**Anders Milhøj**

**En hurtig opsamling af elementær statistik**

**Sommerskolen 2018**

*Denne note indeholder en kort repetition af indholdet i et indledende statistikkursus. Erfaringen siger, at studerende med selv omfattende baggrund i statistik/økonometrikurser hurtigt glemmer de kritiske punkter i pensum, heraf behovet for en opsamlingsnote af kerneindholdet. Noten er tilstræbt at være intuitiv, så den sådan set kan læses ud fra blot gymnasiekundskaber.*

Indhold

[Statistiske modeller med binomialfordelingen 2](#_Toc518540081)

[Poissonfordelingen 7](#_Toc518540082)

[Uafhængighedstest 8](#_Toc518540083)

[Stikprøver 11](#_Toc518540084)

[Normalfordelingen 13](#_Toc518540085)

[Sammenligning af to normalfordelinger 16](#_Toc518540086)

[Regressionsanalyse 16](#_Toc518540087)

[Multipel regressionsanalyse 23](#_Toc518540088)

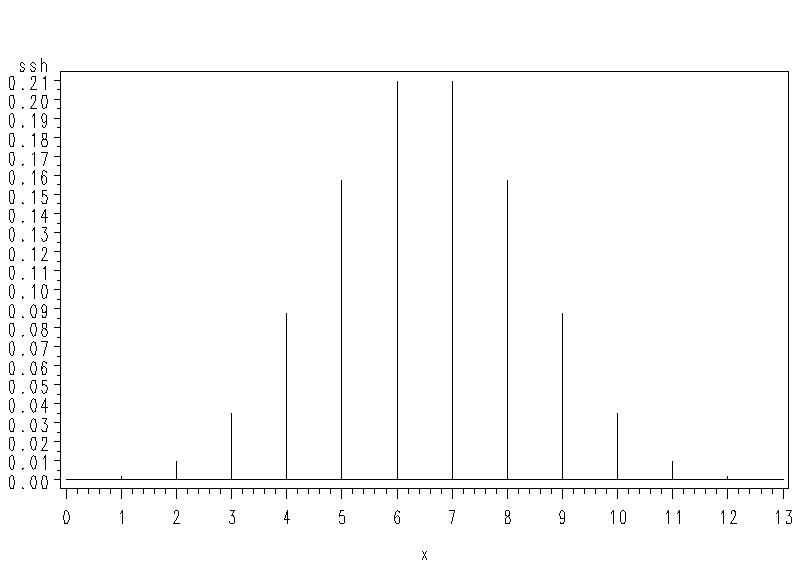
[Variabeludvælgelse 28](#_Toc518540089)

# Statistiske modeller med binomialfordelingen

Hvis en student prøver 13 gange, om hun kan smage forskel på Tuborg og Carlsberg (eller skal vi sige Coca Cola Light og Coca Cola Zero for at styrke ædrueligheden) og hun har samme sandsynlighed p for at gætte/smage rigtigt, er antallet af rigtige svar binomialfordelt.

Her opfattes sandsynligheden p som en parameter vi ikke kender. Den udtrykker den sikkerhed, hvormed hun rent faktisk kan smage rigtigt. Værdien p = 0.5 svarer til at gætte, mens værdier p > 0.5 betyder at hun med en vis sikkerhed kan vurdere rigtigt. Dette er en statistisk model, der binder et tilfældighedsmoment (hun vurderer kun med en vis sikkerhed rigtigt, så der er et usikkerhedsmoment tilbage) og en parameter, her sandsynligheden p, der udtrykker det, vi egentlig er interesseret i at bestemme.

Binomialfordelingen for p = 0.5 ses på figuren.



Hvis nu 9 dvs 69,2% af svarene er rigtige – er det så et udtryk for, at hun rent faktisk kan smage forskel, eller er det blot en tilfældighed?

Sandsynligheden for at der er 9 ELLER flere rigtige valg i de 13 forsøg, er



For p = 0.5 er denne sandsynlighed lig med 13,3%. Det betyder, at hvis forsøget gentages for en lang række studenter, hvor alle blot gætter, vil man for 13,3% af studenterne se 9 eller flere rigtige svar. Havde derimod 11 svar været rigtige, ville den tilsvarende sandsynlighed kun have været 1,1%. Ud fra disse tal vil man sædvanligvis slutte, at kun 9 rigtige godt kan være et udslag af tilfældigheder, for det sker ret tit, mens hele 11 rigtige til gengæld er så stort, at det næppe kan være et udslag af tilfældigheder.

Der er nu udført en statistisk analyse med et statistisk test! Tallene 13.3% og 1.1% er p-værdier (signifikanssandsynligheder) i de to situationer. Normalt vil man teste på et 5% niveau, således at signifikanssandsynligheder større end 5% leder til accept af hypotesen - her hypotesen om p = 0.5, dvs at det gættes. Signifikanssandsynligheder mindre end 5% leder til forkastelse. Grænsen ved 5%, testets signifikansniveau, bør varieres, så det sættes lavere, når der er mange observationer og højere, når der kun er få observationer. I dette tilfælde med kun 13 studenter burde signifikansniveauet nok være sat til 10%.

Ganges tallene med 10 får man følgende billede. Her er der 130 forsøg, og det ses, at 90 rigtige er vildt usandsynligt, hvis der gættes. Grænsen mellem gæt og viden vurderet ud fra et 5% niveau går ved lige omkring 75 rigtige. Det er altså lettere at skelne mellem gætteri og viden i en multiple choise prøve med mange spørgsmål end i prøver med kun få spørgsmål.



Formelt ville man sige, at hypotesen om p = 0.5 testes ved at vurdere afvigelsen mellem den observerede hyppighed og værdien p = 0.5. Det gøres ved



idet nævneren er standardafvigelsen (kvadratroden af variansen) i en binomialfordeling med sandsynlighed p. Denne tilsvarende varians (standardafvigelsen kvadreret) er p(1-p) når n = 1, og variansen på en empirisk hyppighed i n forsøg findes ved at dividere denne varians med n. Teststørrelsens fordeling kan tilnærmes ved en normalfordeling, når n er stor nok, fx over 30, og p ikke er for tæt ved 0 eller 1.

Læg mærke til at teststørrelsen u vokser proportionalt med kvadratroden af antal observationer, . For den samme observerede forskel mellem hyppighed, og den hypotetiske sandsynlighed, her 0.5, vil teststørrelsen u altså være numerisk lille for små værdier af n, men vil vokse vilkårligt stor for store værdier af n. Derved vil der altid kunne opnås signifikant afvisning af hypotesen, blot n er stor nok. Derved vil selv ubetydelige forskelle blive signifikante i store stikprøver. Det er en del af forklaringen på, at man bør vælge at sætte signifikansniveauet lavere end de 5% i store datasæt.

