvensen ökar något med ökande skevhet för t -testet.

I Figur 9 visas motsvarande simuleringar när $\beta_1 < 0$. Här medför ökad skevhet istället något högre förkastningsfrekvens för permutationstest och lägre förkastningsfrekvens för t -test.



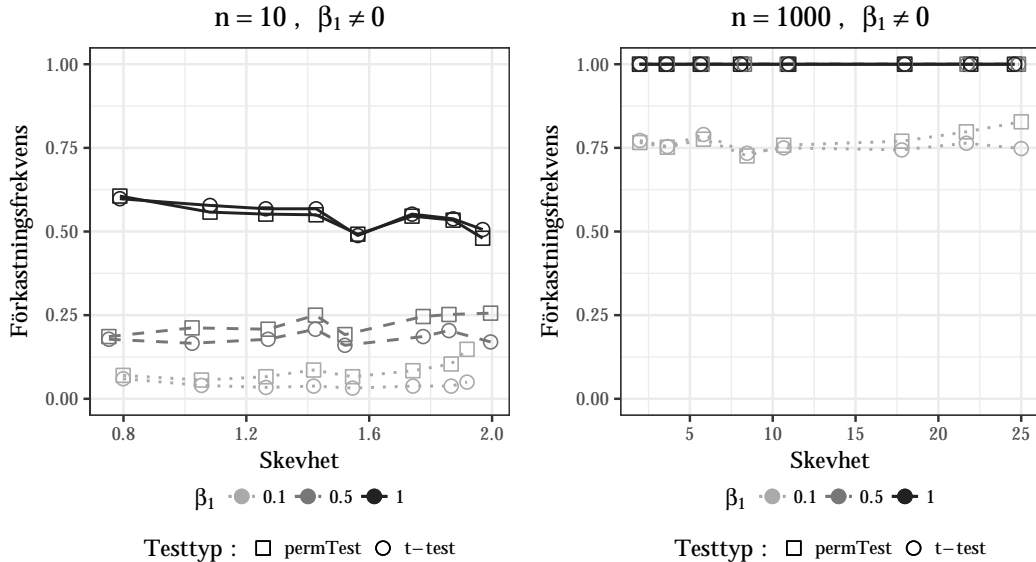
Figur 8: Förkastningsfrekvens som funktion av ökande skevhet för två olika stickprovsstorlekar i fallet då $\beta_1 > 0$. Den förklarande variabeln genererades från en exponentialfördelning med väntevärde 1 och feltermerna från en skev fördelning.



Figur 9: Förkastningsfrekvens som funktion av ökande skevhet för två olika stickprovsstorlekar i fallet då $\beta_1 < 0$. Den förklarande variabeln genererades från en exponentialfördelning med väntevärde 1 och feltermerna från en skev fördelning.

5.1.2.2 Med X från en likformig fördelning

Med den förklarande variabeln från en likformig fördelning mellan 0 och 3 ser vi i Figur 10 att förkastningsfrekvensen för de båda testen är mycket lik den i det föregående fallet med negativa β_1 -värden. Både t -test och permutationstest visar på förkastningsfrekvenser inom det satta referensintervallet och de är därmed giltiga, se Figur A.4 i appendix A.1.2.



Figur 10: Förkastningsfrekvens som funktion av ökande skevhet för två olika stickprovsstorlekar i fallet då $\beta_1 \neq 0$. Den förklarande variabeln genererades från en likformig fördelning på $[0, 3]$ och feltermerna från en skev fördelning.

5.1.3 Utstickare

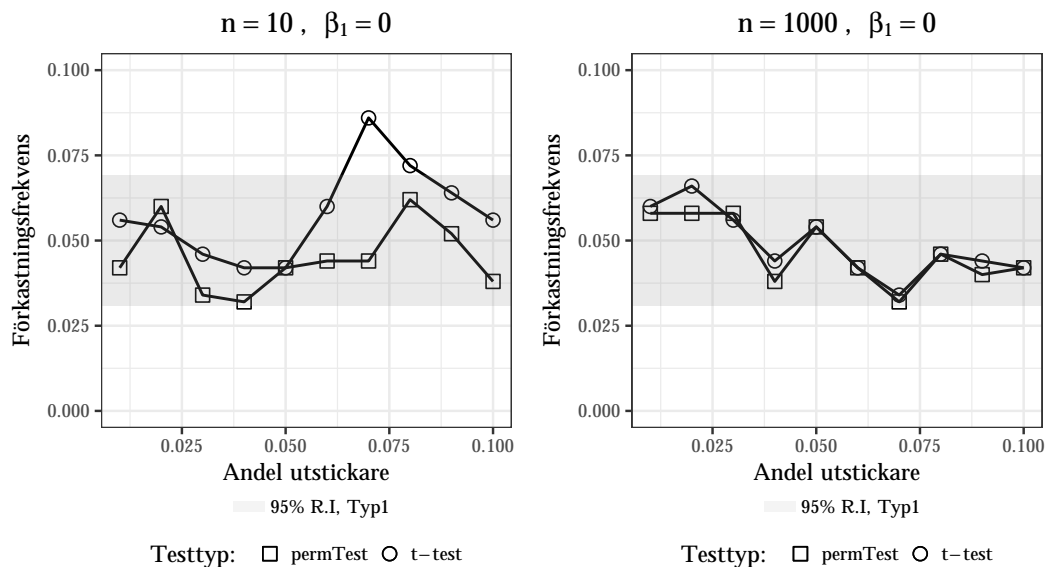
I Figur 11 presenteras förkastningsfrekvensen för de två testtyperna som funktion av andel utstickare för fallet $\beta_1 = 0$. För $n = 10$ ser vi att t -testet ett flertal gånger ger en förkastningsfrekvens högre än referensintervallet vilket tyder på att t -testet är icke-giltigt. Det kan motivera användning av permutationstest för mindre stickprovsstorlekar när datan innehåller utstickare i Y -led. För fallet $n = 1000$ däremot befinner sig båda testen inom referensintervallet. Viktigt att notera är att eftersom referensintervallet är på 95% är det 5% sannolikhet att ett giltigt test hamnar utanför referensintervallet. För fallet $\beta_1 \neq 0$ presterar testen likartat med avseende på effektivitet, och för kompletta resultat hänvisas läsaren till appendix A.1.3, Figur A.6.

5.1.4 Icke-konstant varians

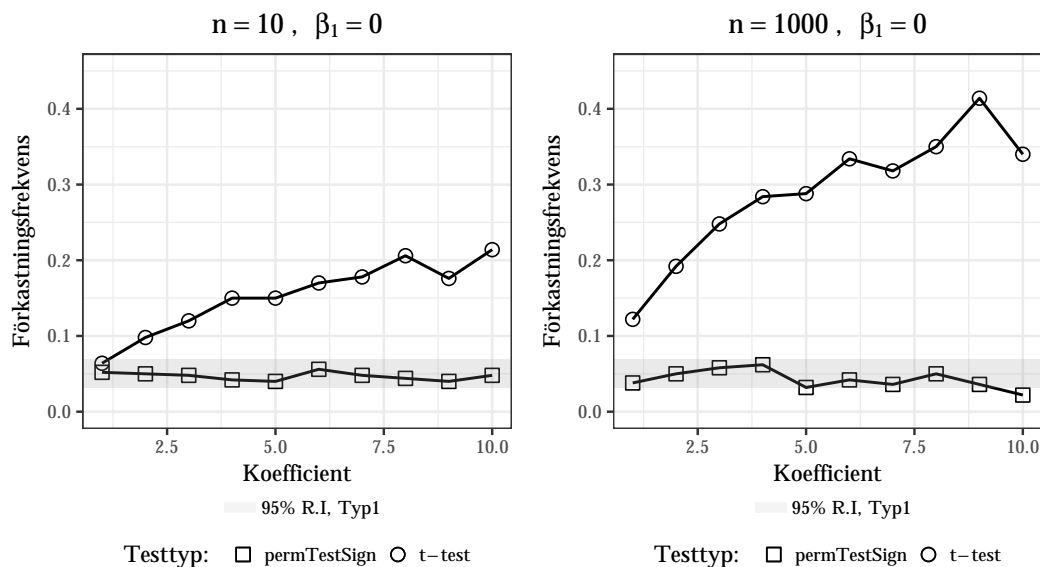
I Figur 12 presenteras förkastningsfrekvensen som funktion av spridning i feltermens varians, i fallet $\beta_1 = 0$. Koefficienten som bestämmer detta förklaras i avsnitt 4.1.4. Vi ser att det normalfördelningsbaserade testet har en för hög förkastningsgrad för båda stickprovsstorlekarna och är därmed icke-giltigt. Permutationstestet förkastar inom referensintervallet för de båda stickprovsstorlekarna, och bevarar således signifikansnivån.

