

Analys av bakterier i tarmfloran och dess påverkan på övervikt

Nicole Osin Ceder

Kandidatuppsats 2016:15
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Analys av bakterier i tarmfloran och dess påverkan på övervikt

Nicole Osin Ceder*

Juni 2016

Sammanfattning

Syftet med detta arbete är att undersöka om förekomsten av olika bakteriearter i tarmfloran har ett samband med övervikt. I en allergistudie vid Södersjukhuset i Stockholm väcktes misstankar om att det finns ett sådant samband och därmed har vår uppgift blivit att undersöka detta närmare. Fecesprover har samlats in från 70 barn i tidig ålder där man kunnat mäta i vilken mängd olika bakteriearter har förekommit i tarmen. Dessa barn har sedan blivit vägda och mätta vid 2, 5 och 10 års ålder för att räkna ut barnens BMI.

Genom olika statistiska metoder har enskilda samband studerats mellan två olika BMI-variabler och fem olika bakteriearters förekomst i tarmen. Multipel logistisk och multipel linjär regression har tillämpats för att undersöka hur förekomsten av de olika bakteriearterna tillsammans påverkar BMI. Upprepande gånger har koefficienten för bakteriearten *S. aureus* visats sig vara signifikant i våra modeller och är den art som verkar minska risken för övervikt.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: nicole-osin-ceder@hotmail.com. Handedare: Jan-Olov Persson.

Sammanfattning

Syftet med detta arbete är att undersöka om förekomsten av olika bakteriearter i tarmfloran har ett samband med övervikt. I en allergistudie vid Södersjukhuset i Stockholm väcktes misstankar om att det finns ett sådant samband och därmed har vår uppgift blivit att undersöka detta närmare. Fecesprover har samlats in från 70 barn i tidig ålder där man kunnat mäta i vilken mängd olika bakteriearter har förekommit i tarmen. Dessa barn har sedan blivit vägda och mätta vid 2, 5 och 10 års ålder för att räkna ut barnens BMI.

Genom olika statistiska metoder har enskilda samband studerats mellan två olika BMI-variabler och fem olika bakteriearters förekomst i tarmen. Multipel logistisk och multipel linjär regression har tillämpats för att undersöka hur förekomsten av de olika bakteriearterna tillsammans påverkar BMI. Upprepande gånger har koefficienten för bakteriearten *S. aureus* visats sig vara signifikant i våra modeller och är den art som verkar minska risken för övervikt.

Abstract

The objective of this thesis is to investigate if different bacterias in the intestinal flora influences overweight in young children. In a previous allergy study at Södersjukhuset in Stockholm, they raised suspicions that there is such a connection, and thus, our task has been to examine this more closely. Faecal samples were collected from 70 children at an early age where one could measure the amount of different species of bacteria occurred in the intestine. These children have been weighed and measured at 2, 5 and 10 years of age in order to calculate children's BMI.

Through various statistical methods we have studied individual relationship between two variables of BMI and five different bacterial species aggregated in the gut. Multiple logistic and multiple linear regression was applied to investigate if the prevalence of different bacterial species together affects BMI. The coefficient of the specie *S. aureus* has repeatedly shown to be significant in our models, so *S. aureus* reduce the risk of obesity.

Förord

Denna uppsats utgör ett självständigt examensarbete i matematisk statistik om 15 hp vid Stockholms Universitet. Ett stort tack till Maria Johansson och Eva Sverremark-Ekström vid institutionen för molekylär biovetenskap som gav mig möjlighet att få utföra detta arbete samt till min handledare Jan-Olov Persson för stort stöd och goda råd under arbetets gång.

Innehåll

1	Inledning	1
1.1	Bakgrund	1
1.2	Syfte	2
1.3	Beskrivning av data	2
1.3.1	Bakterier i tarmfloran	2
1.3.2	Kolonisering av bakterier	3
1.3.3	BMI - Body Mass Index	4
1.4	Metod	4
2	Teori	6
2.1	Kategoriska variabler	6
2.2	2×2-kontingenstabell och Oddsquot	6
2.2.1	Fisher's exakta test	7
2.3	Student's t-test	7
2.4	Bonferroni's korrektion	8
2.5	Logistisk regression	8
2.5.1	Tolkning av de skattade koefficienterna	8
2.5.2	Modellkontroll	9
2.6	Linjär regression	9
2.6.1	Modellkontroll	10
3	Analys	11
3.1	Responsvariabel	11
3.1.1	Avvikelse från överviktsgränsen	13
3.2	Korrelation mellan förekomsten av bakteriearterna	15
3.3	Samband mellan BMI och bakterieförekomst	16
3.3.1	Responsvariabeln Övervikt	16
3.3.2	Multipl logistisk regression	17
3.3.3	Responsvariabeln BMI	18
3.3.4	Multipl linjär regression	19
3.4	Modellkontroll för regressionsmodellerna	19
4	Slutsats	22
5	Diskussion	23
6	Appendix	25

1 Inledning

På Sachsska Barnsjukhuset vid Södersjukhuset i Stockholm har en grupp från Institutionen för Molekylär Biovetenskap vid Stockholms Universitet genomfört en studie för att undersöka vilka faktorer som påverkar hur barn utvecklar allergier. Frivilliga föräldrar deltog i studien mellan åren 1997 och 2000 genom att lämna uppgifter om sig själva och deras barn. Förutom information om familjernas allergihistoria samlades även fecesprov (avföringsprov) från barnen in för att undersöka förekomsten av olika bakteriearter i tarmfloran och hur det påverkar barnens allergiutveckling. Barnens vikt och längd har även samlats in under flera tillfällen. Under studiens gång noterades att barn med högt BMI hade liknande bakteriearter i sin tarmflora och detta har lett till att vi i denna uppsats vill undersöka om det faktiskt finns ett sådant samband. [7]. Det har tidigare gjorts studier på råttor där man har påvisat att förekomsten av bakteriearten *Laktobaciller* bidrog till mindre vikt hos råttorna och vi vill undersöka om det föreligger ett samband mellan förekomsten av en bakterieart och barns vikt även på human data. [9]

1.1 Bakgrund

Människan bär på en mängd bakterier i kroppen och i tarmfloran hittar vi hundratals olika bakteriearter och med hjälp av fecesprov kan man ta reda på vilka dessa bakteriearter är. Barnen i studien fick lämna fecesprov vid fyra tidpunkter en kort tid efter födseln (vecka 1, vecka 2, månad 1, månad 2). Från fecesproverna kunde DNA renas fram med hjälp av en metod kallad *Real Time PCR (Polymerase Chain Reaction)*. Genom att rena utdraget av DNA kan man få fram en relativ mängd av en bakterieart baserad på hur mycket DNA som totalt renats i processen med en detektionsgräns på $5 \cdot 10^{-6}$ %. I studien på Södersjukhuset valde man att studera endast fem bakteriearter i varje prov, *Bifidobacterium (B.) adolescentis*, *B. breve*, *B. bifidum*, *Staphylococcus (S.) aureus* samt *Laktobaciller*. Den sistnämnda är ett samlingsnamn för en grupp av bakterier *L. casei*, *L. paracasei* och *L. rhmnosus*. [6]

Inom medicin talar man ofta om kolonisering av bakteriearter, vilket innebär att en bakterie har fått fäste och växer på en ny plats. Bakterierna i tarmfloran skyddar oss mot sjukdommar och främmande och skadliga bakterier vilket gör att de är viktiga för vår hälsa. De måste överleva färdan genom mage och tarm och även kolonisera sig i tarmen. [5] När en bakterieart har förekommit i tarmen under en längre tid säger man att arten är *koloniserad*. Några år efter att man föds har en bakterieflora i tarmen byggts upp, vilket innebär att vissa bakteriearter har koloniserat sig och vissa inte. Det finns ingen regel för hur man bestämmer när en art är koloniserad men i allergistudien sa man att en art var koloniserad om den fanns detekterad i fler än ett fecesprov.

Barnen följdes upp vid 2, 5 och 10 års ålder för att väga och mäta dem. Barnens BMI räknades ut, som är ett mått som används för att kunna definera övervikt och fetma, och defineras,

$$\text{BMI} = \frac{\text{vikt [kg]}}{\text{längd}^2 \text{ [meter]}}.$$

För en vuxen människa är gränsen för övervikt ett BMI lika med 25 kg/m^2 , medan överviktsgränsen för barn upp till 18 år varierar med åldern. Under barns uppväxt förändras kroppen ständigt och på olika sätt för pojkar och flickor. Därför spelar både ålder och kön roll för vilket BMI som anses vara övervikt. I artikeln "*Establishing a standard definition for child overweight and obesity*

worldwide: international survey" av T.J Cole et al. beskrivs hur överviktsgränserna har tagits fram för barn vid olika åldrar och för respektive kön. I tabell 1 listas överviktsgränserna vid 2, 5 och 10 års ålder.

Tabell 1: Överviktsgränsen vid de olika åldrarna för respektive kön.