Den næste figur viser kønsfordelingen, målt ved andel piger, ved alle fødsler i hver af landets 98 kommuner - vist nok i 2009.



Desuden vises størrelsen u, beregnet for hypotesen p = 0.5, se næste figur.



De mest ekstreme kønsfordelinger er

|  | **Andel piger** |  |  | **Andel piger** |
| --- | --- | --- | --- | --- |
| Fanø | 0.416667 |  | Læsø | 0.545455 |
| Roskilde | 0.428571 |  | Solrød | 0.546296 |
| Dragør | 0.433333 |  | Lemvig | 0.558685 |
| Hvidovre | 0.441739 |  | Tårnby | 0.563380 |
| Glostrup | 0.442623 |  | Ærø | 0.638889 |

I den standardiserede normalfordeling for teststørrelsen u svarer værdier numerisk større end 2 til afvigelser, der er bemærkelsesværdige ved et 5% test mod et dobbeltsidet alternativ. Figurerne viser, at der er 4 kommuner med væsentligt under 50% pigefødsler og kun enkelt, Ærø, med væsentlig flere end 50% pigefødsler. Det ses dog også en klar tendens til, at fordelingen ligger forskudt lidt mod venstre for nul svarende til det velkendte fænomen, at der fødes lidt færre piger end drenge. Men da antallet af fødsler ikke så stort, er det kun i få kommuner, at hypotesen om ligelig kønsfordeling afvises - og der er endda en kommune, hvor hypotesen afvises pga for mange pigefødsler, dvs den "forkerte" vej.

For regionerne får man følgende tabel, der viser en signifikant lavere hyppighed af pigefødsler i de fleste regioner og en klar signifikans for hele landet. Her betragtes langt flere fødsler, så signifikans på 5% lettere opnås, selvom den observerede hyppighed for en pigefødsel, der historisk set er 48.8% jo ikke er langt fra 0.5.

|  | **Drenge** | **Piger** | **pct** | **u** |
| --- | --- | --- | --- | --- |
| Region Syddanmark | 6781 | 6439 | 0.48707 | -2.97448 |
| Region Sjælland | 4424 | 4102 | 0.48112 | -3.48725 |
| Region Nordjylland | 3160 | 2965 | 0.48408 | -2.49162 |
| Region Midtjylland | 7516 | 7355 | 0.49459 | -1.32025 |
| Region Hovedstaden | 10934 | 10406 | 0.48763 | -3.61440 |
| **Hele landet** | **32815** | **31267** | **0.48792** | **-6.11509** |

# Poissonfordelingen

Fordelingen, hvor punktsandsynligheden er

 x = 0,1,2,3,4,…….(uden øvre grænse)

kaldes Poissonfordelingen. Den kan opnås som tilnærmelse til binomialfordelingen ved rent matematisk at se, hvad der sker, hvis p er lille og n er stor, idet np holdes tæt ved λ. Det er en grænseovergang, der kan bevises på gymnasieniveau.

Poissonfordelingen er brugbar, når der behandles mange uafhængige hændelser (n stor) med små sandsynligheder (p = λ/n lille). Den bruges desuden ofte som udgangspunkt for en statistisk model, hvor egenskaberne ”uafhængige hændelser”, ”mange hændelser” og ”samme lille sandsynlighed” kan antages, uden nødvendigvis at være fuldstændigt sande. Et andet eksempel er begivenheder, der af og til indtræffer. Fx antal opkald til vagtlægen i Nordjylland i en række fem minutters intervaller en mandag eftermiddag. Der er mange mennesker i Nordjylland, og hver af dem har en meget lille sandsynlighed for at ringe til vagtlægen i et fem-minutters interval. Derfor benyttes en Poissonfordeling.

Et praktisk forsøg på gymnasieniveau er at betragte en klump radioaktivt materiale. Her er der et meget stort antal atomer (n styk) og hvert af disse har en meget lille sandsynlighed for at henfalde (så de opfattes af Geigertælleren) indenfor 10 sekunder (sandsynlighed p). Antal målte henfald på 10 sekunder kan altså betragtes som en Poissonfordeling.

Et sådant materiale om baggrundsstrålingen for 27 observationer i 10-sekundersintervaller giver følgende observationer, der har gennemsnit 3.6667, giver

2 2 3 4 3 4 1 5 7 3 3 3 3 1 2 6 7 9 5 1 2 4 5 4 3 4 3

Figurerne viser data sammenholdt med Poissonapproksimationen for λ = 3.6667





# Uafhængighedstest

I en TV Avis (vist nok 16/11 - 2008) sammenholdt man hyppigheden af salmonella i 20 pakker dansk kød med hyppigheden i 20 pakker udenlandsk kød. Af de danske pakker var 4 salmonellainficerede, mens 17 var inficerede blandt de 20 pakker udenlandsk kød. Spørgsmålet er, om denne forskel i hyppighed er signifikant. Opstillet i en tabel er det

|  | **salmonella** | | |
| --- | --- | --- | --- |
|  | **Ja** | **Nej** | **Total** |
| **Dansk** | 4 | 16 | **20** |
| **Udenlandsk** | 17 | 3 | **20** |
| **Total** | **21** | **19** | **40** |

Her er spørgsmålet ikke, hvor hyppig salmonella egentlig er, men kun om hvordan hyppigheden er for dansk contra udenlandsk kød. Spørgsmålet er heller ikke, hvor meget kød der dansk hhv. udenlandsk, for data er jo indsamlet, så der er lige mange af hver mulighed. I dette tilfælde er spørgsmålet om salmonella sandsynligheden 4/20 for dansk kød er væsentlig forskellig fra 17/20 for udenlandsk kød.

I datamaterialet er den samlede hyppighed for salmonella 21/40, mens dansk og udenlandsk kød er ligeligt repræsenteret. Havde der været uafhængighed i salmonella forekomsten i dansk og udenlandsk kød ville sandsynligheden for fx salmonella i dansk kød været 21/40 og da der i alt er 20 pakker dansk kød ville man derfor forvente at se 10.5 tilfælde af salmonella i dansk kød i dette dataeksempel med i alt 40 pakker. Det tal skal sammenholdes med de observerede 4 tilfælde. Tilsvarende for de andre celler i tabellen, hvor der ved en ligelig fordeling af de 21 salmonellatilfælde ville forventes 10.5 pakker inficeret udenlandsk kød og 9.5 uinficerede pakker af både dansk og udenlandsk kød.