Figur 13 visar förkastningsfrekvensen i fallet $\beta_1 \neq 0$ för stickprovsstorlekarna $n = 10$ respektive $n = 1000$. Genom att basera slutsatser på dessa figurer kan den felaktiga slutsatsen dras att t -test presterar bra för denna feltermstyp då det har en hög styrka, men så är inte fallet. Eftersom föregående simuleringarna visar på att t -testet inte bevarar signifikansnivån för denna typ av felterm, är testet inte giltigt och bör därför inte användas, trots den högre styrkan.

Viktigt att notera är att permutationstestet som använts för avvikelser med icke-konstant varians, är med teckenbyte - se avsnitt 3.3.



Figur 11: Förkastningsfrekvens som funktion av ökande andel utstickare för två stickprovsstorlekar i fallet då $\beta_1 = 0$. Den förklarande variabeln genererades från exponentialfördelning med väntevärde 1 och feltermerna från en normalfördelning med en given andel utstickare.

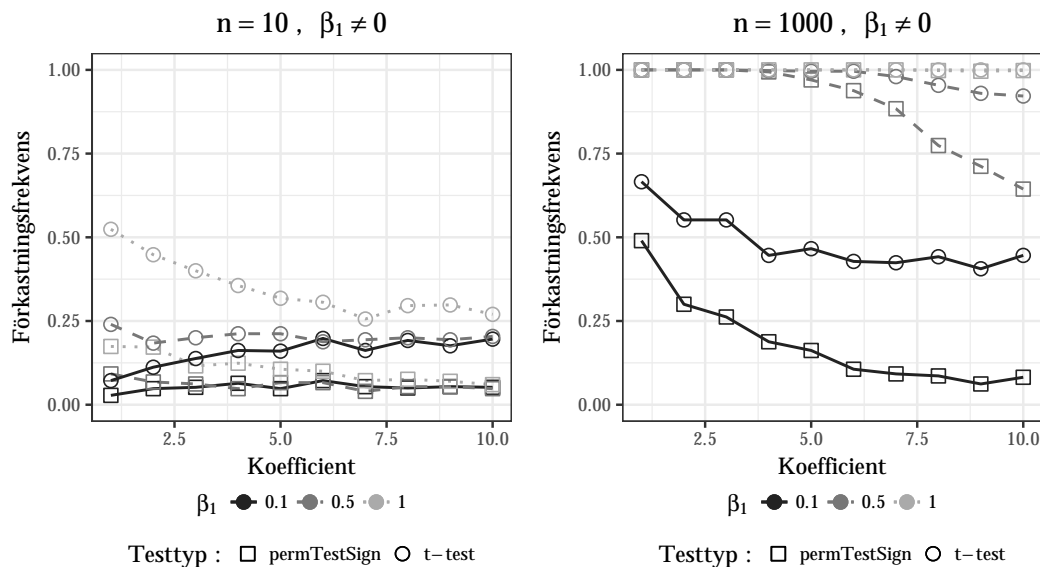


Figur 12: Förkastningsfrekvens som funktion av ökande lutningskoefficient för två stickprovsstorlekar i fallet då $\beta_1 = 0$. Den förklarande variabeln genererades från en exponentialfördelning med väntevärde 1 och feltermerna från en normalfördelning med varians beroende på den förklarande variabeln.

5.2 Multipel linjär regression

5.2.1 Tunga svansar

I appendix A.2.1, Figur A.7, presenteras förkastningsfrekvens som funktion av svanstyngd (kurtosis) i fallet då $\beta_1 = 0$. Vi ser att båda testtyperna bevarar signifikansnivån, eftersom de håller en förkastningsfrekvens inom referensintervallet. Figur A.8 visar att de båda testtyperna är lika effektiva. Dessa resultat styrker alltså slutsatserna från simuleringar i det univariata fallet.



Figur 13: Förkastningsfrekvens som funktion av ökande lutningskoefficient för två stickprovsstorlekar i fallet då $\beta_1 \neq 0$. Den förklarande variabeln genererades från en exponentialfördelning med väntevärde 1 och feltermerna från en normalfördelning med varians beroende på den förklarande variabeln.

5.2.2 Skevfördelade feltermer

I Figur A.9, appendix A.2.2, ser vi att förkastningsfrekvensen för t -testet i fallet med skevfördelade feltermer och en multivariat modell i ett fall når över referensintervallet. Detta kan vara en indikation på att testet inte är giltigt, men eftersom referensintervallet är valt till 95% är det 5% risk att ett giltigt test hamnar utanför intervallet. Därmed är det rimligt att tro att både t -testet och permutationstestet, som håller en förkastningsfrekvens inom referensintervallet, är giltiga.

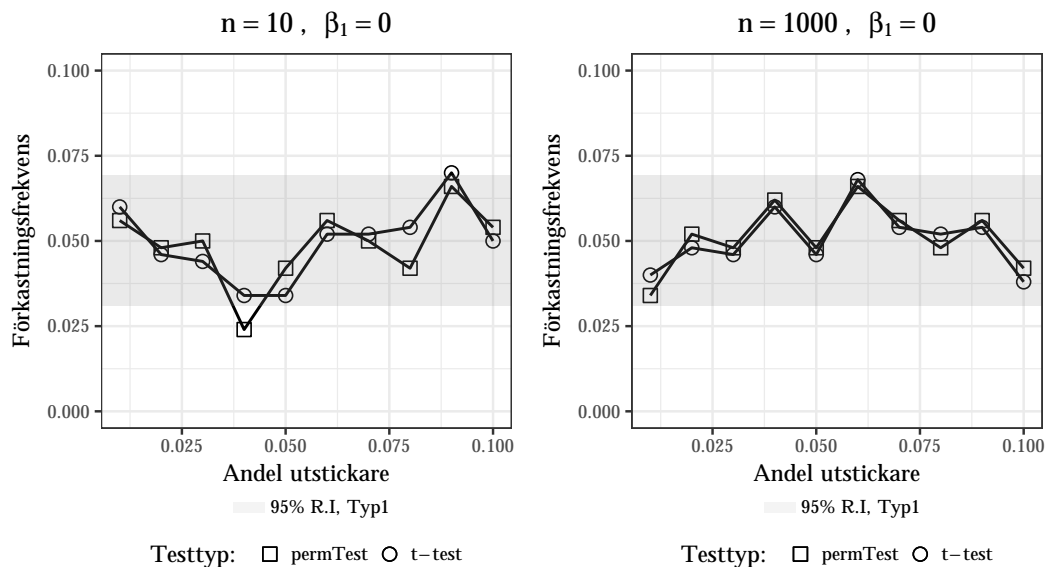
Simuleringarna visade även att de båda testtyperna presterar mycket likartat med avseende på effektivitet, se Figur A.10 i appendix A.2.2.