	2 år	5 år	10 år
Pojkar	18.41	17.42	19.84
Flickor	18.02	17.15	19.86

Överviktsgränsen är högre för pojkar än för flickor vid 2 och 5 års ålder medan överviktsgränsen för flickor är högre vid 10 års ålder, detta är ett exempel på att pojkar och flickor utvecklas olika i kroppen under uppväxten. Utifrån dessa gränser definieras övervikt för barnen i detta arbete. [2]

1.2 Syfte

Syftet med detta arbete är att undersöka om det finns ett samband mellan förekomsten av olika bakteriearter i tarmfloran och barns kroppsvikt. Minskar risken för övervikt om en viss bakterieart finns koloniserad i tarmen?

1.3 Beskrivning av data

I detta arbete används ett dataset bestående av information om 70 barn födda vaginalt efter vecka 36 mellan åren 1997 och 2000 på olika sjukhus runt om i Stockholm. Vid fyra olika tidpunkter efter födseln (vecka 1, vecka 2, månad 1, månad 2) har man med hjälp av fecesprov från barnen samlat information om mängden av olika bakteriearter i tarmfloran. De arter vi kommer att studera i detta arbete är de fem bakteriearterna som de valde att undersöka i allergistudien. Barnens vikt och längd har uppmätts vid åldrarna 2, 5 och 10 år. [7]

1.3.1 Bakterier i tarmfloran

I datasetet anges hur stor procentandel varje bakterieart utgör av den totala mängden bakteriearter i varje fecesprov. Metoden för att ta fram den relativa mängden har en detektionsgräns på $5 \cdot 10^{-6}$ %, och om en art inte har detekterats i ett fecesprov anges andelen 10^{-6} % i datasetet. Det kan förekomma uteblivna fecesprov vilket har lett till att det saknas data för bakteriearterna vid den tidpunkten. [6]

Eftersom det finns hundratals bakterier i tarmfloran utgör varje bakterieart en liten procent av den totala mängden bakterier. Därmed förekommer det ofta att en art inte går att detektera i ett fecesprov på grund av detektionsgränsen. Med hjälp av tabell 2 försöker vi skaffa förståelse för hur ofta en art är detekterad och i vilken mängd de utgör i datasetet för varje fecesprov.

Tabell 2: Intervallet av bakteriemängden för de detekterad arterna samt medianvärdet.

Tid	Bakterie	Andel (Min, Max) [%]	Median [%]	Antal det.
Vecka 1	Laktobaciller	$(10^{-5}, 0.68)$	0.0004	17
	B. adolescentis	$(10^{-5}, 0.87)$	0.0030	26
	B. breve	$(3 \cdot 10^{-3}, 2.70)$	0.2698	18
	B. bifidum	$(10^{-5}, 1.86)$	0.1274	21
	S. aureus	$(10^{-5}, 0.04)$	0.0027	43
Vecka 2	Laktobaciller	$(10^{-5}, 0.46)$	0.0005	18
	B. adolescentis	$(10^{-5}, 2.23)$	0.0057	21
	B. breve	$(10^{-5}, 4.98)$	0.3938	18
	B. bifidum	$(10^{-5}, 9.54)$	0.8192	20
	S. aureus	$(3 \cdot 10^{-5}, 0.08)$	0.0043	46
Månad 1	Laktobaciller	$(10^{-5}, 0.36)$	0.0013	20
	B. adolescentis	$(3 \cdot 10^{-5}, 2.05)$	0.0190	18
	B. breve	$(4.3 \cdot 10^{-3}, 10.80)$	0.2551	20
	B. bifidum	$(10^{-5}, 8.80)$	0.4193	20
	S. aureus	$(10^{-5}, 0.07)$	0.0022	48
Månad 2	Laktobaciller	$(10^{-5}, 0.23)$	0.0011	29
	B. adolescentis	$(10^{-5}, 5.40)$	0.0072	26
	B. breve	$(10^{-4}, 23.59)$	0.3085	22
	B. bifidum	$(10^{-5}, 7.75)$	0.1133	28
	S. aureus	$(2 \cdot 10^{-5}, 0.03)$	0.0009	46

Notera att *S. aureus* är detekterad hos betydligt fler barn vid de olika tidpunkterna jämfört med övriga bakteriearter men har en väldigt låg median. Detta skulle kunna betyda att *S. aureus* lättare får ett fäste i tarmen än de övriga arterna och förekommer i en jämnare nivå vilket gör att arten blir mer stabil i tarmfloran. Bakterierna *B. breve* och *B. bifidum* har det största intervallet mellan minsta och största procentandel och ändå ett högt medianvärde. Detta beror på att när de väl förekommer i tarmen utgör arterna en större mängd än de övriga bakteriearterna.

1.3.2 Kolonisering av bakterier

I datasetet finns en variabel som beskriver om en bakterieart är koloniserad eller ej. För att bestämma om en art är koloniserad i tarmen räcker det att arten är detekterad i minst två fecesprov. Därför kan vi, trots vissa uteblivna prov, bestämma för de flesta barn vilka arter som är koloniserade och ej. Till exempel kan vi bestämma att en art är koloniserad om den är detekterad i två fecesprov medan de andra två proven saknas. Däremot om två fecesprov inte kan detektera en art och de andra två proven saknas, kan vi inte bestämma vare sig arten är koloniserad eller ej.

Tabell 3: Andelen kolonierade bakteriearter.

Bakterieart	koloniserade
Laktobaciller	34% (21/62)
B. adolescentis	38% (25/65)
B. breve	32% (21/66)
B. bifidum	36% (24/66)
S. aureus	77% (51/66)

I tabell 3 kan vi se att flest barn har koloniserade *S. aureus*-bakterier i tarmen. Som vi såg i tabell 2 hade många barn en detekterad *S. aureus*-bakterie vid de olika tidpunkterna vilket då medför att bakteriearten är koloniserad hos många barn.

Genom att studera om en art är koloniserad kan vi göra ett antagande om att arten fått fäste i tarmen hos ett barn. Lika så om en art inte är koloniserad då vi antar att arten inte fått ett fäste. Eftersom en art kan förekomma i tarmen vid en tidpunkt utan att den är koloniserad där spelar det därför ingen roll hur stor andel arten utgör i just den tidpunkten eftersom den lika väl inte behöver förekomma vid nästa provtagning. Mängden av en bakterieart i tarmfloran hos en vuxen är mer stabil under en längre tid jämfört med hos ett barn. Det tar en tid innan ett barn hinner bygga upp en stabil tarmflora och därför varierar mängden av arterna mycket och vi blir därför intresserade av vilka arter som fått fäste i tarmen istället för mängden av varje art vid olika tidpunkter. Därav är kolonisering fortsättningsvis den variabel som blir intressant för att undersöka syftet i detta arbete.

1.3.3 BMI - Body Mass Index

Genom att väga och mäta barnen har man kunnat räkna ut deras BMI vid 2, 5 och 10 års ålder. På samma sätt som det saknas fecesprov i studien saknas det även BMI-värden för vissa barn vid olika tidpunkter då mätningarna skulle ha skett.

BMI-värdena som räknas ut är kontinuerliga men på samma sätt som för bakteriearterna så finns ytterligare en variabel för BMI i datasetet. Nämligen en kategorisk variabel som beskriver om ett barn är överviktigt eller ej, vilket har bestämts med hjälp av överviktsgräserna. Vi skiljer på de två olika BMI-variablerna genom att kalla den kategoriska, eller binära, variabeln för **Övervikt**, som då antar värdet 1 (överviktig) eller 0 (normalviktig) och den kontinuerliga variabeln för **BMI**, som antar BMI-värdena för varje barn. **Övervikt** ger oss möjlighet att se om det finns något gemensamt för de överviktiga barnen i deras tarmflora.

I tabell 4 ser vi fördelningen av antalet överviktiga pojkar och flickor vid respektive ålder.

Tabell 4: Andelen överviktiga barn.

Kön	2 år	5 år	10 år
Pojkar	15% (6/40)	16.2% (6/37)	15.6% (5/32)
Flickor	17.2% (5/29)	10.3% (3/29)	18.5% (5/27)
Totalt	15.9% (11/69)	13.6% (9/66)	16.9% (10/59)

Det saknas fler BMI-värden ju äldre barnen blir. Detta beror säkerligen på att det är svårt att följa upp individer i en studie som sträcker sig under så lång tid. Andelen överviktiga barn är däremot ungefär lika för de tre olika åldrarna fastän det totala antalet observerade BMI-värden är olika. Det är 15.9%, 13.6% och 16.9% av det totala antalet barn som är överviktiga vid de olika åldrarna.

1.4 Metod

I detta arbete kommer i huvudsak multipel logistisk och multipel linjär regression att tillämpas för att undersöka om ett samband existerar mellan förekomsten av olika bakteriearter och BMI. Men innan dess undersöker vi om det går att påvisa enskilda samband mellan BMI och förekomsten av

de olika bakteriearterna i tarmfloran. Metoder som t-test och 2x2 tabeller tillämpas för att utreda detta.

2 Teori

2.1 Kategoriska variabler

Kategoriska variabler används ofta inom statistiken och är en variabel uppdelad i kategorier som beskriver olika kvalitéer. Det finns ordnade samt oordnade kategoriska variabler. För de ordnade variablerna finns det en ordning mellan de olika kategorierna medan det inte finns det för de oordnade variabler.