Forskellene mellem de forventede antal og de rent faktisk observerede antal kan samles til en teststørrelse ved



der i dette tilfælde beregnes til 16.94. Dette tal skal sammenholdes med en χ2-fordeling med en frihedsgrad. Dette er den klassiske teststørrelse, der let kan beregnes uden brug af IT.

Som et alternativ til denne teststørrelse kan man bruge det eksakte likelihood ratio test, der er givet ved



Denne teststørrelse er anvendt i den nye udgave af faget "Statistik og Sandsynlighedsregning" på andet år af politstudiet. Talværdien er 18.43, så konklusionen er den samme.

Disse teststørrelser, hvor observerede antal sammenholdes med forventede, kan benyttes i mange sammenhænge. Det ses, at den er proportional med antal observationer, således at der også her gælder at signifikans på 5% kan opnås for selv ubetydelige afvigelser fra uafhængighed, når blot datasættet er stort nok.

At der kun er én frihedsgrad for en tabel med fire celler, skyldes at rækkesummerne og søjlesummerne alle er fastsat uden usikkerhed. Det skyldes, at forsøgsplanlægningen har bestemt, at der 20 prøver fra hver oprindelsessted, og salmonellaforekomsten i verdenen har fastsat, at der i alt er 21 ud af de 40 prøver, der har salmonella - spørgsmålet er altså kun, hvordan de inficerede kødpakker er fordelt. Når rækkesummer og søjlesummer er opgivet på forhånd, er der kun et enkelt tal, der kan flyttes i tabellen, for alle andre tal i tabellen "flytter med". Deraf den ene frihedsgrad.

Fordelingen stammer fra, at en standardiseret normalfordeling kvadreret følger en χ2-fordeling med en frihedsgrad. Værdier over knapt fire svarer til signifikante afvigelser fra uafhængigheden.

Et andet eksempel er kønsfordelingen ved fødslerne i Frederiksberg og Københavns kommuner.

|  | **Drenge** | **Piger** | **Total** |
| --- | --- | --- | --- |
| **Frederiksberg** | 765 | 765 | **1530** |
| **København** | 4370 | 4024 | **8394** |
| **Total** | **5135** | **4789** | **9924** |

Selvom Frederiksberg er lidt atypisk ved trods de mange fødsler at have en helt lige kønsfordeling, har teststørrelsen værdien 2.20, hvilket med p = 13.8% ikke er signifikant.

For alle de danske regioner får man tilsvarende tabellen

|  | **Drenge** | **Piger** | **Total** |
| --- | --- | --- | --- |
| **Region Hovedstaden** | 10934 | 10406 | **21340** |
| **Region Midtjylland** | 7516 | 7355 | **14871** |
| **Region Nordjylland** | 3160 | 2965 | **6125** |
| **Region Sjælland** | 4424 | 4102 | **8526** |
| **Region Syddanmark** | 6781 | 6439 | **13220** |
| **Total** | **32815** | **31267** | **64082** |

I denne tabel er teststørrelsen 4.63. Her der fire frihedsgrader. Det skyldes, at antal fødsler i de enkelte regioner antages givet - det eneste interessante er kønsfordelingen. Desuden antages, at kønsfordelingen i landet som helhed er givet. Derfor er række- og søjlesummerne igen faste, så der kun kan varieres på fire tal blandt de 10 tal i 5 × 2 tabellen.

# Stikprøver

I en stikprøve udtager man en stikprøve på n individer ud af en totalpopulation på N individer. Normalt udtager man uden tilbagelægning, dvs at samme person ikke kan undersøges to gange, hvilket besværliggør alle usikkerhedsberegninger. Men hvis N er stor i forhold til n, er der ingen praktisk forskel.

Hvis man vil undersøge hyppigheden af en egenskab i populationen bruger man selvfølgelig den tilsvarende hyppighed i stikprøven som estimat. Variansen herpå er, når man tager højde for at der udtages uden tilbagelægning,



hvor faktoren  skyldes, at der ikke er tilbagelægning.

Ud fra den tilsvarende standardafvigelse, får man konfidensintervaller for en stor population, N = ∞, ud fra følgende formel

.

Tabellen giver nogle eksempler

|  |  |  |
| --- | --- | --- |
| stikprøvestørrelse, n | hyppighed, p | interval som |
| 1000 | 50% | ± 3.16 procentpoint |
| 1000 | 5% eller 95% | ± 1.38 procentpoint |
| 500 | 20% eller 80% | ± 3.58 procentpoint |
| 100 | 50% | ± 10 procentpoint |
| 100 | 5% eller 95% | ± 4.36 procentpoint |

For stikprøver på 1000 personer er usikkerheden altså af størrelsesordnen ± 3 procentpoint, hvilket er forbløffende præcist. For mindre stikprøver er usikkerheden selvfølgelig større. For hyppigheder nær nul og en, er præcisionen større målt ved usikkerheden i procentpoint, men den relative usikkerhed i forhold til den hyppighed, der skal måles, er selvfølgelig større.

Ofte kan man opdele populationen i homogene delpopulationer, kaldet strata. Fx er befolkningen jo inddelt i mænd i kvinder. Hvis man antager, at 20% af kvinderne interesserer sig for fodbold, og 80% af mændene af mændene gør det, er det for befolkningen som helhed 50%, der gør det. Man kan så med fordel indsamle en stikprøve på 500 kvinder og 500 mænd i stedet for 1000 personer med uspecificeret køn. Dermed risikerer man ikke skævheder pga. af en tilfældigt opstået skæv kønsfordeling i stikprøven, da den risiko er elimineret ved den præcist jævne kønsfordeling, så estimatoren for den faktiske andel i den samlede befolkning bliver mere præcist.

Visse skævheder kan tilsvarende rettes op ved vægtning, fx kan en skæv kønsfordeling rettes op ved at tage et ligeligt gennemsnit af mænds og kvinders hyppighed uanset hvor mange mænd og kvinder, der rent faktisk er i stikprøven. Individerne fra det underrepræsenterede køn i stikprøven får derved større vægt i beregningen.