5.2.3 Utstickare

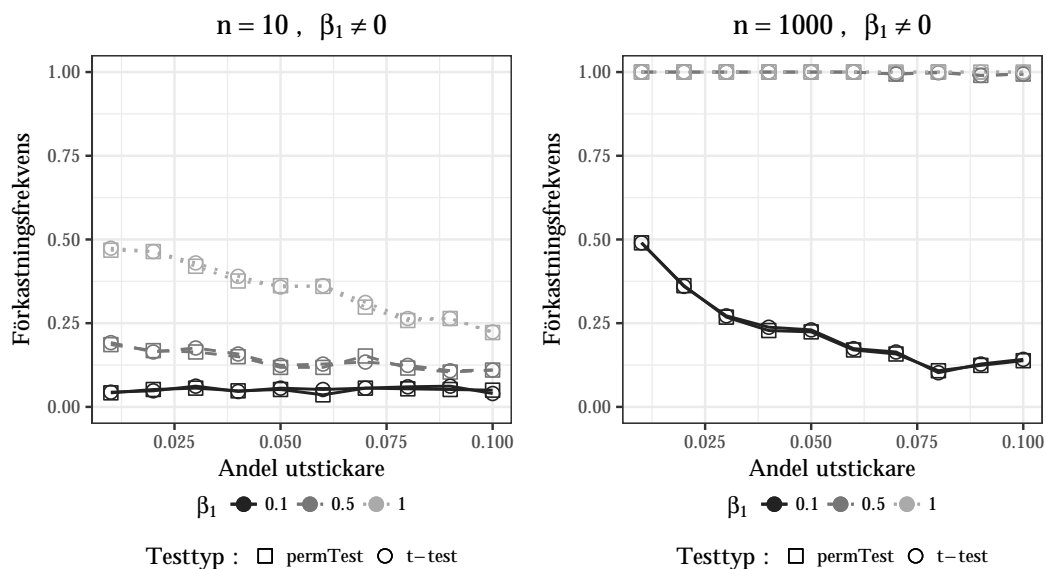
Figur 14 visar att båda testtyperna är giltiga, eftersom de inte överskrider referensintervallet för signifikansnivån. Normalfördelningsbaserade testet har genomgående en förkastningsfrekvens under signifikansnivån och det är därmed konservativt. Figur 15 visar att permutationstest och normalfördelningsbaserat test presterar likvärdigt med avseende på effektivitet. Båda testtyperna får minskad styrka för ökande andel utstickare i feltermen.

5.2.4 Icke-konstant varians

Resultat från simuleringar med en multivariat modell och icke-konstant varians i feltermerna visas i appendix A.2.4. Här visas förkastningsfrekvensen för både permutationstest som kräver utbytbarhet och permutationstest med teckenbyte. Det vanliga permutationstestet inte är teoretiskt motiverat och därför är det förväntat att inte prestera bra. Resultaten visar som väntat att permutationstest med teckenbyte är giltigt, medan det vanliga permutationstestet förkastar i för hög grad och är därav icke-giltigt. Ej presenterat i figurerna är t -testet som inte heller i det multivariata fallet bevarade signifikansnivån.



Figur 14: Förläsningsfrekvens som funktion av ökande andel av utstickare för två stickprovsstorlekar i fallet då $\beta_1 = 0$. De förklarande variablerna genererades från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5 och feltermerna enligt en normalfördelning med en given andel utstickare.



Figur 15: Förläsningsfrekvens som funktion av ökande andel av utstickare för två stickprovsstorlekar i fallet då $\beta_1 \neq 0$. De förklarande variablerna genererades från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5 och feltermerna enligt en normalfördelning med en given andel utstickare.

6 Diskussion

6.1 Resultatdiskussion

Simuleringarna i denna studie visar att permutationstestet som väntat är giltigt även med icke-normalfördelade feltermerna. Även t -testet, vars kriterier för att vara giltigt inte är uppfyllda, verkar vara robust för många av de undersökta typerna av feltermers avvikelser. Ett fall där detta dock inte gäller och t -testet är icke-giltigt är då feltermerna har icke-konstant varians. Vad som är viktigt att beakta är att det i det här arbetet endast har studerats en begränsad mängd feltermerna som bryter mot t -testets förutsättningar, således kan inga generella slutsatser dras om vilka avvikelser som t -testet bevarar signifikansnivån. Detta kan vara en anledning att föredra permutationstest framför t -test.

För feltermerna med icke-konstant varians överskrider t -testet signifikansnivån. Detsamma gäller för permutationstest där vi permuterar residualerna. Däremot bevarar permutationstestet med teckenbyte signifikansnivån för dessa feltermerna och är således att föredra.

Simuleringsresultaten från fallet med X från en exponentialfördelning, $\beta_1 > 0$ och feltermerna från en skev fördelning, se Figur 8, visar på att permutationstestet med ökande skevhet är sämre på att hitta faktiska samband när dessa finns. Däremot fortsätter t -testet prestera på samma nivå, vilket är oväntat. I fallet då X istället följer en likformig fördelning som i Figur 10 presterar de båda testen likartat. Från detta dras slutsatsen att fördelningen i X har en påverkan på testens styrka specifikt för fallet då feltermerna är skevfördelade. Att även tecknet på β_1 påverkar hur permutationstestet presterar, i det specifika fallet då X är exponentialfördelad och residualerna följer en skev fördelning, styrks av resultaten då $\beta_1 < 0$, se Figur 9. För att sammanfatta bevarar båda testen signifikansnivån men styrkan beror på hur den förklarande variabeln är fördelad och om vi har ett positivt eller negativt samband.

Det är relevant att ta fram bra metoder för att undersöka data med denna typ av avvikelse i residualernas fördelning. Ett exempel för detta finns i Andelin m. fl. (2016), en studie av precision av en glukosmätare. I Figur 2 i Andelin m. fl. (2016) ser vi ett exempel på skevfördelade residualer i en modell som visserligen är icke-linjär, men som ändå visar på en situation som styrker att skevhet rimligtvis är ett problem som är viktigt att kunna hantera.

För feltermerna med utstickare indikerar simuleringarna överlag på att t -testet bevarar signifikansnivån, med undantag för fallet med en simpel modell och $n = 10$. Eftersom signifikansnivån inte bevaras för små stickprovsstorlekar, kan det vara bra att undersöka ifall kraftigare utstickare eller andra typer av utstickare kan ha inverkan på t -testets robusthet.

Resultaten från fallet då feltermerna följer en fördelning med tunga svansar visar att det normalfördelningsbaserade t -testen fungerar bra med denna typ av avvikelse.

Simuleringarna med multivariat modell styrker överlag slutsatserna från simuleringar med en simpel modell, vilket tyder på att korrelation mellan förklarande variabler inte påverkar testens prestation. Vad som skulle kunna vara intressant att studera är påverkan från ännu kraftigare korrelation. Att permutationstest endast är asymptotiskt exakt för fallet med en multivariat modell verkar inte vara något problem för att motivera användning av testet.

Genom arbetets gång har det visat sig att resultatet för t -testet för den minsta stickprovsstorleken $n = 10$ varierar mer med olika simuleringar än för större stickprovsstorlekar. Detta styrker slutsatsen att detta test bör användas med försiktighet då feltermerna inte kan antas normalfördelade, särskilt då stickprovsstorleken är liten. Därtill är det värt att nämna att om feltermerna är både icke-symmetriska och har icke-konstant varians är inget av de permutationstest vi diskuterar i denna rapport giltiga.

Sammanfattningsvis ser vi att permutationstest presterar bättre eller likvärdigt med t -testet, men aldrig sämre för de fall vi undersökt i denna rapport.

6.2 Allmän diskussion

6.2.1 R^2 -statistika i permutationstest

I litteratur som behandlar permutationstest i linjära modeller används ofta andra statistikor än T -statistikan för att utvärdera en skattad regressionskoefficient. Ett exempel på en sådan statistika

är förklaringsgraden R^2 som är ett mått på hur väl variansen i responsvariabeln förklaras av den linjära anpassningen. I dessa fall konstrueras ett p -värde genom att mäta andelen permutationer av residualtermen som ger högre R^2 -värde än den ursprungliga skattningen. R^2 -statistikan ger samma resultat i termer av förkastningsfrekvens som t -statistikan med symmetriskt p -värde då båda utgår från en symmetrisk referensfördelning.