När en kategorisk variabel endast har två möjliga utfall, det vill säga två kategorier, kallar vi den för en *binär variabel*. [1]

2.2 2×2-kontingenstabell och Oddskvot

En 2×2-kontingenstabell kan bildas av två binära variabler A och B, där varje cell i tabellen anger en frekvens. Vi illustrerar detta med ett exempel där både A och B är uppdelade i två olika kategorier, 1 och 2. Vi har exempelvis en population av intresse där varje individ tillhör en kategori för varje variabel. Detta illustreras i tabell 5 för $N = \sum_{i,j} n_{ij}$ individer, där n_{+1} , n_{+2} , n_{1+} och n_{2+} står för rad- och kolumnsummorna.

Tabell 5: 2×2-tabell.

		A		
		1	2	
B	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	N

Med hjälp av en kontingenstabell kan man räkna ut en **oddskvot**, *odds ratio* (OR).

$$OR = \frac{\frac{P(A=1|B=1)}{P(A=2|B=1)}}{\frac{P(A=1|B=2)}{P(A=2|B=2)}} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}. \quad (1)$$

Oddskvoten är kvoten av två odds, där odds beräknas med hjälp av betingade sannolikheter som vi kan se i 1, och beskriver sambandet mellan de två variablerna A och B. Vi tolkar oddskvoten,

$OR > 1$: B=1 ökar oddsen att A=1.

$OR = 1$: B=1 påverkar inte A=1, lika sannolikt att A=1 som att A=2.

$OR < 1$: B=1 minskar oddsen att A=1.

Oddskvoten beror inte på hur stor populationen är utan endast på förhållandet mellan frekvenserna i tabellen.

2.2.1 Fisher's exakta test

Fisher's exakta test används vid 2×2 -tabeller för att undersöka oberoende mellan två binära variabler, detta motsvarar att oddskvoten är lika med ett. Med hjälp av Fisher's exakta test erhålls ett exakt p-värde även då antalet observationer är litet.

Frekvenserna i tabellcellerna $\{n_{ij}\}$ kommer från två oberoende binomiala stickprov eller från ett multinomialt stickprov. Rad- och kolumnsummorna är fixa vilket gör att cellfrekvenserna antar en hypergeometrisk fördelning. För en given rad- och kolumnsumma kan vi med hjälp av cellfrekvensen n_{11} , som blir vår testvariabel, bestämma de övriga cellerna i tabellen. Under nollhypotesen, $OR=1$, kan vi bestämma sannolikheten att n_{11} antar ett specifikt värde t genom,

$$p(t) = P(n_{11} = t) = \frac{\binom{n_{1+}}{t} \binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}. \quad (2)$$

Vi förkastar nollhypotesen om sannolikheten är liten att erhålla ett minst lika extremt värde på n_{11} som det observerade värdet. [1]

2.3 Student's t-test

Ett t-test hjälper oss att avgöra om det föreligger någon skillnad mellan medelvärdena för två grupper. Låt X_1 och X_2 vara två oberoende normalfördelade stickprov med medelvärde μ_1 respektive μ_2 . Teststatistikan T definieras

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}},$$

där T är t-fördelad under nollhypotesen med $n_1 + n_2 - 2$ frihetsgrader och den poolade variansen,

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

där s_1 och s_2 står för de skattade varianserna för olika stickproven och n_1 och n_2 står för storleken på respektive stickprov.

Teststatistikan testas statistiskt på signifikansnivån α om det föreligger skillnad i gruppmedelvärdena.

$$H_0 : \mu_1 = \mu_2,$$

$$H_1 : \mu_1 \neq \mu_2.$$

Om vi kan förkasta nollhypotesen, H_0 , kan vi avgöra om gruppmedelvärdena statistiskt ej är lika varandra. [3]

2.4 Bonferroni's korrektion

Vid ett flertal statistiska test, där varje test har en signifikansnivå α , kan några av dessa erhålla ett p-värde mindre än α av ren slump trots att nollhypotesen är sann. Med hjälp av Bonferroni's korrektion kan vi ta hänsyn till detta.

Vid multipla jämförelser måste vi kontrollera för de falska signifikanser som kan uppstå. Den kritiska signifikansnivån α sätts ofta till 0.05, vilken man med Bonferroni's korrektion vill sänka. Genom att välja en kritisk signifikansnivå,

$$\alpha' = \frac{\alpha}{n},$$

där n är antalet multipla test som utförs, så sänker vi felrisken för varje test. [4]

2.5 Logistisk regression

Logistisk regression är en metod för att undersöka om det finns ett samband mellan en kategorisk beroende variabel Y som endast kan anta två olika värden och k oberoende förklarande variabler $X_1, \dots, X_k$. De förklarande variablerna X_i kan vara antingen kvalitativa eller kvantitativa. I en logistisk regression är man intresserad av ett samband mellan sannolikheten att Y ska anta ett av sina värden med X_i . Till exempel om Y antar värdet 0 eller 1 så är vi intresserade av $P(Y = 1)$.

Multipel logistisk regression definieras av funktionen som vi kallar *logit* och sambandet med de förklarande variablerna,

$$\text{logit}(P(Y = 1)) = \log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

Vilket kan skrivas om,

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}.$$

2.5.1 Tolkning av de skattade koefficienterna

Kvoten mellan oddsen för $Y=1$ om x_i ökar med en enhet och oddsen för $Y=1$ om x_i är konstant kallas för oddskvoten (OR).

Låt oss titta på OR för x_1 .

$$\text{OR} = \frac{\exp(\beta_0 + \beta_1(x_1 + 1) + \dots + \beta_k x_k)}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} = \exp(\beta_1).$$

Med hjälp av de skattade β -koefficienterna som vi erhåller när vi utför en logistisk regression kan vi beräkna den skattade oddskvoten som beskriver hur oddsen ökar/minskar då x_i ökar med en enhet.

Mer om hur koefficienter och oddskvoter skattas går att läsa i *Categorical Data Analysis*, av Alan Agresti (2002).

2.5.2 Modellkontroll

För logistisk regression finns det olika ”goodness of fit”-test att genomföra för att undersöka hur väl en modell är anpassad till observerad data. För ogrupperad data (som vi har i denna uppsats) är Hosmer-Lemeshow’s test ett sådant. Hosmer-Lemeshow’s test skattar först sannolikheter som sedan delas in i ungefär lika stora grupper g . Hosmer-Lemeshow’s teststatistika $\hat{C}$ utformas med hjälp av Pearson’s χ^2 -teststatistika för att jämföra observerade och anpassade värden.

$$\hat{C} = \sum_{k=1}^g \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right],$$

där

$$\begin{aligned} o_{1k} &= \sum_{j=1}^{c_k} y_j, \\ o_{0k} &= \sum_{j=1}^{c_k} (m_j - y_j), \\ \hat{e}_{1k} &= \sum_{j=1}^{c_k} m_j \hat{\pi}_j, \\ \hat{e}_{0k} &= \sum_{j=1}^{c_k} m_j (1 - \hat{\pi}_j). \end{aligned}$$

Vi har att c_k är antalet olika kombinationer de förklarande variablerna kan ha i grupp k , y_j är de observerade **Övervikts**-värdena, $\hat{\pi}$ är de anpassade värdena och m_j är antalet observationer med samma kombination av förklarande variabler för varje j . Teststatistikan $\hat{C}$ är approximativt χ^2 -fördelad med $g-2$ frihetsgrader.

Mer om Hosmer-Lemeshow’s test går att läsa i kapitel 5 i *Applied Logistic Regression*, av David W. Hosmer, Stanley Lemeshow, Rodney X (2013).

2.6 Linjär regression

Multipel linjär regression är en modell som tillämpas när vi har en kontinuerlig responsvariabel. Med den linjära modellen undersöks hur värdena på k stycken förklarande variabler påverkar värdet på responsvariabeln. Det linjära sambanden formuleras,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i,$$

för observation $i=1, \dots, n$ där de slumpmässiga och felen $\epsilon_i \sim N(0, \sigma^2)$ är oberoende och lika fördelade.

Mer om linjär regression går att läsa i *Lineära Statistisk Modeller*, av Rolf Sundberg (2015).

2.6.1 Modellkontroll

Förutsatt att man anpassar en linjär regressionsmodell gäller det att vissa antaganden ska vara uppfyllda. Antagandena som bör vara uppfyllda är att de slumpmässiga felen, ϵ_i , är oberoende och normalfördelade med konstant varians. För att kontrollera dessa antaganden tar vi hjälp av olika plottar. Vi studerar plottar över residualer och anpassade värden samt normalkvantilplottar. [10]

3 Analys

I avsnitt 1.3 beskrivs datasetet som används för analysen i denna uppsats, som sammanfattningsvis består av de fem olika bakteriearternas förekomst samt BMI vid 2, 5 och 10 års ålder. Vi kommer ihåg att vi valt att studera förekomsten, koloniseringen, av varje art framför att studera varje andel som en art utgör i de fyra olika fecesproven. Vi låter koloniseringen av en art sammanfatta arternas förekomst vid de fyra olika tidpunkterna som man mätt mängden av dem.