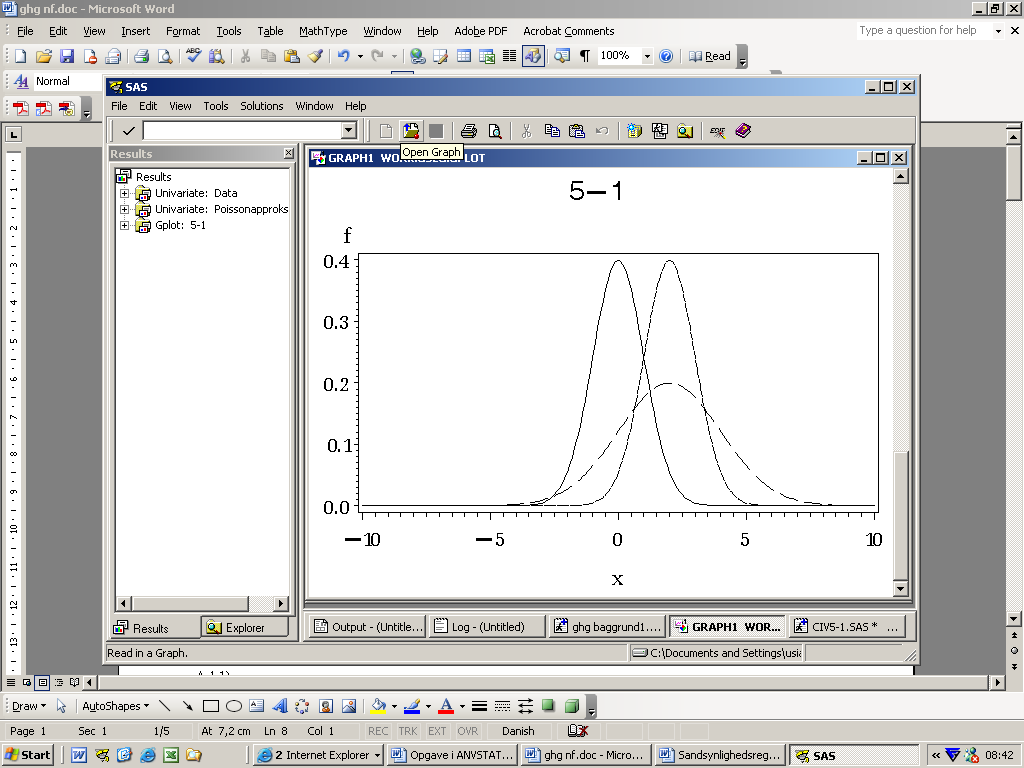
I praksis skyldes stikprøveusikkerhed mange andre fejlkilder, idet det er meget svært at indsamle stikprøver på korrekt vis. Alle danskere kan ikke nås pr. telefon og mange nægter at svare, stikprøver indsamlet på gaden er ikke repræsentative for den samlede befolkning etc.

# Normalfordelingen

I en normalfordeling med middelværdi μ og varians σ2 beregnes sandsynlighederne ud fra

P(X ≤ x) = 

Den funktion, der integreres, kaldes tæthedsfunktionen. Figuren viser tre eksempler med hhv μ = 0 σ2 = 1, μ = 2 σ2 = 1 og μ = 2 σ2 = 2.



Denne definition rejser jo en del spørgsmål:

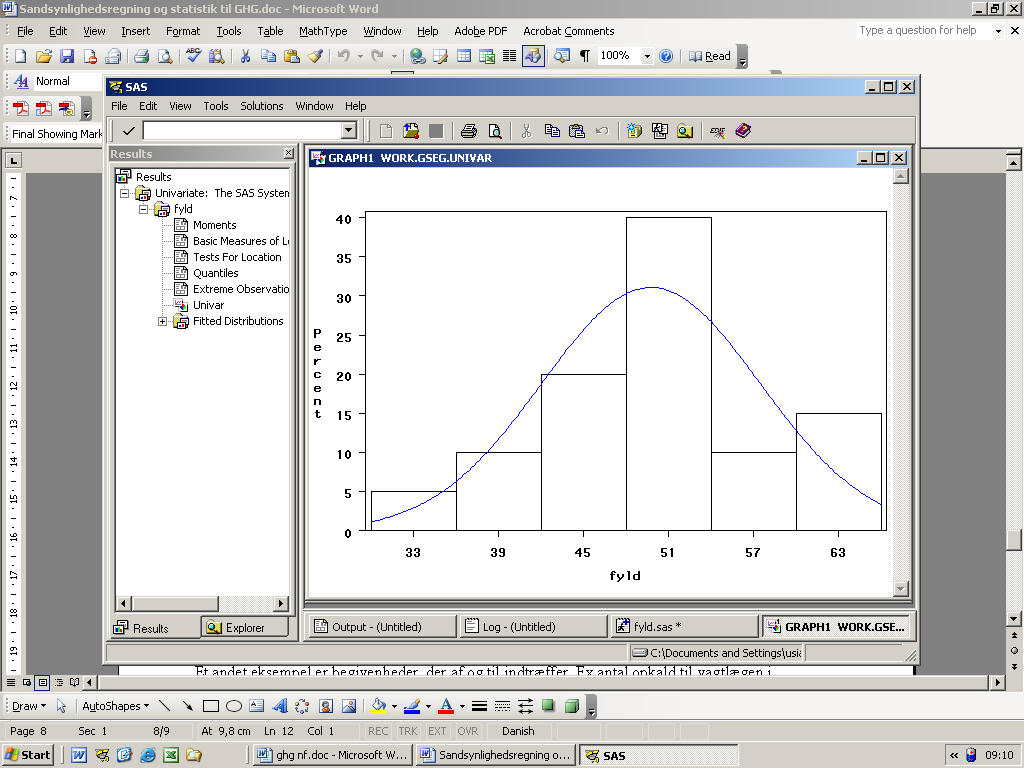
1. Hvad er et integral?
2. Hvad vil det sige, at nedre grænse i integralet er minus uendelig?
3. Kan man overhovedet finde en stamfunktion denne funktion?
4. Hvad betyder middelværdi og varians?
5. Hvad "normalt" er der dog ved den fordeling?

etc….

Men disse spørgsmål er sådan set af underordnet betydning i forhold til anvendelser af normalfordelingen, så det er næsten klogest at forbigå dem i tavshed!

Et typisk eksempel på anvendelse af normalfordelingen er måleusikkerhed. Her betragtes 20 observationer af afstanden mellem en nyåbnet ølflaskes top og indholdets overflade. Bryggeriet tilstræber, at denne afstand skal være 50 mm.