I detta arbete undersöktes även permutationstest med R^2 -statistikan parallellt med T -statistikan, de båda statistikorna gav mycket liknande resultat. För att lättare kunna jämföra permutations-testet med det parametriska t -testet, presenteras endast T -statistikan.

6.2.2 Konfidensintervall via permutationstest

I praktiken är det ofta önskvärt att tillsammans med ett p -värde för en parameterskattning $\hat{\beta}$ även presentera ett konfidensintervall för hur säker denna skattning är. Om residualtermen i vår linjära modell inte kan antas normalfördelad gäller det även att vanliga konfidensintervall som grundas på att responsvariabeln är normalfördelad inte kan användas. För att lösa detta problem finns metoder som baseras på permutationsförfaranden framtagna för att bygga ett konfidensintervall även i de fall då feltermen - och därmed responsvariabeln - i en linjär regressionsmodell inte uppfyller ett normalfördelningsantagande. Konfidensintervallet konstrueras genom att hitta alla parametervärden b för vilka nollhypotesen $H_0 : \beta = b$ inte kan förkastas. Värt att notera är att konfidensintervall beräknade via permutationstest, precis som de hypotestest via permutationer som studerats i denna rapport, ofta är beräkningstungt.

Simuleringsbaserade studier av konfidensintervall som tas fram via permutationer i jämförelse med intervall som baseras på ett normalfördelningsantagande är ett möjligt område för fortsatta studier.

6.2.3 Befintliga implementationer av permutationstest

Som slutanvändare kan det vara av intresse att veta vad för befintliga implementationer som finns lättillgängliga. I språket R finns det i dagsläget primärt tre paket för permutationstest. För permutationstest i linjär regression finns det relativt nya paketet *permuco* (Frossard och Renaud, 2018) som inkluderar implementation av flertalet permutationstest för linjära modeller, däribland Freedman-Lanes metod som använts i detta arbete. Ett annat alternativ för permutationstest i linjär regression är *lmPerm* (Wheeler och Torchiano, 2016) och för permutationstest i allmänhet paketet *coin* (Hothorn m. fl., 2017).

6.3 Etiska aspekter

Vi finner i detta projekt inga problematiska etiska aspekter. Våra simuleringsresultat är reproducerbara genom användning av frö (seed). I övrigt är vi tydliga med att denna metod inte utvecklats av oss utan refererar till dess ursprungliga upphovspersoner.

Referenser

- Aickin, Mikel (2010). “Invalid Permutation Tests”. English. I: *International Journal of Mathematics and Mathematical Sciences* 2010, s. 1–10.
- Andelin, M. m.fl. (2016). “Assessing the Accuracy of Continuous Glucose Monitoring (CGM) Calibrated With Capillary Values Using Capillary or Venous Glucose Levels as a Reference”. I: *Journal of Diabetes Science and Technology*. URL: <http://journals.sagepub.com/doi/abs/10.1177/1932296815626724>.
- Anderson, Marti J. och Pierre Legendre (1999). “An Empirical Comparison of Permutation Methods for Tests of Partial Regression Coefficients in a Linear Model”. I: *Journal of Statistical Computation and Simulation*. URL: <https://www.tandfonline.com/doi/abs/10.1080/00949659908811936>.
- Anderson, Marti J. och John Robinson (2001). “Permutation Tests for Linear Models”. I: *Australian & New Zealand Journal of Statistics* 43.1, s. 75–88. DOI: 10.1111/1467-842X.00156. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-842X.00156>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-842X.00156>.
- Demmel, James W. (1997). *Applied Numerical Linear Algebra. First Edition*. Siam.
- Freedman, David och David Lane (1983). “A Nonstochastic Interpretation of Reported Significance Levels”. I: *Journal of Business & Economic Statistics* 1.4, s. 292–298. ISSN: 07350015. URL: <http://www.jstor.org/stable/1391660>.
- Frossard, Jaromil och Olivier Renaud (2018). *permuco: Permutation Tests for Regression, (Repeated Measures) ANOVA/ANCOVA and Comparison of Signals*. R package version 1.0.1. URL: <http://CRAN.R-project.org/package=permuco>.
- Good, Phillip (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Second Edition*. Springer.
- Hothorn, Torsten m.fl. (2017). *coin: Conditional Inference Procedures in a Permutation Test Framework*. R package version 1.2-2. URL: <http://CRAN.R-project.org/package=coin>.
- Komsta, Lukasz och Frederick Novomestky (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14. URL: <https://CRAN.R-project.org/package=moments>.
- Nichols, Thomas E. och Andrew P. Holmes (2001). “Nonparametric permutation tests for functional neuroimaging: A primer with examples”. I: *Human Brain Mapping* 15.1, s. 1–25. DOI: 10.1002/hbm.1058. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.1058>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.1058>.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rawlings, John O., Sastry G. Pantula och David A. Dickey (1998). *Applied Regression Analysis: A Research Tool. Second Edition*. Springer.
- Rice, John A. (2007). *Mathematical Statistics and Data Analysis. Third Edition*. Brooks Cole.
- Tancredi, Mauro m.fl. (2017). “Prevalence of primary aldosteronism among patients with type 2 diabetes”. I: *Clinical Endocrinology* 87.3, s. 233–241. DOI: 10.1111/cen.13370. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cen.13370>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cen.13370>.
- Venables, W. N. och B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: <http://www.stats.ox.ac.uk/pub/MASS4>.
- Wheeler, Bob och Marco Torchiano (2016). *lmPerm: Permutation Tests for Linear Models*. R package version 2.1.0. URL: <http://CRAN.R-project.org/package=lmPerm>.
- Wickham, Hadley (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-0-387-98140-6. URL: <http://ggplot2.org>.
- Winkler, Anderson M. m.fl. (2014). “Permutation Inference for the General Linear Model”. I: *Neuroimage*. URL: <https://www.sciencedirect.com/science/article/pii/S1053811914000913?via%5C%3Dihub>.

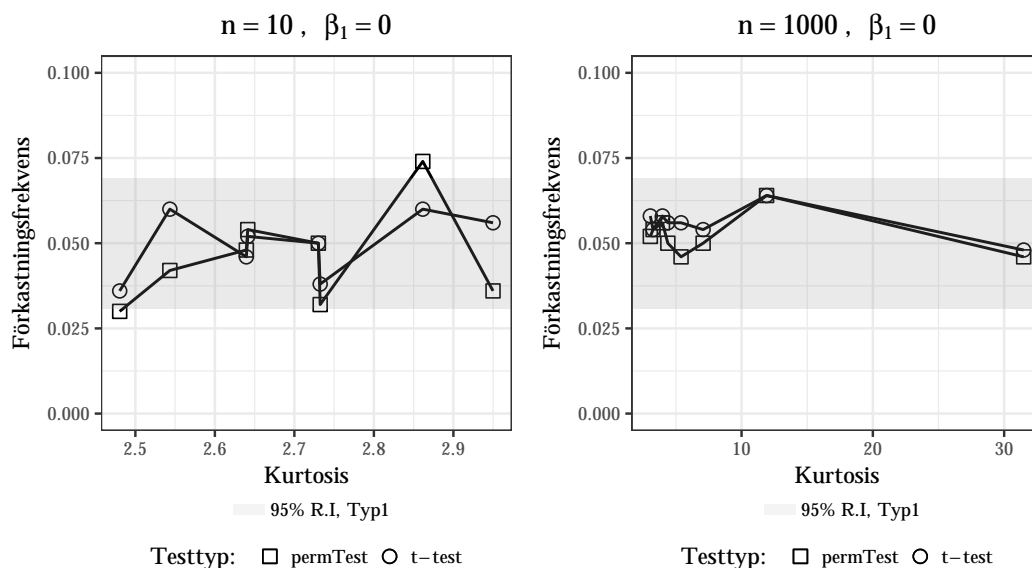
A Övriga resultat

Resultat av simuleringsstudien som av praktiska skäl inte presenterats i rapporten huvuddel, presenteras i sin fullständighet i detta appendix.