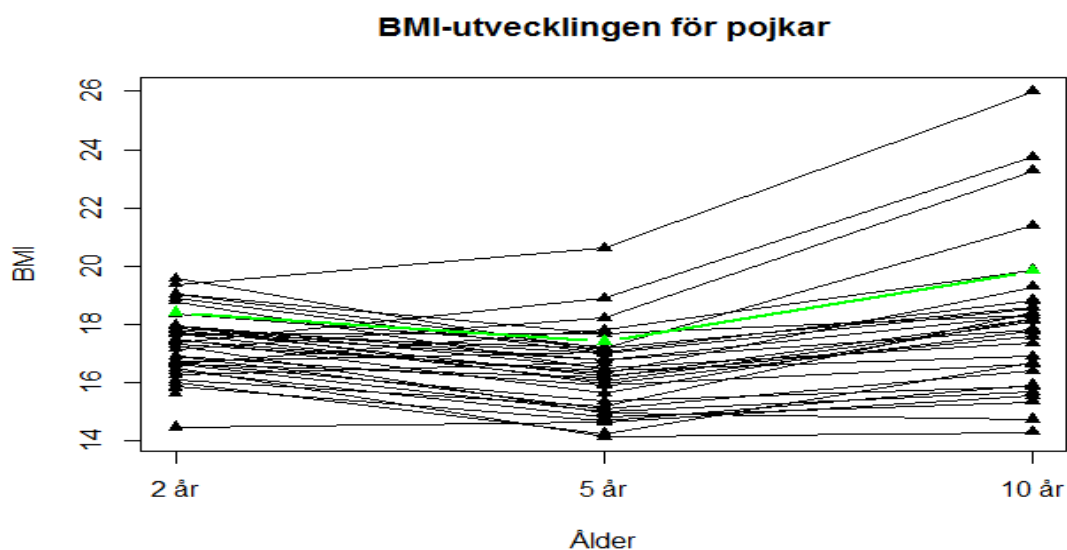
3.1 Responsvariabel

Responsvariabeln i detta arbete är BMI vid de olika åldrarna eftersom vi vill studera utfallet på barnens vikt beroende på förekomsten av de olika bakteriearterna. **BMI** är ett mått som anges i kg/m^2 och finns i datasetet som en variabel för de olika åldrarna som barnen blivit vägda och mätta. Som nämdes i avsnitt 1.3.3 så finns även en binär BMI-variabel **Övervikt** för de olika åldrarna. Vi har därmed två olika typer av responsvariabler för BMI, både en kontinuerlig och en binär sådan.

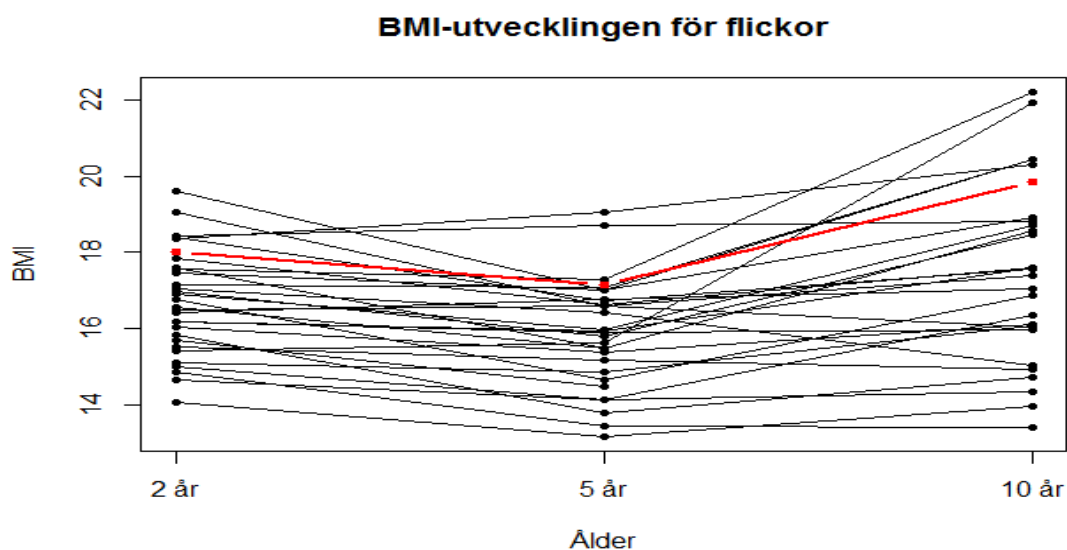
I tabell 1 visas överviktsgränsen för de olika åldrarna och som synes är de olika för pojkar och flickor, i alla fall vid 2 och 5 års ålder. Då överviktsgränsen är olika kan det bli svårt att jämföra barnens BMI-värden om vi inte tar hänsyn till kön. Problematiken är att en överviktig flicka kan ha lägre BMI än en normalviktig pojke vid 2 och 5 års ålder. Vi skulle kunna dela upp BMI-variabeln på två, en för respektive kön, men erhåller då två variabler med relativt få observationer. Därav kan variabeln **Övervikt** med fördel användas som responsvariabel för att inte behöva ta hänsyn till kön och samtidigt innehålla så många observationer som möjligt.

Vi går dock miste om en del information om vi endast studerar **Övervikt**. Genom att studera **BMI** kan vi eventuellt upptäcka samband mellan förekomsten av vissa bakteriearter och ett ökat/minskat BMI. Det behöver inte nödvändigtvis vara så att BMI-värdena ligger över överviktsgränsen för att ett sådant samband ska upptäckas.

Vi tittar närmare på hur BMI utvecklas över åldrarna och delar upp de i två grupper för de olika könen. I figur 1 och 2 kan vi se hur BMI förändras mellan de tre åldrarna för pojkar respektive flickor. Den utmarkerade linjen i de två figurerna är överviktsgränsen för de olika könen.



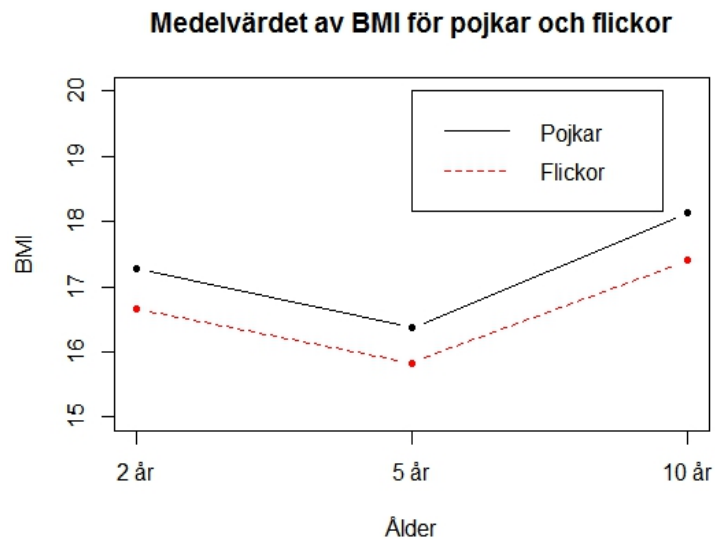
Figur 1: BMI-utvecklingen för pojkar med utsatt överviktsgräns.



Figur 2: BMI-utvecklingen för flickor med utsatt överviktsgräns.

I figurerna kan vi se hur de flesta barn ligger under överviktsgränsen i de tre olika åldrarna. Vi ser också hur spridningen på BMI ökar med åldern och det blir märkbart tydligt vilka barn som är överviktiga vid 10 års ålder eftersom det ligger längre ifrån överviktsgränsen. Vi noterar också att det i de flesta fall är samma barn som är överviktiga vid de olika åldrarna.

För att se hur det skiljer sig mellan BMI för pojkar och flickor tittar vi på medelvärdet vid de olika åldrarna.



Figur 3: Medelvärdet av BMI för pojkar respektive flickor.

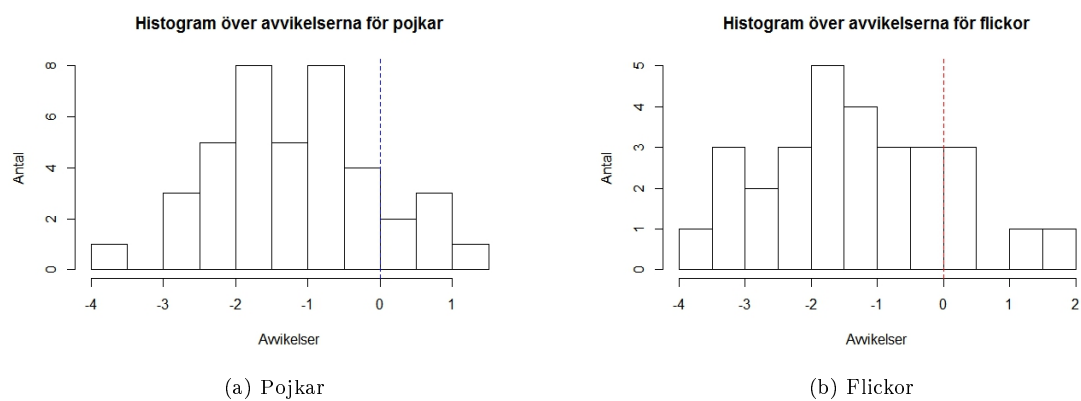
Flickors BMI-medelvärde är tydligt lägre än pojkars. Skillnaden mellan pojkar och flickor är nägorlunda konstant vid de olika åldrarna.