Gennemsnittet er  = 49,75, og variansen er s2 = 59,36. Observationerne er vist i et histogram med en indtegnet normalfordelingskurve. Det er muligt at efterprøve, om normalfordelingen kan antages at beskrive datas fordeling, f.eks. på gammeldags vis ved hjælp af sandsynlighedspapir eller mere moderne ved et computertegnet fraktildiagram, men det forbigås her.



Gennemsnittet kan opfattes om normalfordelt med en varians, der kan tilnærmes til 59,36/20 = 2,97. Ofte vil man udtrykke denne fordeling i form af et konfidensinterval på formen

,

der afgrænser et interval med sandsynlighed 95% i fordelingen af gennemsnittet. Slå selv efter at der gælder Φ(-1.96) = 0.025! I dette tilfælde får man



Spørgsmålet er så, om middelværdien kan antages at være 50 og ikke kun 49,75, som data jo faktisk viste. Hvis gennemsnittet kan opfattes om en observation af en normalfordeling med middelværdi 50 og varians 2,97, er den observerede værdi 49,75 så usædvanligt lille? Nej, for beregnet i normalfordelingen med middelværdi 50 og varians 2,97 er

P(= 44%

Afvigelsen mellem det observerede gennemsnit er så lille, at der med en sandsynlighed på 44% kan forekomme større afvigelser (endda med sandsynligheden 88%, hvis afvigelser i den modsatte retning regnes med). Så tappemaskinen kan godt antages at være justeret rigtigt, så den giver en afstand på 50 mm.

Bemærk at dette test bygger på, at variansen på gennemsnittet kendes uden usikkerhed – der er jo blot indsat en varians på 2,97. Faktisk er denne varians estimeret ud fra data, hvilket påvirker sandsynlighedsberegningen. Forskellen er dog ubetydelig, men håndteringen af den udgør begyndelsen på vigtige dele af statistikken, så dette test være udføres oftest på en lidt mere kompliceret måde, med en t-fordeling og beregning af antal frihedsgrader.

Estimatet for variansen er



og antallet af frihedsgrader er n - 1, da de n afvigelser fra gennemsnittet ikke varierer helt frit, fordi

,

hvilket giver et enkelt bånd svarende til den ene frihedsgrad, der tabes.

At teststørrelsen ikke er normalfordelt, men i stedet t-fordelt, skyldes, at nævneren, hvor kvadratroden af variansen indgår, ikke er et fast tal, men også er påvirket af estimationsusikkerhed. Derved bliver afvigelser fra middelværdien større, end hvis variansen havde været et fast tal. At det så netop er t-fordelingen er vanskeligt at bevise på gymnasieniveau. Forskellen er ubetydelig i tilfælde med flere end ganske få observationer, i eksemplet med 20 observationer ændrer sandsynligheden på 44% sig ikke, medmindre der medtages decimaler. Tallet 1.96 i konfidensintervallet ændres til 2.09 - i praksis kan man altså roligt bruge tallet 2 uden at tabe væsentlig præcision.

Sammenligning af to normalfordelinger

Antag at der for to sæt observationer fra hver sin normalfordeling og at alle observationer er uafhængige. Man vil så især interessere sig for om middelværdierne er ens, hvilket jo er sammenligning af de to gennemsnit  og . Ved at dividere med et estimat for standardafvigelsen for forskellene mellem disse gennemsnit får man teststørrelsen



Den kan sædvanligvis opfattes som approksimativt standardiseret normalfordelt. Et eksakt fordelingsresultat kan kun opnås, hvis de to varianser er kendte på forhånd, eller hvis de to varianser antages ens, så deres fælles værdi kan estimeres. Resultatet er i det sidste tilfælde igen en t-fordeling med n1 + n2 - 2 frihedsgrader, da der estimeres to gennemsnit undervejs.

I visse tilfælde passer observationerne i de to normalfordelinger sammen parvis. Fx hvis man vejer samme person før og efter en slankekur. Ved at teste på den måde, der blev beskrevet ovenfor, får man ikke inddraget den ekstra information at samme person indgår i to gange i data. I stedet er det oplagt at betragte differenserne, altså vægttabene, som en enkelt række observationer, når effekten af slankekuren skal beskrives, i stedet for sammenligne gennemsnit af start- og slutvægte for en række tykke og tynde personer. Disse to gennemsnit er selvfølgelig upræcise pga. den biologiske variation, så sammenligningen bliver upræcis af den grund. Hvis man skal undersøge, om slankekuren overhovedet har en effekt, skal man altså teste om gennemsnittet af differenserne er nul og på den måde se bort fra den biologiske variation, der jo er elimineret ved differensdannelsen.

# Regressionsanalyse

I regressionsanalysen sammenholder man to variable, idet man opfatter x-variablen som faste givne tal, mens y variablen varierer omkring en ret linie. Nedenfor ses et eksempel med højde og vægt for en række skolebørn – data er datasættet CLASS, der hænger de fleste SAS-brugere langt ud af halsen! I det konkrete tilfælde kan man selvfølgelig diskutere, hvad der er faste tal, og hvad der varierer - det er også muligt at modellere begge størrelser som varierende ved hjælp af korrelationer. Det er dog langt det enkleste at udføre regressionsanalyser, og blot man husker, at man derved undersøger for sammenhæng og ikke nødvendigvis for kausalitet, begås der ingen fejl. Det skal blot huskes, at de estimerede regressionskoefficienter er parametre i den betingede fordeling af responsvariablen givet de forklarende variable.



Ved at opstille en statistisk model af formen

yi = α + βxi + ei

hvor restleddet ei sædvanligvis antages at være normalfordelte med samme varians (kaldet σ2) og uafhængige.

Der estimeres ved mindste kvadraters metode, dvs at linien bestemmes ved at minimere summen af de kvadrerede lodrette afstande.



Dette problem kan klares ved at differentiere partielt efter α og β, hvorved eksplicitte formler kan opstilles. Hældningen estimeres ved

 ,

hvis varians er

,

hvor s2 er estimatet



for restledsvariansen σ2.

Derfor kan hypotesen β = 0 estimeres ved teststørrelsen



der er t-fordelt med n - 2 frihedsgrader. At der tabes to frihedsgrader ud af de oprindelige n uafhængige observationer, skyldes, at der estimeres to parametre, α og β. Derved opfylder residualerne de to restriktioner



så der er kun n - 2 frit varierende tal, der kvadreres i udtrykket for s2.