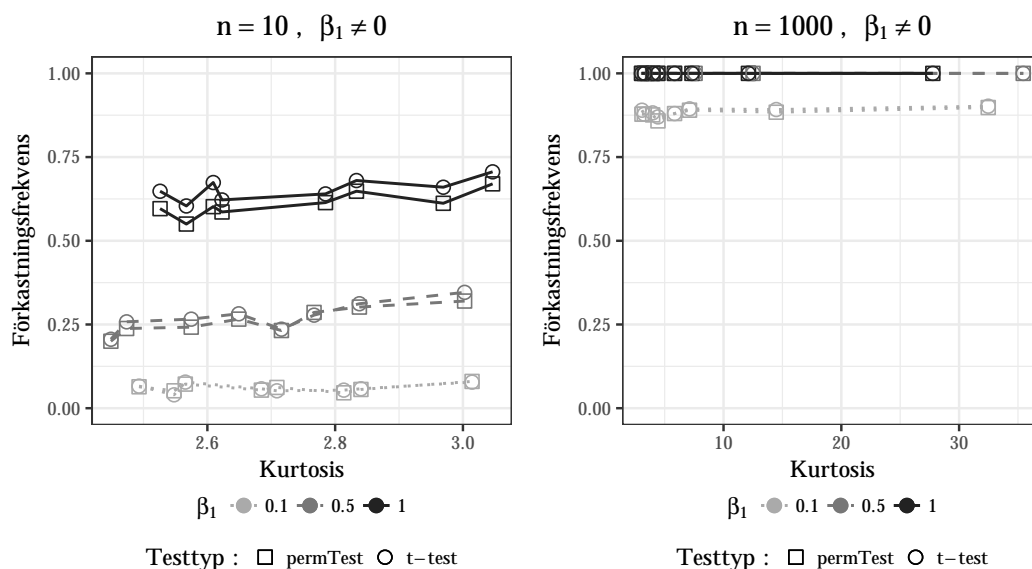
A.1 Simpel linjär regression

A.1.1 Tunga svansar

A.1.1.1 Med X från en exponentialfördelning



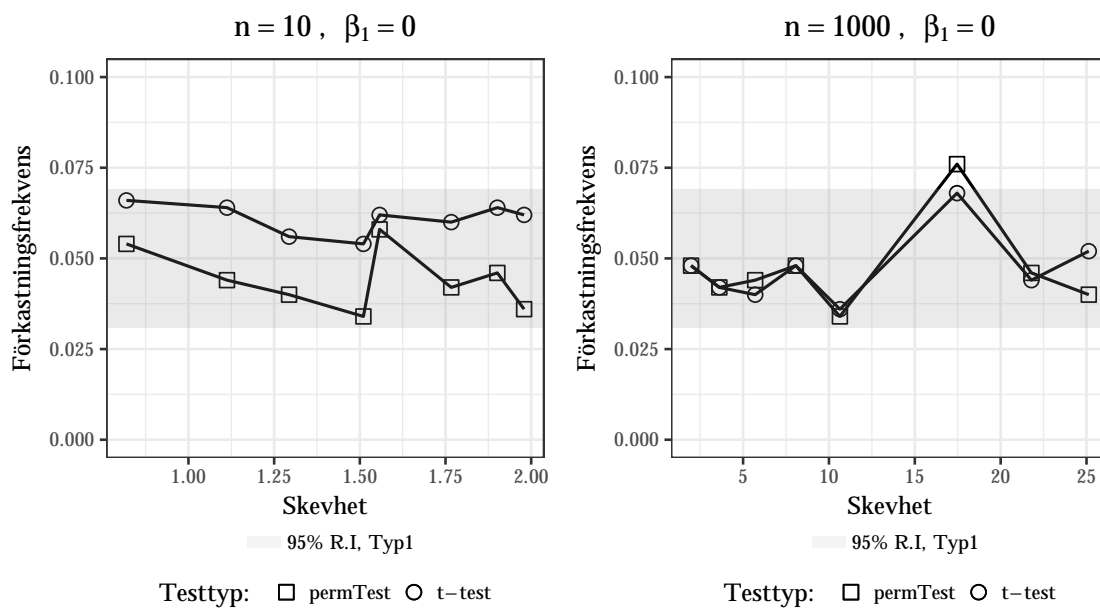
Figur A.1: Förkastningsfrekvens som funktion av ökande kurtosis för två olika stickprovsstorlekar i fallet då $\beta_1 = 0$. Den förklarande variabeln är genererad från en exponentialfördelning med väntevärde 1. De båda testen presterar likartat.



Figur A.2: Förkastningsfrekvens som funktion av ökande kurtosis för två olika stickprovsstorlekar i fallet då $\beta_1 \neq 0$. Den förklarande variabeln är genererad från en exponentialfördelning med väntevärde 1. De båda testen presterar likartat.

A.1.2 Skevfördelade feltermar

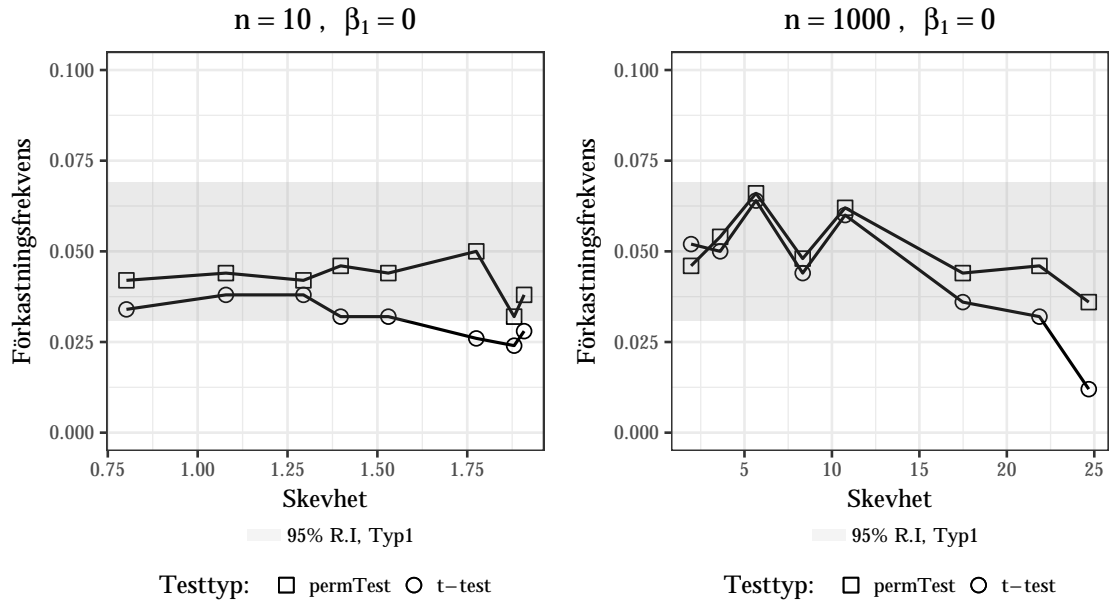
A.1.2.1 Med X från en exponentialfördelning



