Kapitel 9

Ensidet variansanalyse

Den ensidede variansanalysemodel minder om den simple regressionsmodel vi studerede i kapitel 8. Forskellen ligger i, at den forklarende variabel, som i regressionsanalysen er *kvantitativ,* dvs. med (reelle) talværdier, i den ensidede variansanalysemodel er erstattet med en inddeling af observationerne i grupper, også kaldet en *faktor* eller en *kvalitativ* variabel.

I forlængelse af Toyota Hiace eksemplet kunne vi tænke på en situation, hvor vi kun ser på biler af en bestemt alder (f.eks. 4 år), men til gengæld på flere forskellige typer af biler. Biltypen kunne så være en faktor eller kvalitativ variabel med værdier i en endelig mængde af formen {ToyotaHiace, FordTransit, VWtransporter, ... }. Elementerne i denne mængde kaldes faktorens *niveauer.* Generelt er faktorens niveauer altså de navne (eller numre), som man bruger til at betegne grupperne.

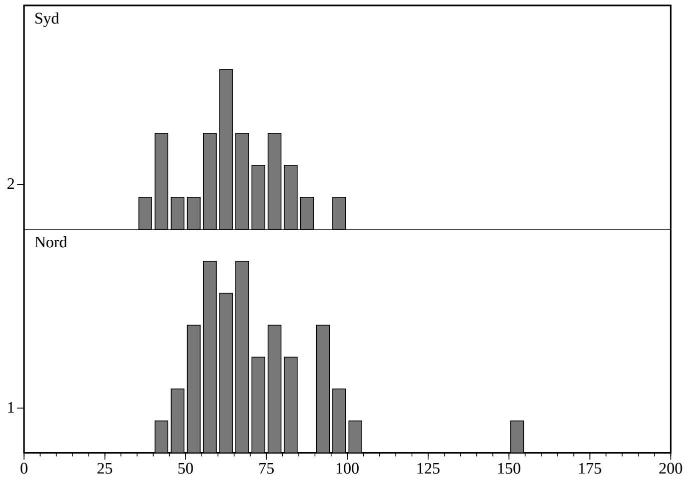
Eksempel 9.1. Det bliver for kedeligt at se på biler hele tiden, nu går vi over til at se på huse. Nedenstående tal stammer fra en hjemmeopgave i teoretisk statistik på HA fra 1984.

Omr.	kr/m^2	Omr.	kr/m^2	Omr.	kr/m^2	Omr.	kr/m^2
N	65.24	N	61.60	N	68.57	N	66.43
N	76.64	N	44.77	N	56.57	N	59.37
N	70.37	N	52.08	N	67.74	N	59.02
N	51.28	N	62.16	N	67.32	N	74.09
N	94.56	N	80.00	N	59.59	N	48.00
N	64.40	N	83.00	N	60.46	N	53.24
N	59.33	N	79.83	N	74.07	N	63.54
N	54.28	N	67.50	N	93.07	N	59.57
N	91.84	N	49.58	N	79.03	N	153.33
N	96.00	N	81.93	N	77.50	N	90.90
N	101.99	N	98.99	S	81.73	S	65.17
S	67.83	S	59.88	S	55.30	S	85.84
S	68.86	S	73.75	S	63.33	S	79.33
S	43.33	S	52.88	S	71.15	S	83.33
S	44.25	S	64.54	S	78.94	S	62.85
S	41.28	S	38.46	S	64.18	S	55.99
S	45.81	S	60.54	S	99.13	S	75.24

Tallene vedrører samtlige villaer som var annonceret til salg i Berlingske Tidende 4/9 1983, og som var beliggende enten i kystområdet mellem

Klampenborg og Helsingør (N) eller i kystområdet mellem Brøndby og Køge (S). For hver villa er angivet forholdet mellem den månedlige bruttoydelse i kr. og og boligarealet i m^2. Rækkehuse, kædehuse, gårdhavehuse o.l. er ikke medtaget. Der er 42 villaer i område N og 26 i område S, ialt 68.