I SAS køres:

ods graphics/imagemap=on;

**proc** **reg** data=sashelp.class plots=FitPlot;

model weight=height;

id name;

**run**;

Hvis hypotesen om, at hældningskoefficienten β er lig med et andet tal end nul, fx β = 17, bliver teststørrelsen

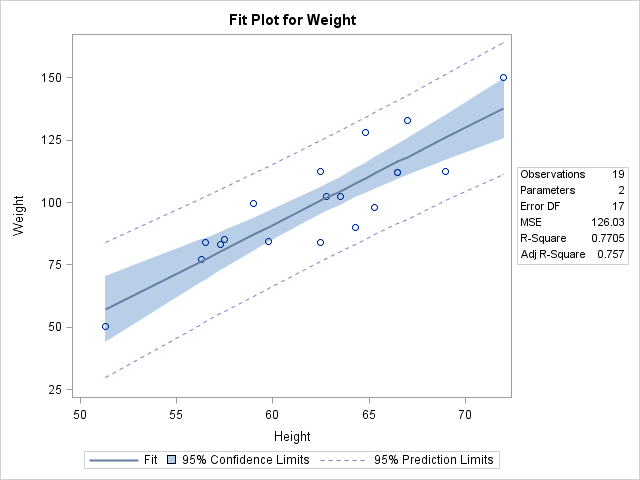


I SAS gøres det ved at tilføje en linie med en TEST-statement:

test height=**17**;

Den fittede regressionsmodel vises i et plot, der specificeres ved optionen plots=FitPlot. Der kan genereres mange flere plots ved fx plots=all, men de forbigås her. Pga ordren ods graphics/imagemap=on og id name er det muligt at se oplysninger om de enkelte datapunkter på grafen i Resultsvinduet, når musen placeres på et datapunkt.

På grafen angiver det snævre markerede område 95% sikkerhedsgrænser for selve den estimerede linie, dvs den usikkerhed, der skyldes, at de eksakte værdier af regressionskoefficienterne, α og β ikke kendes eksakt. De to stiplede linier er 95% sikkerhedsgrænser for en værdi af y-variablen, her vægten, når teenagerens højde kendes.



I tekst outputtet vises de estimerede parametre.

| *Parameter Estimates* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Variable* | *DF* | *Parameter Estimate* | *Standard Error* | *t Value* | *Pr > |t|* |
| *Intercept* | 1 | -143.02692 | 32.27459 | -4.43 | 0.0004 |
| *Height* | 1 | 3.89903 | 0.51609 | 7.55 | <.0001 |

Især er hældningskoefficienten, β, interessant. Den estimeres til 3.90 med en standardafvigelse på 0.52. Hypotesen om at denne parameter er nul, testes ved den oplyste t-teststørrelse, der blot er den estimerede værdi divideret med standardafvigelsen. Denne værdi, t = 7.55 og den tilsvarende signifikanssandsynlighed, der er meget lille, vises også i tabellen. Et 95% konfidensinterval konstrueres ved

3.90 ± 2.11 × 0.51 = [2.82, 4.98]

hvor 2.11 er 97.5% fraktilen i t-fordelingen med 19 - 2 = 17 frihedsgrader.

Variansanalysetabellen viser visse kvadratsummer, der indgår i beregningerne samt den estimerede residualvarians s2.

| *Analysis of Variance* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Source* | *DF* | *Sum of Squares* | *Mean Square* | *F Value* | *Pr > F* |
| *Model* | 1 | 7193.24912 | 7193.24912 | 57.08 | <.0001 |
| *Error* | 17 | 2142.48772 | 126.02869 |  |  |
| *Corrected Total* | 18 | 9335.73684 |  |  |  |

Den totale kvadratafvigelsessum 9335.74 er defineret ved

SAKy = SSDy = 

( SAK = Sum af Afvigelsernes Kvadrat, SSD = Sum of Squares of Deviations)

og angiver altså kvadratsummen af responsvariablen værdier, når der kun korrigeres for gennemsnittet af de observerede y-er. Den er spaltet op i et bidrag fra modellen på 7193.25 og et uforklaret bidrag på 2124.48.

Den uforklarede variation på 2124.48 er givet ved residualkvadratsummen



(RKS = Residual KvadratSum, RSS = Residual Sum of Squares)

I OLS estimationen forsøges dette bidrag altså minimeret, og de estimerede parameterværdier, , er indsat. Den estimerede residualvarians er så RKS divideret med frihedsantallet, der er n - 2 = 17, hvilket angives til 126.03 under betegnelsen Mean Square.

Modelbidraget 7193.25 er givet ved forskellen mellem disse to tal, men det kan også skrives som



således, at det angiver en kvadratafvigelse mellem den skrå linie, der er estimeret ud fra den lineære sammenhæng mellem x og y, og så en vandret linie, der er bestemt ud fra gennemsnittet af responsvariablens observerede værdier.

Determinationskoefficienten, der normalt kaldes R2 altså "R i anden", udtrykker, hvor stor en andel af den samlede variation, SAKy, der forklares ved hjælp af modellen. Denne del er altså modelkvadratafvigelsen SAKy - RKS.



Dette tal er vist i tabellen, der også angiver den estimerede residualstandardafvigelse 11.23, der er kvadratroden af residualvariansen



|  |  |  |  |
| --- | --- | --- | --- |
| *Root MSE* | 11.22625 | *R-Square* | 0.7705 |
| *Dependent Mean* | 100.02632 | *Adj R-Sq* | 0.7570 |
| *Coeff Var* | 11.22330 |  |  |

Med en tilsvarende notation kan estimatet for hældningskoefficienten skrives



hvor

SAPxy = SPDxy =

SAP = Sum af Afvigelsers Kvadrat og SPD = Sum of Products of Deviations

og

SAKx = SSDx = 

( SAK = Sum af Afvigelsernes Kvadrat, SSD = Sum of Squares of Deviations)

Heraf ses også ligheden - og ikke mindst forskellen - til den estimerede korrelationskoefficient mellem x og y



Målt i enheder, som en fysiker ville betragte formlen for hældningskoefficienten , har estimatet enheden vægt/højde, der i dette amerikanske datasæt er pund/tommer og ikke kg/cm, som vi danskere nok ville bruge. Men pointen er, at regressionshældningen overfører en højde til en vægt, hvilket afspejles i enheden. Korrelationskoefficienten  derimod er enhedsløs, da enhederne forkorter ud. Den er altså kun et mål for sammenhængen mellem de to variable, x og y, og ikke en del af funktion, der overfører x til y.