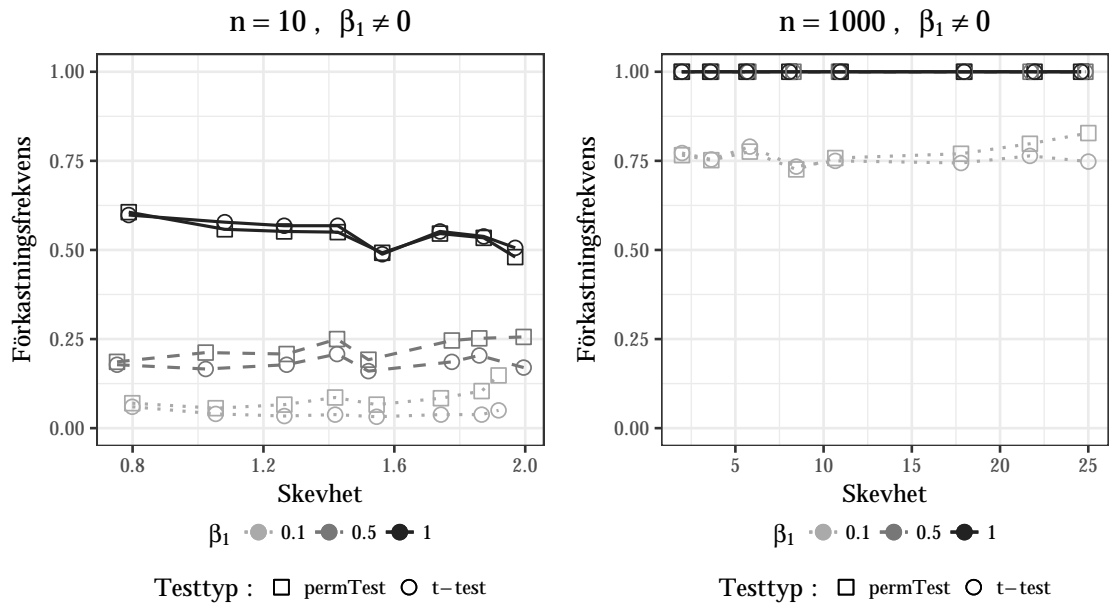
Figur A.3: Förkastningsfrekvens som funktion av ökande skevhet för två olika stickprovsstorlekar i fallet då $\beta_1 = 0$. Den förklarande variabeln är genererad från en exponentialfördelning med väntevärde 1. De båda testen presterar likartat.

Motsvarande graf för $\beta_1 \neq 0$ presenteras i avsnitt 5.1.2.

A.1.2.2 Med X från en likformig fördelning

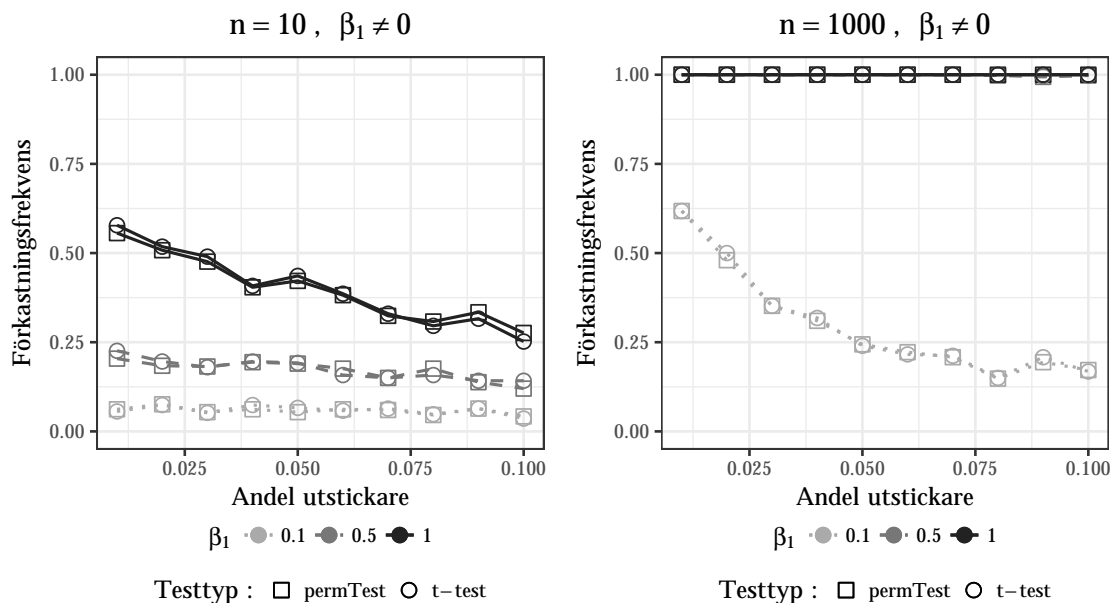


Figur A.4: Förkastningsfrekvens som funktion av ökande skevhet för två olika stickprovsstorlekar i fallet då $\beta_1 = 0$. Den förklarande variabeln är genererad från en likformig fördelning på $[0, 3]$. De båda testen presterar likartat.



Figur A.5: Förkastningsfrekvens som funktion av ökande skevhet för två olika stickprovsstorlekar i fallet då $\beta_1 \neq 0$. Den förklarande variabeln är genererad från en likformig fördelning på $[0, 3]$. De båda testen presterar likartat.

A.1.3 Utstickare



Figur A.6: Förkastningsfrekvens som funktion av ökande andel utstickare för två stickprovsstorlekar i fallet då $\beta_1 \neq 0$. Den förklarande variabeln är genererad från en exponentialfördelning med väntevärde 1. De båda testen presterar likartat.

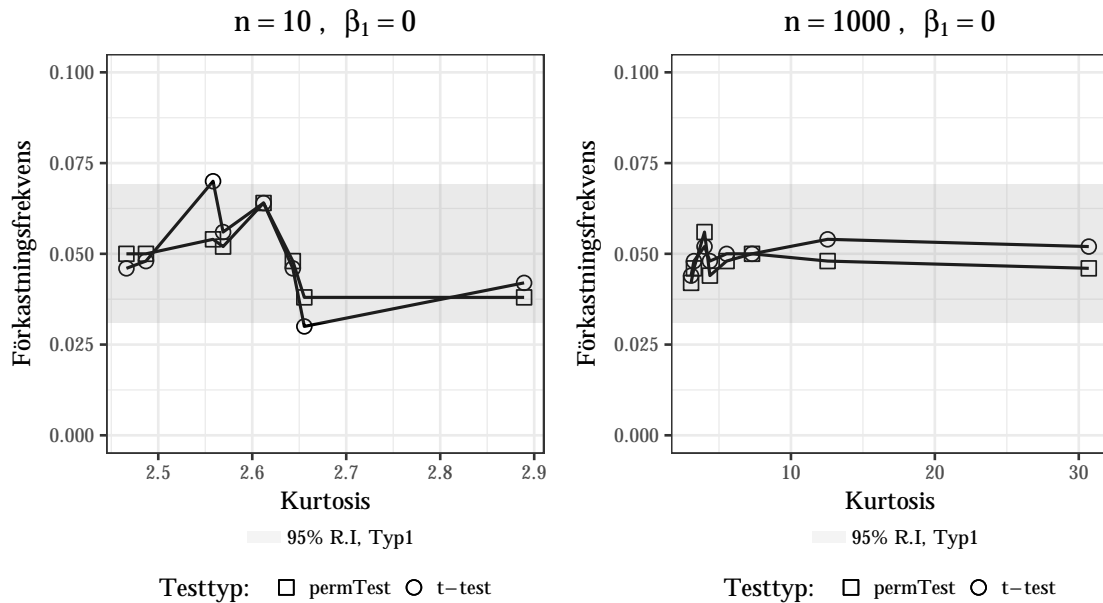
Motsvarande graf för $\beta_1 = 0$ presenteras i avsnitt 5.1.3