3.1.1 Avvikelser från överviktsgränsen

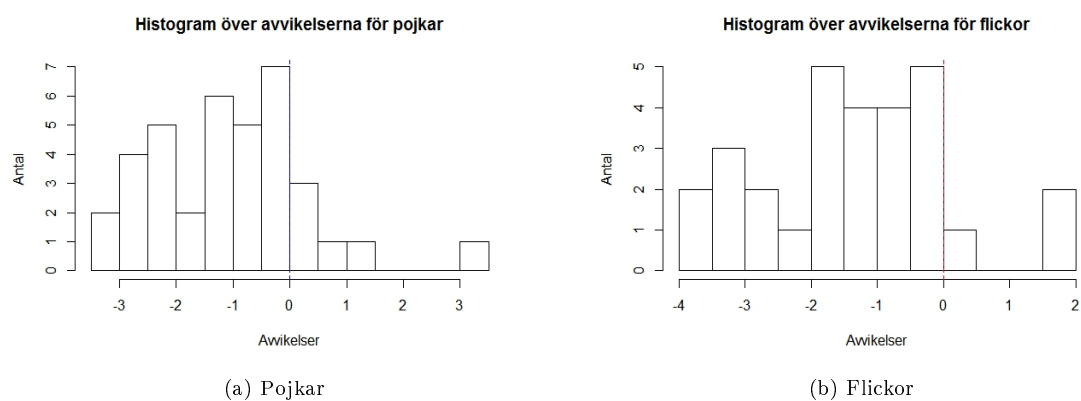
Som nämnades tidigare beror BMI-värdena på ålder och kön. Vi vill undvika att dela upp variabeln **BMI** på kön för att behålla så många observationer som möjligt. Enligt tabell 1 skiljer det sig inte så mycket mellan överviktsgränsen för pojkar och flickor vid 10 års ålder och vi kan nästan anta att de är lika. För BMI vid 10 års ålder tar vi inte hänsyn till kön vilket underlättar för oss då vi inte behöver justera för detta. Vid 2 och 5 års ålder är överviktsgränsen olika beroende på kön. Vi skapar en ny variabel **Avvikelse** som vi erhåller genom att dra bort rätt överviktsgräns från BMI-värdet beroende på kön och ålder. Detta görs endast för åldrarna 2 och 5 år.

Histogram för **BMI** vid 10 års ålder och **Avvikelse** vid 2 och 5 års ålder visas för att kunna se spridningen av variablerna för de olika könen.

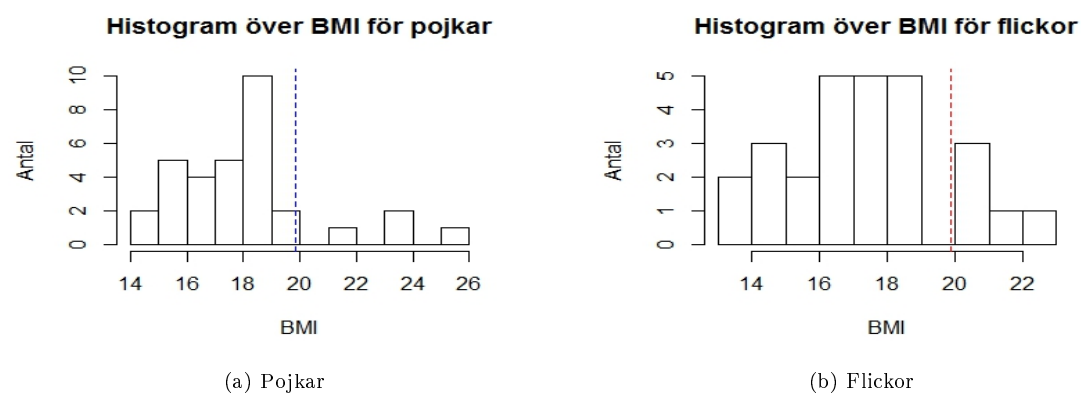
Figur 4: Histogram över BMI-avvikelserna vid 2 års ålder



Figur 5: Histogram över BMI-avvikelserna vid 5 års ålder



Figur 6: Histogram över BMI vid 10 års ålder



Histogrammen påvisar att BMI-avvikelserna inte skiljer sig så mycket åt mellan könen vid åldrarna 2 och 5 år. Vid 10 års ålder kan vi se att spridningen på BMI är större hos pojkar än hos flickor, det är nog också anledningen till att pojkars medelvärde var högre än flickors.

Barns kroppsvikt förändras ständigt under uppväxten och ju äldre de blir ju stabilare blir vikten. Därför kommer vi i detta arbete att använda oss av responsvariabeln **BMI** vid 10 års ålder eftersom det är den äldsta åldern i datasetet och därför mer stabil jämfört med de övriga åldrarna. Detta gäller även för variabeln **Övervikt** som också studeras endast vid 10 års ålder i fortsättningen.

3.2 Korrelation mellan förekomsten av bakteriearterna

I senare avsnitt tillämpas regressionsanalys för att undersöka sambandet mellan förekomsten av de olika bakteriearterna och BMI. Då är det viktigt att de oberoende förklarande variablerna inte är starkt korrelerade med varandra eftersom det blir problematiskt att hålla isär effekterna av dem på den oberoende responsvariabeln. Detta kallas multikollinearitet, alltså att två (eller flera) kovarianser är högt korrelerade med varandra, vilket vi måste undersöka för våra variabler.

Våra förklarande variabler i detta arbete är förekomsten av de fem bakteriearterna. För att ta reda på hur dessa korrelerar med varandra tar vi hjälp av oddskvoter, då våra variabler är binära. Oddskvoten testas för om den är lika med ett, vilket motsvarar oberoende mellan två variabler, och detta görs med hjälp av Fisher's exakta test med 5% felrisk. (Oddskvoterna erhålls från 2×2 -tabeller som läsaren hittar i Appendix.)

Tabell 6: Oddskvot och Fisher's exakta p-värde för de olika bakteriearterna.

	Oddskvot	P-värde
Lac, B. adol	0.62	0.56
Lac, B. breve	1.47	0.57
Lac, B. bifi	1.54	0.56
Lac, S. aur	0.85	1
B. adol, B. breve	0.44	0.27
B. adol, B. bifi	1.49	0.59
B. adol, S. aur	0.90	1
B. breve, B. bifi	1.28	0.78
B. breve, S. aur	0.46	0.22
B. bifi, S. aur	1.16	1

Om oddskvoten är lika med ett finns inget samband mellan variablerna. Som vi kan se i tabell 6 finner vi inget p-värde mindre än $\alpha=0.05$, alltså kan vi inte förkasta nollhypotesen och vi kan inte utesluta att de inte finns ett samband mellan förekomsten av de olika bakteriearterna.

Enligt Maria Johansson, som deltagit i den tidigare nämnda allergistudien på Södersjukhuset, kan det mycket väl finnas ett samband mellan förekomsten av bakteriearterna men att man inte vet så mycket om det i verkliga livet. Hon säger däremot att man i provrör har sett hur Laktobaciller hämmar tillväxten av S. aureus-bakterier. Ett sådant samband går inte att påvisa enligt våra oddskvoter och p-värden.

3.3 Samband mellan BMI och bakterieförekomst

3.3.1 Responsvariabeln Övervikt

Figur 7: 2x2-tabeller för bakteriearterna och Övervikt vid 10 års ålder.

Tabellerna visar på att *S. aureus* och **Övervikt** är signifikant beroende av varandra då vi får ett p-värde lika med 0.03. Detta betyder att det är signifikant skillnad på antalet barn som har koloniserade *S. aureus*-bakterier beroende på om de är normal- eller överviktiga. Utifrån tabellen ser vi att det är 13% (5/38) barn som är överviktiga av dem som har koloniserade *S. aureus*-bakterier jämfört med 71% (5/7) överviktiga av dem som inte har arten koloniserad. Detta ger oss oddskvoten 0.18, vilket säger att oddsen för övervikt minskar med 82% om arten *S. aureus* är koloniserad i tarmfloran jämfört med om arten inte är koloniserad.

16

$\frac{0.05}{5} = 0.01$, alltså att vi endast förkastar nollhypotesen om p-värdet är mindre än 0.01. Vilket leder till att vi i detta fall inte längre kan förkasta nollhypotesen vid jämförelse mellan *S. aureus* och **Övervikt** då p-värdet vi erhåller är 0.03.

3.3.2 Multipel logistisk regression

Multipel logistisk regression är en modell som kan tillämpas på en kategorisk responsvariabel. Modellen anpassas om intresset är att undersöka vilken effekt de förklarande variablerna har på responsen med hänsyn tagna till effekten av övriga variabler. Vi kan nu undersöka hur förekomsten av bakteriearterna tillsammans påverkar **Övervikt**.