Et umiddelbart indtryk af, hvordan kvadratmeterpriserne fordeler sig i de to områder, og hvordan disse to fordelinger ligger i forhold til hinanden, får man ved at tegne to histogrammer, et for hvert af de to områder, på samme skala og med samme inddeling. Resultatet for passende valg af inddeling ser sådan ud:



Fordelingen af priser pr. m^2 i de to områder

Umiddelbart ser der ikke ud til at være ret stor forskel på områderne. Der er måske en svag tendens til at priserne i det nordlige område ligger højere end i det sydlige, i overensstemmelse med de fordomme som mange (især måske nordsjællændere) har. Men det er ikke så oplagt om denne tendens er signifikant. Det er præcis det, man kan undersøge ved hjælp af en ensidet variansanalyse.

Vi opstiller følgende model: De observerede m^2–priser $y_1, \ldots, y_{68}$ antages at være observationer af uafhængige normalfordelte stokastiske variable Y_i, $i = 1, \ldots, 68$, med samme varians σ^2 og middelværdier

$$\mathrm{E}Y_i = \mu_i = \begin{cases} \mu_N & \text{for huse i område N (altså for } i = 1, \ldots, 42), \\ \mu_S & \text{for huse i område S (altså for } i = 43, \ldots, 68). \end{cases}$$

Sagt på en anden måde: Hver af de to grupper N og S antages at være beskrevet ved en model som i kapitel 7 (uafhængige identisk fordelte normale variable), men med det bånd at variansen er den samme i de to modeller.

Modellens forudsætninger ser umiddelbart ud til at være opfyldt, bortset måske fra den iøjenfaldende lille søjle langt ude til højre i det nederste histogram, som skyldes et enkelt hus i område N med en månedlig ydelse pr. m^2 på over 150 kr. Man burde muligvis overveje at udnævne denne observation til at være en "outlier", dvs. en observation der bør fjernes fra datamaterialet. Der kan være tale om en trykfejl i annoncen, eller om et hus der falder helt uden for det normale (lille stråtækt nyrenoveret bindingsværkshus på 12000 m^2 naturgrund med egen havn, velholdt tennisbane, overdækket marmorsvømmepøl osv.). Vi vender tilbage til dette problem, men vælger indtil videre at ignorere det. Og bortset fra denne ene observation er der i hvert fald ikke noget at udsætte på de to histogrammer, som ser pænt normale ud (jvf. histogrammerne på side 70), og også har nogenlunde samme bredde.

9.1. Den ensidede variansanalysemodel.

Ovennævnte model er et eksempel på en *ensidet variansanalysemodel.* Vi kalder den *ensidet* fordi der kun optræder én faktor i modellen, nemlig grupperingen i områderne N og S. At det kaldes *variansanalyse,* det man foretager sig når man analyserer et datamateriale ved hjælp af sådan en model, har noget at gøre med at de centrale formler for opspaltning af den totale kvadratsum SSD_y, som man bruger i analysen, er i familie med formler der fortæller hvordan en empirisk varians i et opdelt talmateriale relaterer sig til varianser og gennemsnit indenfor de grupper man har opdelt i.

Generelt ser den ensidede variansanalysemodel sådan ud: Observationerne består af n samhørende værdier af faktorniveauer g_i og responser y_i, altså (g_i, y_i), $i = 1, \ldots, n$. g'erne skal være elementer i en endelig mængde, som vi for nemheds skyld vælger at betegne $\{1, \ldots, G\}$, hvor G altså er antallet af grupper i den inddeling vi har af datamaterialet, eller antallet af niveauer for den faktor vi betragter. I eksemplet svarer dette blot til at vi vælger at tale om "niveau 1" og "niveau 2" i stedet for "område N" og "område S". I konkrete tilfælde er det helt afgjort en dårlig idé at bruge numre i stedet for informative forkortelser, men i forbindelse med de rent matematiske udledninger giver det en lidt mere gennemskuelig notation.