A.1.4 Icke-konstant varians

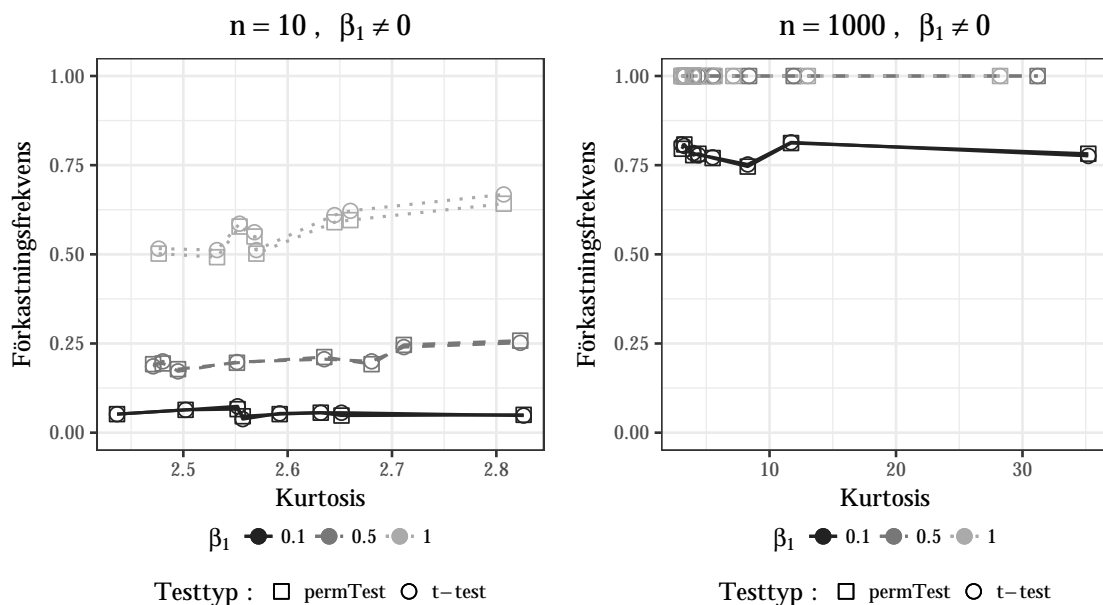
Samtliga grafer presenteras i avsnitt 5.1.4

A.2 Multipel linjär regression

A.2.1 Tunga svansar

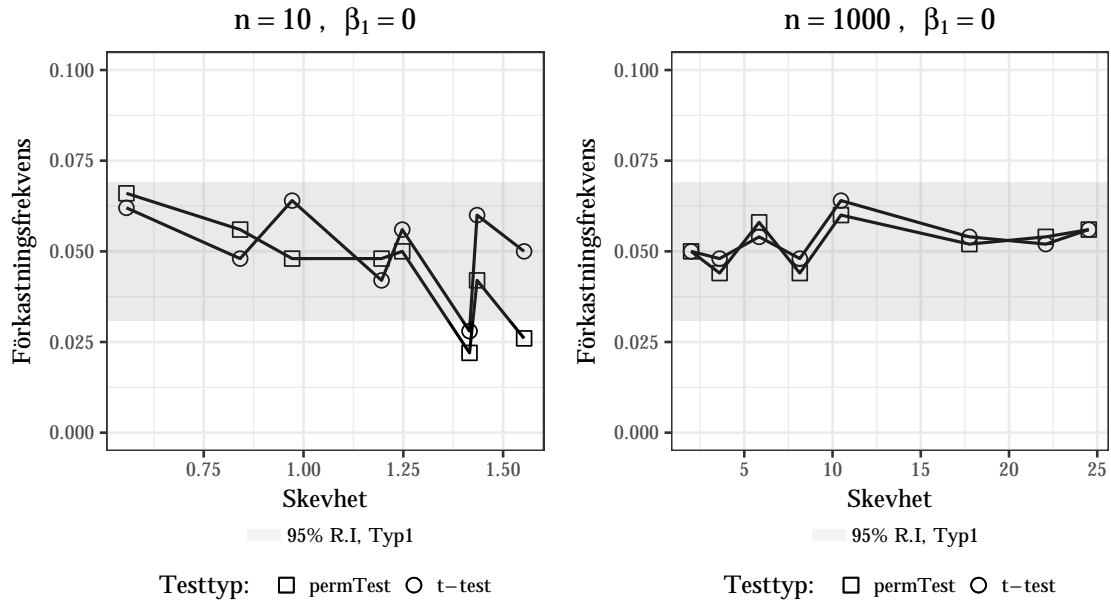


Figur A.7: Förkastningsfrekvens som funktion av ökande kurtosis för två stickprovsstorlekar i fallet då $\beta_1 = 0$. De förklarande variablerna är genererade från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5. De båda testen presterar likartat med vad som sågs för det motsvarande fallet vid simpel linjär regression.

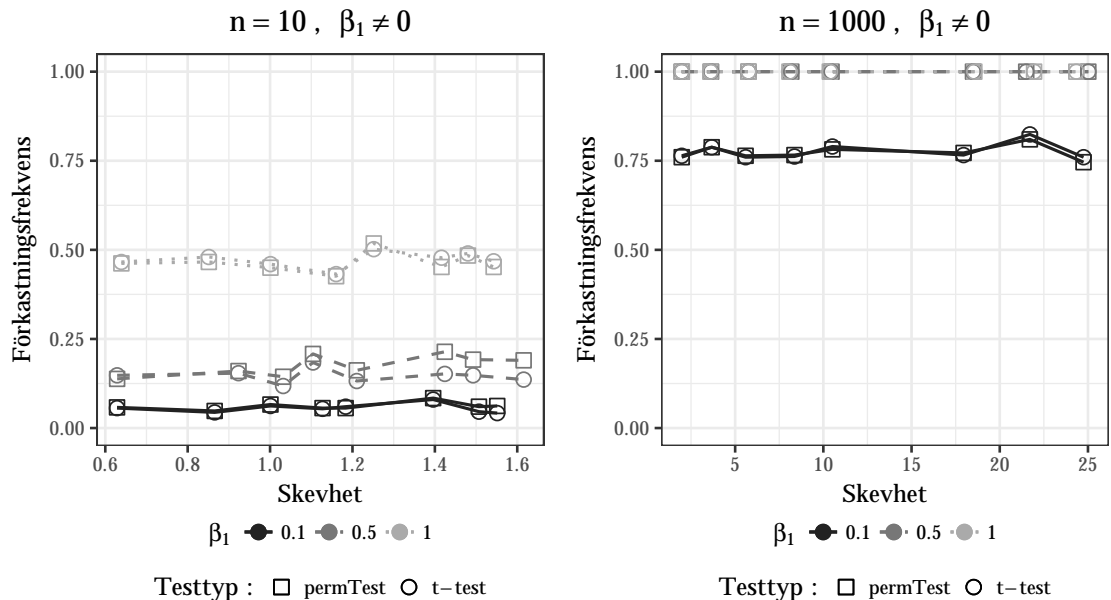


Figur A.8: Förkastningsfrekvens som funktion av ökande kurtosis för två stickprovsstorlekar i fallet då $\beta_1 \neq 0$. De förklarande variablerna är genererade från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5. De båda testen presterar likartat med vad som sågs för det motsvarande fallet vid simpel linjär regression.

A.2.2 Skevfördelade feltermer



Figur A.9: Förkastningsfrekvens som funktion av ökande skevhet för två stickprovsstorlekar i fallet då $\beta_1 = 0$. De förklarande variablerna är genererade från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5. De båda testen presterar likartat med vad som sågs för det motsvarande fallet vid simpel linjär regression.

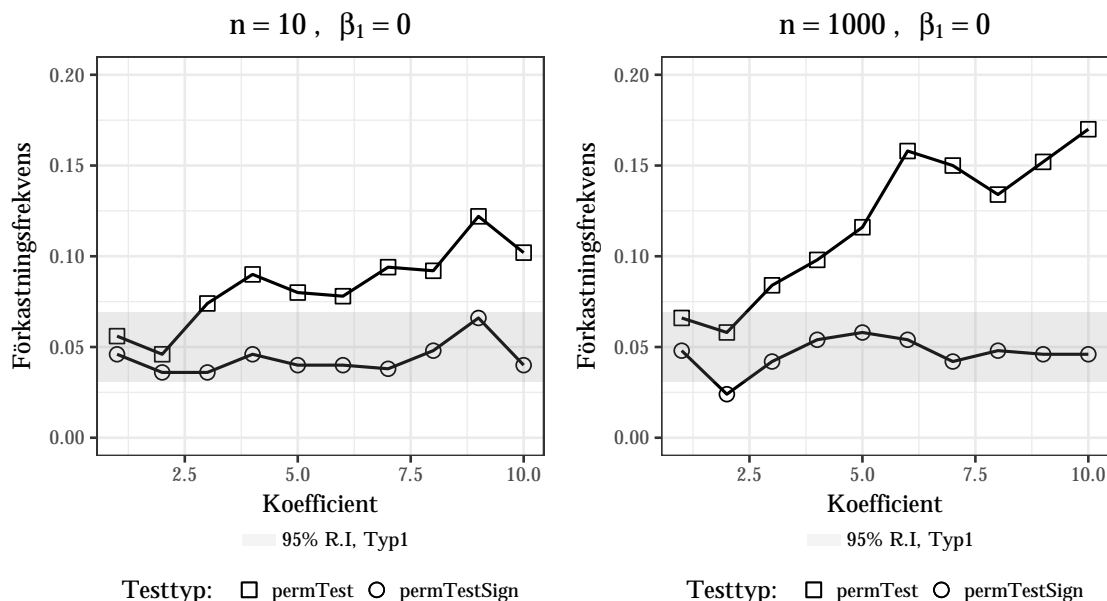


Figur A.10: Förkastningsfrekvens som funktion av ökande skevhet för två stickprovsstorlekar i fallet då $\beta_1 \neq 0$. De förklarande variablerna är genererade från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5. De båda testen presterar likartat med vad som sågs för det motsvarande fallet vid simpel linjär regression.