Logistisk regression tillämpas på responsvariabeln **Övervikt** vid 10 års ålder. De förklarande variablerna i modellen är förekomsten av de fem olika bakteriearterna. Vi kallar dem,

$$x_{lak} = \begin{cases} 0, & \text{om Laktobacillerna ej är koloniserad,} \\ 1, & \text{om Laktobacillerna är koloniserad} \end{cases}$$

$$x_{adol} = \begin{cases} 0, & \text{om B. adolescentis ej är koloniserad,} \\ 1, & \text{om B. adolescentis är koloniserad} \end{cases}$$

$$x_{bre} = \begin{cases} 0, & \text{om B. breve ej är koloniserad,} \\ 1, & \text{om B. breve är koloniserad} \end{cases}$$

$$x_{bif} = \begin{cases} 0, & \text{om B. bifidum ej är koloniserad,} \\ 1, & \text{om B. bifidum är koloniserad} \end{cases}$$

$$x_{aur} = \begin{cases} 0, & \text{om S. aureus ej är koloniserad,} \\ 1, & \text{om S. aureus är koloniserad} \end{cases}$$

Låt $\pi(\mathbf{x}) = P(\text{Övervikt}=1)$, vilket ger oss modellen,

$$\text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_1 x_{lak} + \beta_2 x_{adol} + \beta_3 x_{bre} + \beta_4 x_{bif} + \beta_5 x_{aur}.$$

I tabell 7 visas det erhållna skattningarna och dess p-värde för den logistiska modellen.

Tabell 7: Skattningar för den logistiska regressionsmodellen.

Modell 1	Förkl.var.	β_i skattning	$\exp(\hat{\beta}_i)$	$\Pr(> z)$	Antal
Ålder 10	intercept	0.29	1.34	0.76	47
	Kol _{Lakto}	-0.67	1.51	0.51	
	Kol _{B.adol}	-0.28	0.76	0.77	
	Kol _{B.brev}	-0.70	0.50	0.51	
	Kol _{B.bifi}	0.38	1.46	0.68	
	Kol _{S.aur}	-2.41	0.09	0.01	

Koefficienten för bakteriearten *S. aureus* är signifikant på 5% felrisk, då p-värdet är 0.01. Med den skattade koefficienten för *S. aureus* erhåller vi oddskvoten 0.09, vilket betyder att om arten är koloniserad så minskar oddsen för övervikt med 91% i jämförelse med om arten inte är koloniserad. Arten verkar ha en stark effekt på **Övervikt**. Konfidensintervallet för oddskvoten är [0.01, 0.50] vilket betyder att effekten har en stor variation då *S. aureus* är koloniserad. Koefficienterna för de övriga bakteriearterna är inte statistiskt signifikanta på 5% felrisk och vi kan inte avgöra om de ger effekt på responsvariabeln **Övervikt** eller inte.

3.3.3 Responsvariabeln BMI

Efter att ha studerat sambandet mellan förekomsten av bakteriearterna och **Övervikt** gör vi liknande jämförelser för den kontinuerliga responsvariabeln **BMI**.

Med t-test testar vi för om två gruppmedelvärden är statistiskt skilda från varandra. **BMI** delas upp i två grupper för varje art, koloniserad eller ej, för att sedan beräkna de två medelvärdena i respektive grupp. Resultat från t-testen visas nedan.

Tabell 8: Resultat från t-test (Medel är gruppmedelvärdet för BMI).

Ålder	Art	Koloniserad		Ej Koloniserad		T-test	
		Medel	tot obs	Medel	Tot obs	t-stat	p-värde
10	Laktobaciller	16.97	17	18.26	34	1.89	0.07
	B.adolescentis	17.79	21	17.98	33	0.25	0.80
	B.Breve	17.98	16	17.67	39	-0.41	0.69
	B.bifidum	17.41	19	18.02	36	0.92	0.36
	S.aureus	17.55	43	19.10	12	1.49	0.15

T-testen ger inget signifikant resultat på 5% felrisk, men däremot på 10% felrisk. I detta fall är det förekomsten av Laktobacillerna som visar sig ha ett samband med BMI. Det vill säga, medelvärdet av BMI för de barn med koloniserade Laktobaciller skiljer sig statistiskt från medelvärdet för de barn som inte har arten koloniserad i tarmfloran. Utifrån tabellen ser vi att medelvärdet av BMI för de barn med koloniserade Laktobacillerna är 16.97 mot de ej koloniserade 18.26, vilket betyder att ett barn med koloniserade Laktobaciller har generellt lägre BMI vid 10 års ålder än ett barn som inte har arten koloniserad.

I jämförelse mellan resultatet från t-testen och kontingenstabellerna, i det tidigare avsnittet, är det inte samma art som visar signifikant resultat. Jämförelsevis gav t-testen, för **BMI** och *S. aureus*, ett p-värde på 0.15 medan kontingenstabellen, för **Övervikt** och *S. aureus*, gav ett p-värde på 0.03. Vi erhöll alltså inget signifikant t-test för arten men ändå ett ganska lågt p-värde. Åter igen bör vi ha i åtanke att det finns risk för falska signifikanser då fem olika jämförelser som görs med 5% felrisk vid varje test.

Vidare studeras om en kombination av bakteriearternas förekomst kan tänkas ha ett samband med de olika utfallen av **BMI**.

3.3.4 Multipel linjär regression

Multipel linjär regression tillämpas då istället responsvariabeln är kontinuerlig, vilket **BMI** är. Precis som för multipel logistisk regression vill vi undersöka hur de förklarande variablerna tillsammans kan förklara responsen som i det här fallet är **BMI**.

De förklarande variablerna är desamma som i modell 1, den logistiska regressionmodellen, medan responsen istället är **BMI**.

Detta ger den linjära modellen,

$$BMI = \beta_0 + \beta_1 x_{lak} + \beta_2 x_{adol} + \beta_3 x_{bre} + \beta_4 x_{bif} + \beta_5 x_{aur} \quad (3)$$

Resultatet från den anpassade modellen visas nedan.

Tabell 9: Skattningar för den linjära modellen.

Modell 2	Förkl.var.	β_i skattning	$\Pr(> z)$	Antal	R^2
Ålder 10	intercept	20.00	<2e-16	47	0.17
	Kol _{Lakto}	-1.20	0.13		
	Kol _{B.adol}	-0.38	0.45		
	Kol _{B.brev}	0.36	0.65		
	Kol _{B.bifi}	-0.33	0.63		
	Kol _{S.aur}	-2.03	0.02		

Skattningarna för den linjära modellen ger oss igenkännande resultat från den logistiska regressionsmodellen. Koefficienten för *S. aureus*-bakterien är signifikant på 5% felrisk då vi erhåller ett p-värde lika med 0.02. Den skattade koefficienten är -2.03, vilket betyder att om arten finns koloniserad har den en negativ inverkan på BMI. Barn med koloniserade *S. aureus*-bakterier har igenomsnitt 2.03 lägre BMI än de som inte har arten koloniserad.

3.4 Modellkontroll för regressionsmodellerna

För att avgöra om våra regressionsmodeller är bra anpassade tittar vi närmare på diagnostiken för den logistiska samt den linjära modellen.

För den logistiska regressionsmodellen tar vi hjälp av Hosmer-Lemeshow's test, som är ett "goodness-of-fit"-mått för att ta reda på hur bra anpassad modellen är till data. Eftersom vår data innehåller få observationer har vi valt $g=7$ när vi utför testet. Alltså delas de anpassade och observerade värdena in i sju olika grupper istället för tio som man oftast gör. Vi erhåller teststatistikan och p-värde enligt tabell 10.

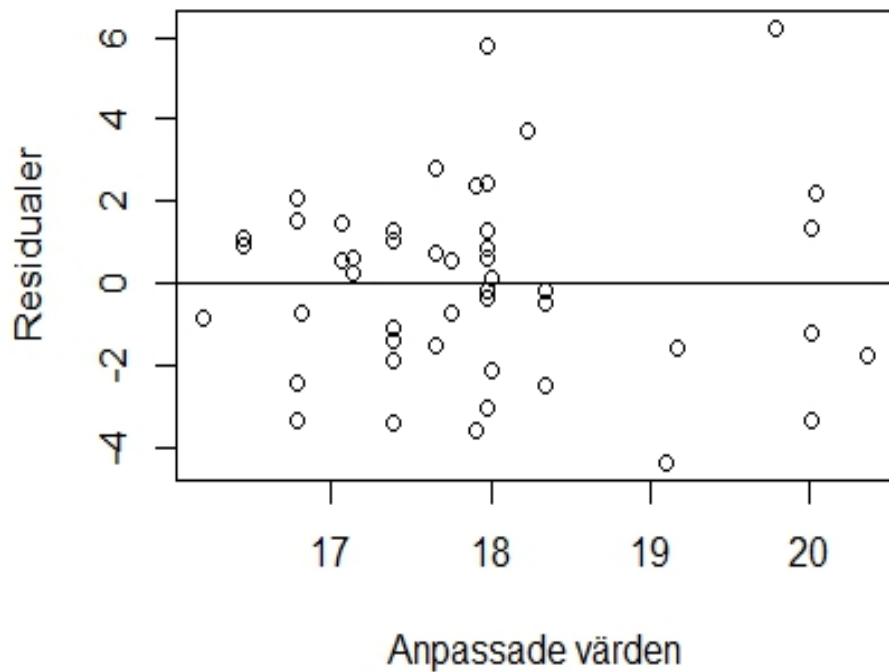
Tabell 10: Resultat från Hosmer-Lemeshow's test.

$\bar{C}$	frihetsgrader	p-värde
4.9104	5	0.43

Som vi ser i tabellen kan vi inte förkasta nollhypotesen, att modellen är väl anpassad, på 5% felrisk. Alltså har vi inte bevis för att det skulle vara en dåligt anpassad modell.