Vi vil foretage yderligere en forenkling, som også kun er til brug i forbindelse med de matematiske udledninger: I stedet for at nummerere observationerne fortløbende indicerer vi dem med et gruppenummer og et løbende indeks indenfor gruppen; dvs. vi betegner i det følgende observationerne

$$y_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, n_g,$$

hvor n_g altså er størrelsen af gruppe g, og

$$n = n_1 + \cdots + n_G \quad .$$

I denne notation kan modellen formuleres således: Responserne y_{gi} antages at være observationer af uafhængige normalfordelte stokastiske variable Y_{gi} med samme varians σ^2 og middelværdier

$$\mathrm{E}Y_{gi} = \mu_g \quad .$$

Likelihoodfunktionen bliver så

$$\begin{aligned} L(\mu_1, \ldots, \mu_G, \sigma^2) &= \prod_{g=1}^{G} \prod_{i=1}^{n_g} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_{gi} - \mu_g)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \mu_g)^2 \right) , \end{aligned}$$

og log–likelihooden (på nær additiv konstant)

$$l(\mu_1, \ldots, \mu_G, \sigma^2) = -\frac{1}{2} \left(n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \mu_g)^2 \right) .$$

Som sædvanlig ser vi, at maksimering af log–likelihooden for fast σ^2 kan foretages ved minimering af en kvadratsum, i dette tilfælde

$$\sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} (y_{gi} - \mu_g)^2 \right) .$$

Løsningen til dette minimeringsproblem kan vi umiddelbart skrive op. Kvadratsummen er jo en sum af G kvadratsummer, der hver for sig er af præcis samme form som den vi minimerede i kapitel 7 (side 60) i forbindelse med estimation i modellen for uafhængige identisk fordelte normale observationer. Blot indgår kun observationerne fra den pågældende gruppe, og parameteren hedder μ_g i stedet for μ. Men da μ_g kun indgår i den g'te kvadratsum kan vi minimere disse G summer hver for sig, og får så umiddelbart, at maksimaliseringsestimatoren for μ_g er gennemsnittet af observationerne i gruppe g,

$$\hat{\mu}_g = \bar{y}_{g\cdot} \quad .$$

Den delvis maksimerede log–likelihood får herefter følgende udseende:

$$l(\hat{\mu}_1, \ldots, \hat{\mu}_G, \sigma^2) = -\frac{1}{2} \left(n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{g\cdot})^2 \right)$$

eller, med den sædvanlige betegnelse $\mathrm{SSD}_{\mathrm{res}} = \sum_g \sum_i (y_{gi} - \bar{y}_{g\cdot})^2$ for residualkvadratsummen,

$$l(\hat{\mu}_1, \ldots, \hat{\mu}_G, \sigma^2) = -\frac{1}{2} \left(n \log \sigma^2 + \frac{1}{\sigma^2} \mathrm{SSD}_{\mathrm{res}} \right) .$$

Fra tidligere (side 62) ved vi, at maksimum for denne funktion af σ^2 antages i punktet

$$\sigma^2_{ML} = \frac{\mathrm{SSD}_{\mathrm{res}}}{n}$$

hvor log–likelihooden antager den maksimale værdi

$$l(\hat{\mu}_1, \ldots, \hat{\mu}_G, \sigma^2_{ML}) = -\frac{1}{2}\left(n \log \frac{\mathrm{SSD}_{\mathrm{res}}}{n} + n\right).$$

Og ganske som i de eksempler vi har set før bruger man i praksis ikke dette estimat for variansen, men et korrigeret estimat, nemlig

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_{\mathrm{res}}}{n - G}.$$

I dette tilfælde kan vi imidlertid — uden henvisning til den generelle teori for lineære normale modeller — direkte indse at $\mathrm{SSD}_{\mathrm{res}}$ er χ^2–fordelt med $n - G$ frihedsgrader og skalaparameter σ^2, og dermed at $\hat{\sigma}^2$ er en central estimator. Den g'te af de G kvadratsummer $\sum_{i=1}^{n_g} \left(Y_{gi} - \bar{Y}_{g\cdot}\right)^2$ er jo, ifølge Ssr. sætning 7.2.2, χ^2–fordelt med $n_g - 1$ frihedsgrader og skalaparameter σ^2, og da disse summer (som funktioner af hvert sit sæt af de oprindelige variable Y_i) er stokastisk uafhængige, følger det let af Γ–fordelingens foldningsegenskab at $\mathrm{SSD}_{\mathrm{res}}$ er χ^2–fordelt med $(n_1 - 1) + \cdots + (n_G - 1) = n - G$ frihedsgrader og skalaparameter σ^2.