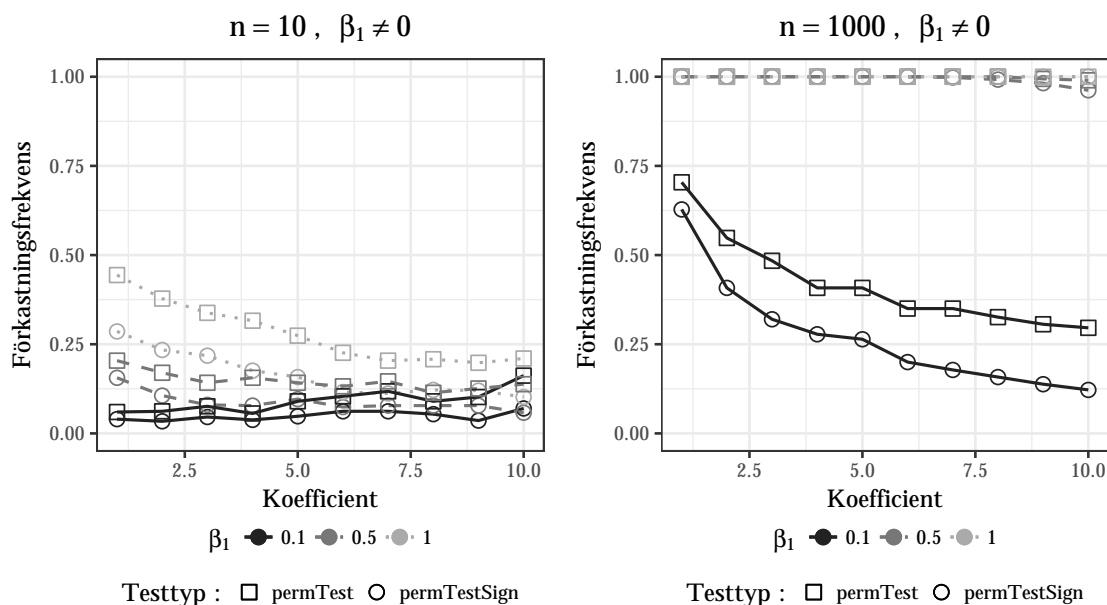
A.2.3 Utstickare

Samtliga grafer presenteras i avsnitt 5.2.3

A.2.4 Icke-konstant varians



Figur A.11: Förkastningsfrekvens som funktion av ökande lutningskoefficient för två stickprovsstorlekar i fallet då $\beta_1 = 0$. De förklarande variablerna är genererade från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5. Det vanliga permutationstestet är icke-giltigt medan permutationstest med teckenbyte bevarar signifikansnivån och är därmed giltigt. t -testet, som inte presenteras i figuren, presterar likartat med vad som sågs för det motsvarande fallet vid simpel linjär regression.



Figur A.12: Förkastningsfrekvens som funktion av ökande lutningskoefficient för två stickprovsstorlekar i fallet då $\beta_1 \neq 0$. De förklarande variablerna är genererade från en bivariat normalfördelning med väntevärde 0, varians 1 och korrelationskoefficient 0.5. I Figur A.11 sågs det att endast permutationstest via teckenbyte var giltigt för fallet $\beta_1 = 0$ vilket gör att endast detta test kan användas för fallet $\beta_1 \neq 0$. För detta fall presterar testet likartat med vad som sågs för det motsvarande fallet vid simpel linjär regression.

B Exempel på utbytbara beroende slumpvariabler

Låt \mathbf{X} vara en n -dimensionell kolumnvektor med normalfördelade variabler. Den simultana täthetsfunktionen av \mathbf{X} är,

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n) = \frac{\exp(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}))}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}}, \quad (9)$$

där $\boldsymbol{\Sigma} = \mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$ och $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$.

Antag att variablerna X_i och X_j har lika varians och lika parvis korrelation,

$$\mathbf{E}[(X_i - \mu)(X_j - \mu)] = \begin{cases} \sigma^2, & \text{om } i = j \\ \rho, & \text{om } i \neq j \end{cases} \quad (10)$$

Antag vidare att medelvärdet för variablerna är lika, $\boldsymbol{\mu} = (\mu, \mu, \dots, \mu)^T$.

Låt P vara en permutationsmatris, sådant att $P\mathbf{X} = (x_{i_1}, x_{i_2}, \dots, x_{i_n})^T$. Där i_1, i_2, \dots, i_n är de permuterade indexen för vektorn \mathbf{X} . Ekvationen (9) är då lika med,

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2, \dots, x_n) &= \frac{\exp(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T P^T P \boldsymbol{\Sigma}^{-1} P^T P(\mathbf{X} - \boldsymbol{\mu}))}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} = \\ &= \frac{\exp(-\frac{1}{2}(P(\mathbf{X} - \boldsymbol{\mu}))^T (P \boldsymbol{\Sigma} P^T)^{-1} (P\mathbf{X} - P\boldsymbol{\mu}))}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} = \\ &= \frac{\exp(-\frac{1}{2}(P\mathbf{X} - P\boldsymbol{\mu})^T (\mathbf{E}[P(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T P^T])^{-1} (P\mathbf{X} - P\boldsymbol{\mu}))}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} = \\ &= \frac{\exp(-\frac{1}{2}(P\mathbf{X} - P\boldsymbol{\mu})^T (\mathbf{E}[(P\mathbf{X} - P\boldsymbol{\mu})(P\mathbf{X} - P\boldsymbol{\mu})^T])^{-1} (P\mathbf{X} - P\boldsymbol{\mu}))}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} = \\ &= \frac{\exp(-\frac{1}{2}(P\mathbf{X} - \boldsymbol{\mu})^T (\mathbf{E}[(P\mathbf{X} - \boldsymbol{\mu})(P\mathbf{X} - \boldsymbol{\mu})^T])^{-1} (P\mathbf{X} - \boldsymbol{\mu}))}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}}. \end{aligned}$$

Från antagandet i (10) följer det att kovariansmatrisen

$$\mathbf{E}[(P\mathbf{X} - \boldsymbol{\mu})(P\mathbf{X} - \boldsymbol{\mu})^T] = [\mathbf{E}[(X_{i_k} - \mu)(X_{i_j} - \mu)]_{k,j=1}^n] = [\mathbf{E}[(X_k - \mu)(X_j - \mu)]_{k,j=1}^n] = \mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T].$$

Eftersom $\det(P \boldsymbol{\Sigma} P^T) = \det(P) \det(\boldsymbol{\Sigma}) \det(P^T) = (\pm 1) \det(\boldsymbol{\Sigma}) (\pm 1) = \det(\boldsymbol{\Sigma})$, följer det att $f_{\mathbf{X}}(x_1, \dots, x_n) = f_{\mathbf{X}}(x_{i_1}, \dots, x_{i_n})$.

C Kod

Våra slutgiltiga implementeringar finns tillgängliga med instruktioner på <https://github.com/Samev/BScPermtestLinReg>.