För den linjära modellen plottar vi först residualerna mot de anpassade värdena för att se hur de uppfyller modellantagandena om slumpmässiga residualer och konstant varians.

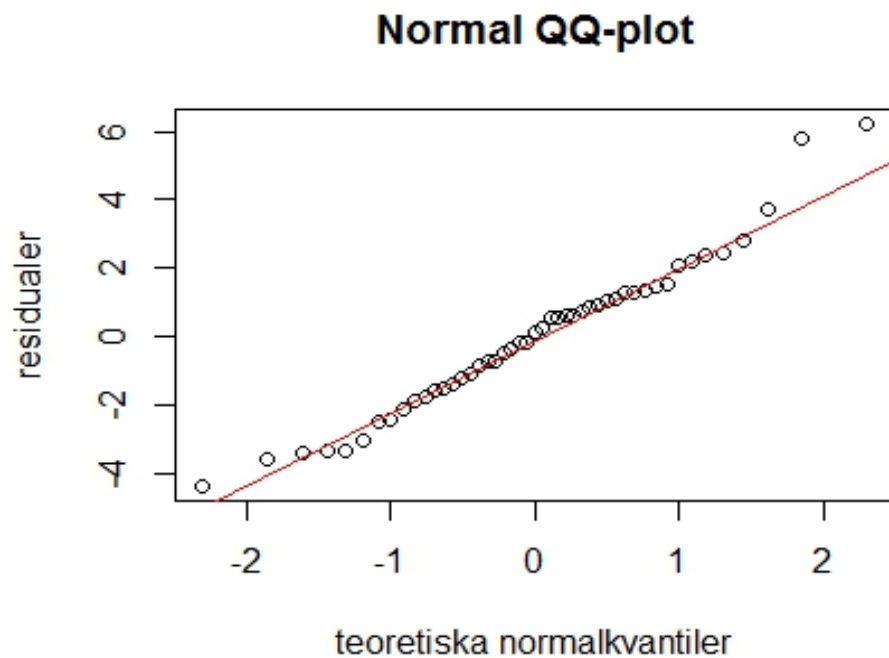
Figur 8: Residualplott för modell 2, BMI vid 10 års ålder.



Plotten bör inte kunna påvisa ett mönster utan bör vara slumpmässigt utspridda kring väntevärdet noll och som vi ser verkar så vara fallet. Variansen bör också vara konstant. Man kan dock ana att variansen ökar något när de anpassade värdena ökar.

Med hjälp av en normal QQ-plott kan vi se hur bra residualerna följer en normalfördelning.

Figur 9: Normalfördelnings-plott för modell 2, BMI vid 10 års ålder.



För den linjära modellen kan vi se att residualerna följer normalfördelningen ganska bra eftersom de ligger längst den diagonala linjen. Dock avviker residualerna vid svansarna en del.

4 Slutsats

I detta arbete har vi undersökt sambandet mellan förekomsten av olika bakteriearter i tarmfloran och BMI hos 70 barn. Vi har haft tillgång till både kontinuerliga och binära variabler för bakteriearterna och BMI. På grund av osäkerheten i mängden av en bakterieart hos ett barn i de unga åldrarna har vi funnit det mer informativt att använda oss av de binära variablerna för bakteriearterna som har angivit om arterna är koloniserade eller ej. För BMI-variablerna har vi studerat både den binära variabel **Övervikt** samt den kontinuerliga variabeln **BMI**. För både **Övervikt** och **BMI** har vi först undersökt om det funnits enskilda samband mellan dem och arternas förekomst. Vid test för våra 2×2 -tabeller visar sig **Övervikt** vara signifikant skild mellan övervikt och normalvikt beroende på om *S. aureus* är koloniserad eller ej. Men för t-testen visade det sig vara skillnad på **BMI** beroende på om arten *Laktobaciller* var koloniserad eller ej. De olika testen antyder att förekomsten av *S. aureus* och *Laktobaciller* bidrar till minskad risk för övervikt/högt BMI hos barn.

Efter undersökningen av varje enskild bakterieart och dess inflytande på **Övervikt** och **BMI** har vi senare anpassat multipel logistisk regression och multipel linjär regression för att undersöka hur bakteriearterna tillsammans påverkar responsvariablerna. Först anpassades multipel logistisk regression på **Övervikt** och de fem bakteriearterna vilket resulterade i att koefficienten för *S. aureus* blev signifikant med 5% felrisk. När vi sedan anpassade den multipla linjära modellen erhöll vi samma resultat, nämligen att koefficienten för *S. aureus* ensamt blev signifikant med 5% felrisk. Trots att t-testen, där vi jämförde arternas förekomst med **BMI**, visade signifikanta resultat för *Laktobacillerna* var det istället *S. aureus* som blev signifikant i den linjära modellen. Detta kan ha att göra med den ökade felrisken i och med multipla test. Att *Laktobaciller* blev signifikant på 10% felrisk i t-testen kan vara ett sådant fall av multipla test som påvisar falska signifikanter eftersom felrisken ökar med antalet jämförelser som görs.

Slutsatsen av våra olika tester och modeller visar ett återkommande resultat att bakteriearten *S. aureus* verkar ha ett inflytande på BMI, ett negativt sådant. Förekomsten av *S. aureus* i tarmfloran verkar alltså bidra till ett lägre BMI och minskad risk för övervikt.

5 Diskussion

En viktig aspekt i detta arbete är att data vi använt oss av är insamlad i syftet att undersöka allergiutvecklingen hos barn. Vi har använt oss av bakteriearterna som valts ut i allergistudien för att se hur de kan påverka BMI. Kanske skulle andra bakteriearter i tarmfloran kunna vara av intresse att undersöka för att hitta samband med BMI. Med tanke på att det finns hundratals olika arter i tarmfloran skulle det rimligtvis kunna finnas andra arter att studera vidare. Men vi kan åtminstone dra slutsatser för de fem bakteriearter som har undersökts i detta arbete och i allergistudien, vilket påvisar samband mellan BMI och arten *S. aureus*.

Om vi fick designa studien annorlunda, för syftet att undersöka bakteriearternas påverkan på BMI, hade vi kunnat välja ett större antal överviktiga barn. I denna uppsats har andelen överviktiga barn varit relativt liten och vi hade hellre föredragit två ungefär lika stora grupper av normalviktiga och överviktiga barn för att kunna dra säkrare slutsatser i arbetet.

BMI är uppätt hos barn vid 2, 5 och 10 års ålder, och eftersom det händer mycket i ett barns kroppsliga utveckling vid de låga åldrarna kan det vara svårt att avgöra om ett barn kommer vara överviktig vid vuxen ålder. Det är en av anledningarna till att vi fokuserar på barns BMI vid 10 års ålder, då det är den äldsta ålder som har undersökts. Antalet barn är relativt få och eftersom det saknas en del observationer reduceras antalet än mer. En större urvalsgrupp skulle också kunna göra resultaten mer säkra.

Vi kunde inte anta någon korrelation mellan bakteriearterna och vi har också antagit att arterna inte samspelar med varandra. Rimligtvis borde det finnas samspel emellan arterna och kanske även andra bakteriearter (som inte nämns i denna uppsats) så att två, eller flera, arters samspel tillsammans påverkar responsvariabeln. På grund av få observationer och ett högt antal möjliga samspel har vi valt att inte ta hänsyn till detta i denna uppsats.

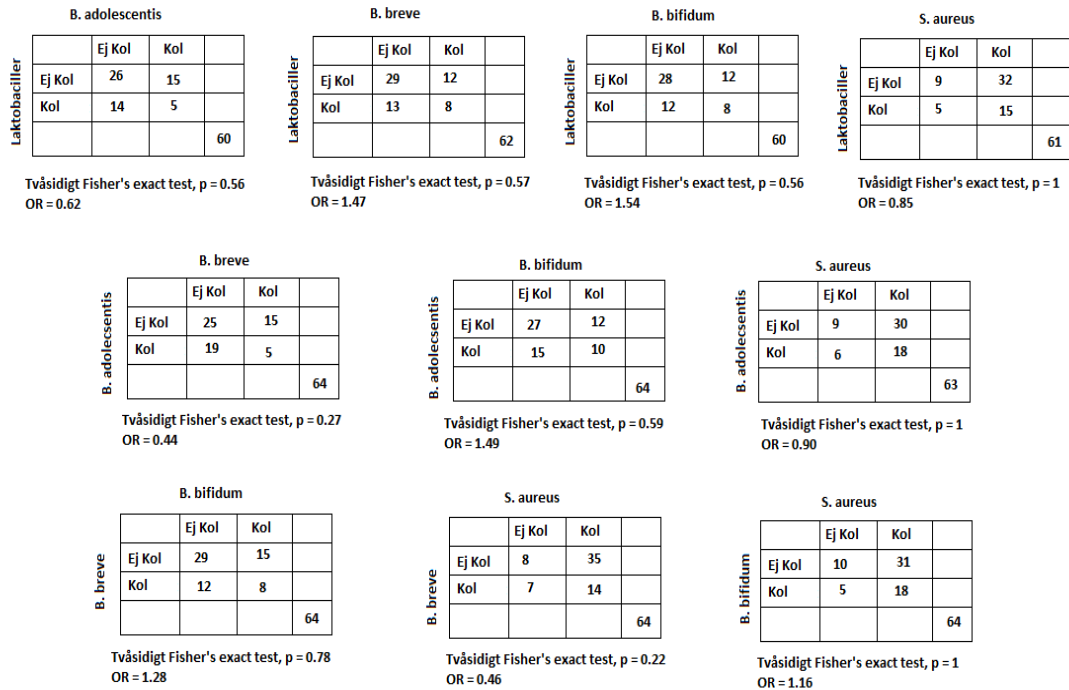
Vi nämnde i tidigare avsnitt att vid ett flertal jämförelser i olika test ökar risken för falska signifikanser. Men i och med att koefficienten för bakteriearten *S. aureus* är signifikant i de båda regressionsmodellerna för olika responsvariabler kan vi vara ganska övertygade om att det verkar finnas ett samband mellan artens förekomst och BMI.