# Multipel regressionsanalyse

I dette eksempel kan der fortsættes ved a tilføje endnu en variabel, der angiver teenagerens køn. I datasættet sashelp.class er denne variabel angivet ved en bogstavsvariabel, sex, der ikke kan bruges i Proc Reg. Den skal derfor omdannes i et datastep, hvor der dannes en dummyvariabel med værdien 0 for drenge og 1 for piger. Denne nye dummyvaribel, gender, kaldes D i formlen

yi = α + βxi + γDi + ei

Parametrene estimeres igen ved mindste kvadraters metode, OLS, ved at minimere kvadratafvigelsessummen



I SAS køres

**data** a;

set sashelp.class;

if sex='M' then gender=**0**;

else gender=**1**;

run;

**proc** **reg** data=a plots=all;

model weight=gender height;

id name;

**run**;

I outputtet findes de estimerede parametre i tabellen

| *Parameter Estimates* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Variable* | *DF* | *Parameter Estimate* | *Standard Error* | *t Value* | *Pr > |t|* |
| *Intercept* | 1 | -126.16869 | 34.63520 | -3.64 | 0.0022 |
| *gender* | 1 | -6.62084 | 5.38870 | -1.23 | 0.2370 |
| *Height* | 1 | 3.67890 | 0.53917 | 6.82 | <.0001 |

Det ses, at koefficienten til dummyvariablen for køn estimeres til en negativ værdi, svarende til at piger vejer mindre end drenge; også selvom der korrigeres for forskelle i højde. Men værdien er ikke signifikant med t-teststørrelsen -1.23 og signifikanssandsynligheden p = 23.7%.

Det er muligt at køre samme eksempel uden omkodningen af kønsvariablen, sex, ved brug af en anden procedure, Proc GLM (Generelle Lineære Modeller), der tillader en såkaldt Class statement. Bemærk at optionen solution skal med for at få udskrevet parameterestimaterne.

**proc** **glm** data=sashelp.class plots=all;

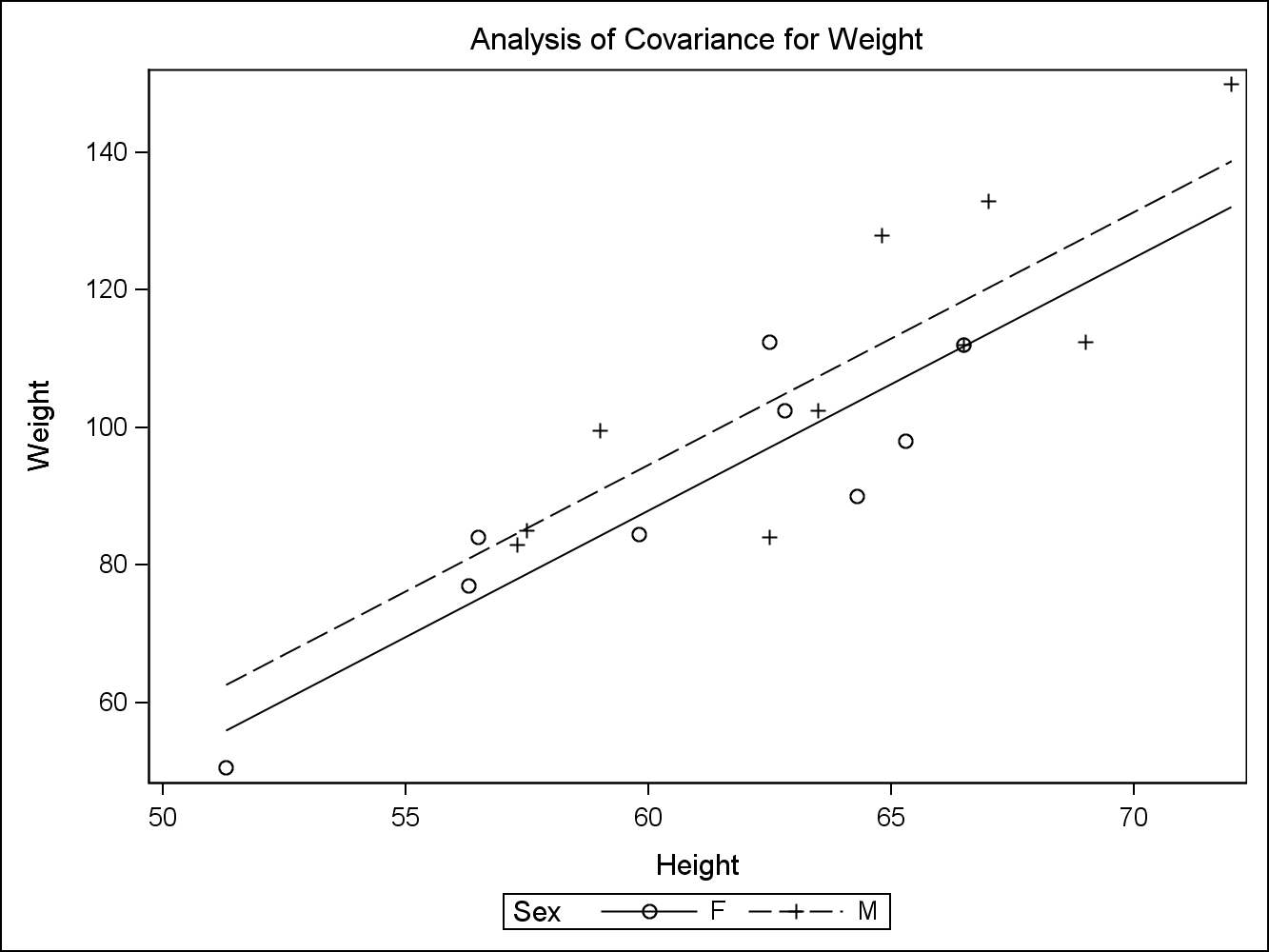
class sex;

model weight=sex height/solution;

id name;

**run**;

Derved dannes et plot, der viser de to estimerede regressionslinier for drenge og piger hver for sig.