Referenser

- [1] ALAN AGRESTI (2002). *Categorical Data Analysis*. Second Edition.
- [2] T.J. COLE, M.C. BELLIZZI, K.M. FLEGAL, W.H. DIETZ (2000). *Establishing a standard definition for child overweight and obesity worldwide: international survey*.
- [3] J. L. DEVORE OCH K.N.BERK (2012). *Modern Mathematical Statistics with Application*.
- [4] L. FAHRMEIR, T. KNEIB, S. LANG OCH B. MARX (2013). *Regression - Models, Methods and Application*.
- [5] O. HAGLUND (2009). *Bra tarmflora tidigt i livet kan förebygga fetma*. <http://www.doktorn.com/artikel/den-livsviktiga-tarmbakterien>
- [6] M.A. JOHANSSON, Y.M. SJÖGREN, J-O PERSSON, C. NILSSON OCH E. SVERREMARK-EKSTRÖM (2011). *Early Colonization with a Group of Lactobacilli Decreases the Risk for Allergy at Five Years of Age Despite Allergic Heredity*.
- [7] M.A. JOHANSSON, S. SAGHAFIAN-HEDENGREN, Y. HAILESELASSIE, S. ROOS, M. TROYE-BLOMBERG, C. NILSSON & E. SVERREMARK-EKSTRÖM (2012). *Early-Life Gut Bacteria Associate with IL-4, IL-10- and IFN- γ Production at Two Years of Age*.
- [8] Y. KHALID (2009). *Bakterierna - våra goda vänner*. Uppsats för forskarutbildningskursen: Aktuell klinisk forskning. Avdelningen för pediatrik. Sahlgrenska akademien, Göteborgs universitet. http://www.goteborgslakaresallskap.se/pdf/sem/Khalid_HT-08.pdf
- [9] N. NORDH (2011). *Bra tarmflora tidigt i livet kan förebygga fetma*. <http://www.vetenskaphalsa.se/bra-tarmflora-tidigt-i-livet-kan-forebygga-fetma/>
- [10] R. SUNDBERG (2015). *Lineära Statistiska Modeller*.

6 Appendix

Figur 10: 2×2-tabeller för bakteriearterna



Figur 11: 2×2-tabeller för bakteriearterna och Övervikt vid 2 och 5 års ålder

Laktobaciller		B. adolescentis		B. breve		B. bifidum		S. aureus	
Övervikt, 2 år		Kol	Ej Kol		Kol	Ej Kol		Kol	Ej Kol
	Övervikt	5	5		4	7		6	5
	Normal	15	36		16	38		18	36
				61			65		
Tvåsidigt Fisher's exact test, p = 0.27 OR = 2.4		Tvåsidigt Fisher's exact test, p = 1 OR = 0.87		Tvåsidigt Fisher's exact test, p = 0.73 OR = 1.36		Tvåsidigt Fisher's exact test, p = 0.30 OR = 7.4		Tvåsidigt Fisher's exact test, p = 0.05 OR = 0.24	
Laktobaciller		B. adolescentis		B. breve		B. bifidum		S. aureus	
Övervikt, 5 år		Kol	Ej Kol		Kol	Ej Kol		Kol	Ej Kol
	Övervikt	3	5		2	6		2	6
	Normal	16	34		17	37		20	34
				58			62		
Tvåsidigt Fisher's exact test, p = 1 OR = 1.28		Tvåsidigt Fisher's exact test, p = 1 OR = 0.8		Tvåsidigt Fisher's exact test, p = 1 OR = 0.73		Tvåsidigt Fisher's exact test, p = 0.70 OR = 0.57		Tvåsidigt Fisher's exact test, p = 0.20 OR = 0.33	

Tabell 11: Sammanfattning av skattningar för den logistiska modellen vid 2 och 5 års ålder.

Modell 1	Förkl.var.	β_i skattning	$\exp(\hat{\beta}_i)$	$\Pr(> z)$	Antal
Ålder 2	intercept	-0.95	0.39	0.28	57
	Kol _{Lakto}	1.14	3.14	0.14	
	Kol _{B.adol}	-0.60	0.55	0.51	
	Kol _{B.brev}	0.02	1.02	0.96	
	Kol _{B.bifi}	0.81	2.24	0.30	
	Kol _{S.aur}	-1.74	0.17	0.03	
Ålder 5	intercept	-0.44	0.64	0.64	54
	Kol _{Lakto}	0.67	1.94	0.45	
	Kol _{B.adol}	-0.64	0.53	0.54	
	Kol _{B.brev}	-0.88	2.42	0.40	
	Kol _{B.bifi}	-0.27	0.76	0.78	
	Kol _{S.aur}	-1.87	0.15	0.04	

Tabell 12: Resultat från Hosmer-Lemeshow's test vid 2 och 5 års ålder.

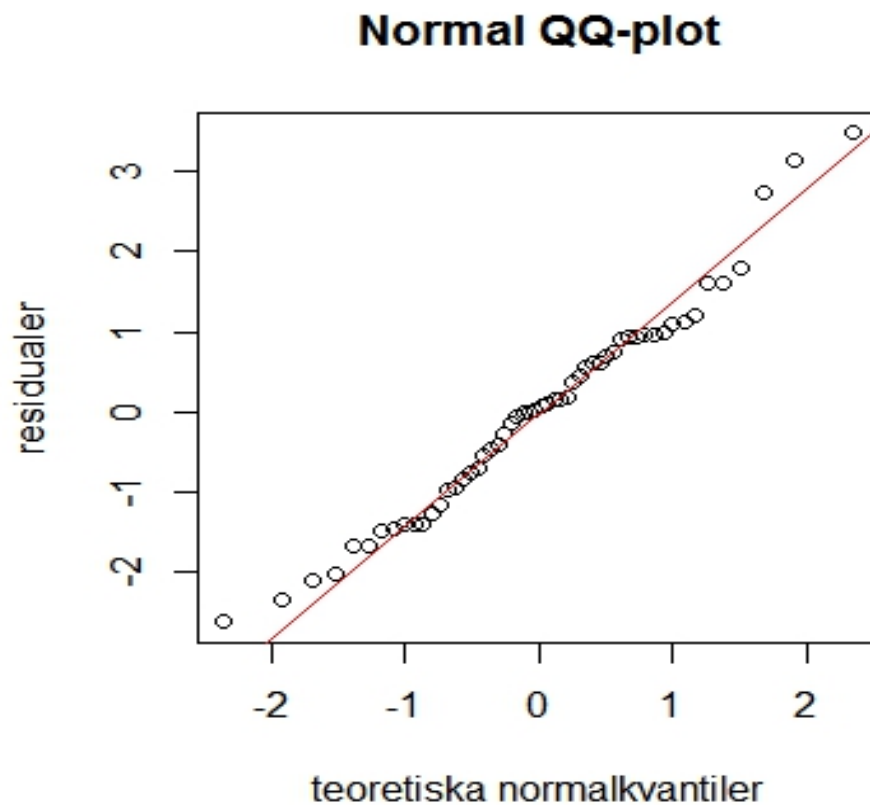
Ålder	$\hat{C}$	frihetsgrader	p-värde
2	7.481	5	0.19
5	5.467	5	0.36

För modell 2 vid åldrarna 2 och 5 år är responsvariabeln **Avvikelse** istället för BMI, som det är vid 10 års ålder.

Tabell 13: Skattningar för den linjära modellen vid 2 och 5 år.

Modell 2	Förkl.var.	β_i skattning	$\Pr(> z)$	Antal	R^2
Ålder 2	intercept	-0.62	0.16	57	0.16
	Kol _{Lakto}	0.10	0.78		
	Kol _{B.adol}	-0.53	0.13		
	Kol _{B.brev}	0.26	0.47		
	Kol _{B.bifi}	0.30	0.38		
	Kol _{S.aur}	-0.83	0.04		
Ålder 5	intercept	-0.42	0.42	54	0.07
	Kol _{Lakto}	-0.33	0.44		
	Kol _{B.adol}	-0.20	0.64		
	Kol _{B.brev}	0.08	0.86		
	Kol _{B.bifi}	-0.02	0.96		
	Kol _{S.aur}	-0.75	0.10		

Figur 12: Normalfördelnings-plott, Modell 2 vid 5 års ålder



Figur 13: Normalfördelnings-plott, Modell 2 vid 2 års ålder