I ovenstående model antages at højden påvirker vægten på samme måde for piger som for drenge - med andre ord de to linier på plottet er parallelle. Denne antagelse kan afprøves ved at inkludere en vekselvirkning i modellen, så den bliver

yi = α + βxi + γDi + δDixi + ei

Det ekstra led er kun forskelligt fra nul for pigerne, da Di = 0 for drengene og Di = 1 for pigerne. For drengene bliver ligningen derfor

yi = α + βxi + ei

mens den for pigerne bliver

yi = (α + γ) + (β + δ)xi + ei

I Proc GLM kodes det ved

**proc** **glm** data=sashelp.class plots=all;

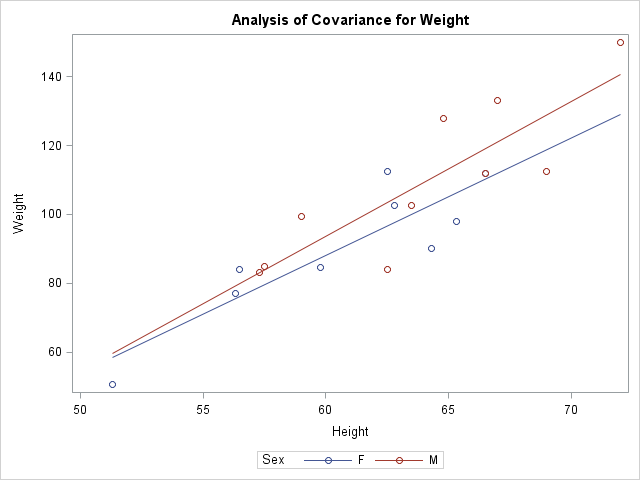
class sex;

model weight=sex height sex\*height/solution;

id name;

**run**;

Rent grafisk viser et plot i outputtet at linierne er tilnærmelsesvis parallelle, hvilket også bekræftes af de estimerede parametre, hvor koefficienten til produktleddet er insignifikant, p = 67%.



| **Parameter** | **Estimate** |  | **Standard Error** | **t Value** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | -141.1010166 | B | 49.10290115 | -2.87 | 0.0116 |
| **Sex F** | 23.7312215 | B | 69.08843486 | 0.34 | 0.7360 |
| **Sex M** | 0.0000000 | B | . | . | . |
| **Height** | 3.9125492 | B | 0.76625758 | 5.11 | 0.0001 |
| **Height\*Sex F** | -0.4881440 | B | 1.10756565 | -0.44 | 0.6657 |
| **Height\*Sex M** | 0.0000000 | B | . | . | . |

Eksemplet kan yderligere fortsættes ved at inddrage teenagernes alder

**data** a;

set sashelp.class;

if sex='M' then gender=**0**;

else gender=**1**;

run;

**proc** **reg** data=a plots=all;

model weight=gender height age;

id name;

**run**;

Det ses, at alderen ikke bliver signifikant, hvilket nok især skyldes, at alle er nogenlunde lige gamle.

| *Parameter Estimates* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *Variable* | *DF* | *Parameter Estimate* | *Standard Error* | *t Value* | *Pr > |t|* |
| *Intercept* | 1 | -116.37082 | 36.17950 | -3.22 | 0.0058 |
| *gender* | 1 | -8.74432 | 5.83497 | -1.50 | 0.1547 |
| *Height* | 1 | 2.87287 | 0.99706 | 2.88 | 0.0114 |
| *Age* | 1 | 3.11311 | 3.23618 | 0.96 | 0.3513 |

# Variabeludvælgelse

Det er muligt at fjerne insignifikante forklarende variable i en multipel regressionsmodel ved automatiske metoder. Tilsvarende er det muligt at finde signifikante forklarende variable. Det gøres ved optionen selection i model statementet. I det følgende eksempel tilføjes flere vekselvirkninger for at illustrere SAS-koden.

**data** a;

set sashelp.class;

if sex='M' then gender=**0**;

else gender=**1**;

veksel\_age=gender\*age;

veksel\_height=gender\*height;

**run**;

**proc** **reg** data=a;

model weight=gender height age veksel\_height veksel\_age/

selection=backward;

**run**;

Til sidst i udskriften vises parametrene i den endelige model

| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | -127.57294 | 31.39331 | 1756.22338 | 16.51 | 0.0010 |
| **Height** | 3.70201 | 0.49123 | 6040.07800 | 56.79 | <.0001 |
| **veksel\_height** | -1.87277 | 0.97190 | 394.87527 | 3.71 | 0.0732 |
| **veksel\_age** | 8.07523 | 4.43426 | 352.69866 | 3.32 | 0.0886 |

Der vises også en tabel over de successive tests, der viste, at de andre forklarende variable kunne fjernes. Bemærk at der anvendes et 10% signifikansniveau. Det kan være en god idé, at overveje betydningen af variable, der er signifikante på 10% niveau, men ikke på 5% niveau, i lyset af modelkontrol, indflydelsesrige observationer og robuste estimationsmetoder.

Vekselvirkningen mellem alder og køn er positiv. Det kan være interessant, at alderens tilsyneladende betyder lidt for pigers vægt, dvs for gender = 1, i dette datasæt med yngre teenagere, men ikke for drengene. Vekselvirkningen mellem højde og køn er negativ, så højden betyder altså lidt mindre for pigers vægt end for drenges vægt.

|  |
| --- |
| **All variables left in the model are significant at the 0.1000 level.** |

|  |
| --- |
|  |
|  |
|  |

| **Summary of Backward Elimination** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | Age | 4 | 0.0021 | 0.8312 | 4.1648 | 0.16 | 0.6914 |
| **2** | gender | 3 | 0.0021 | 0.8291 | 2.3266 | 0.17 | 0.6845 |

Signifikansniveauet i variabeludvælgelsen kan ændres ved optionerne slstay og slentry.

**proc** **reg** data=a;

model weight=gender height age veksel\_height veksel\_age/

selection=backward slstay=**0.05** slentry=**0.05**;

**run**;

|  |
| --- |
| **All variables left in the model are significant at the 0.0500 level.** |

|  |
| --- |
|  |
|  |
|  |

| **Summary of Backward Elimination** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | Age | 4 | 0.0021 | 0.8312 | 4.1648 | 0.16 | 0.6914 |
| **2** | gender | 3 | 0.0021 | 0.8291 | 2.3266 | 0.17 | 0.6845 |
| **3** | veksel\_age | 2 | 0.0378 | 0.7913 | 3.2730 | 3.32 | 0.0886 |
| **4** | veksel\_height | 1 | 0.0208 | 0.7705 | 2.8982 | 1.60 | 0.2243 |

I stedet for backward elimination, hvor forklarende variable fjernes successivt, kan der anvendes selection=forward, hvor forklarende variable successivt tilføjes. Det kan også udføres trinvist med selection=stepwise, der er en blanding af forward og backward udvælgelse. Optionen slentry kan anvendes i denne forbindelse, da der som default i visse tilfælde anvendes et urealistisk højt signifikansniveau på 50%. Der er ingen garanti for at de forskellige metoder ender med samme model, dvs med de samme forklarende variable, men det vil som oftest være tilfældet.