9.2. Konfidensgrænser for middelværdiparametre.

Konfidensgrænser for en af middelværdiparametrene μ_g kan udledes præcis som i kapitel 7 (side 64), blot med den modifikation at variansestimatet $\hat{\sigma}^2$ også er baseret på variationen indenfor de andre grupper. Uden at gå i detaljer med udledningen refererer vi resultatet: Et eksakt 95%–konfidensinterval for parameteren μ_g er givet ved

$$\mu_g = \bar{y}_{g\cdot} \pm t_{n-G}(97.5)\sqrt{\frac{\hat{\sigma}^2}{n_g}}.$$

I mange tilfælde er selve parametrene μ_g af mindre interesse, medens man er mere interesseret i differenserne imellem dem, i særdeleshed disses fortegn. Når man f.eks. udfører landbrugsforsøg eller industrielle forsøg til sammenligning af forskellige produktionsmetoder, hvor responsen er et eller andet mål for metodens effektivitet, er det næsten altid for at identificere og verificere konklusioner af typen "metode g' er (så og så meget) bedre end metode g''", og udsagn af den slags kan bedst formuleres ved hjælp af estimater og konfidensintervaller for parametre af formen $\mu_{g'} - \mu_{g''}$, såkaldte *kontraster.* Et konfidensinterval for en sådan parameter kan, ifølge tidligere bemærkninger, konstrueres ved følgende argument: ML–estimatet for $\mu_{g'} - \mu_{g''}$ er $\bar{y}_{g'}. - \bar{y}_{g''}.$. Den tilsvarende stokastiske variable $\bar{Y}_{g'}. - \bar{Y}_{g''}.$ er normalfordelt med middelværdi $\mu_{g'} - \mu_{g''}$ (dvs. den er central) og varians

$$\left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)\sigma^2.$$

I en model med kendt varians ville vi således have eksakte 95% konfidensgrænser

$$\mu_{g'} - \mu_{g''} = \bar{y}_{g'}. - \bar{y}_{g''}. \pm 1.96\sqrt{\sigma^2\left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)}.$$

Ved i dette udtryk at erstatte den kendte varians med sit estimat og normalfordelingens 97.5% fraktil 1.96 med den relevante T–fordelings 97.5% fraktil, får vi konfidensgrænserne

$$\mu_{g'} - \mu_{g''} = \bar{y}_{g'}. - \bar{y}_{g''}. \pm t_{n-G}(97.5)\sqrt{\hat{\sigma}^2\left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)}.$$

9.3. Test for homogenitet.

I mange tilfælde er det første man bør gøre i forbindelse med en ensidet variansanalyse — efter at have foretaget passende modelkontrol, herunder Bartlett's test, se senere — at teste hypotesen

$$\mu_1 = \cdots = \mu_G = \mu$$

om *fuldstændig homogenitet.* Under denne hypotese er vi tilbage i modellen fra kapitel 7, hvorfra vi kender log–likelihoodfunktionens maksimale værdi under hypotesen:

$$l(\hat{\mu}, \ldots, \hat{\mu}, \frac{\mathrm{SSD}_y}{n}) = -\frac{1}{2}\left(n \log \frac{\mathrm{SSD}_y}{n} + n\right).$$

Kvotientteststørrelsen bliver derfor

$$\begin{aligned}
-2\log q &= 2\left(l(\hat{\mu}_1, \ldots, \hat{\mu}_G, \frac{\mathrm{SSD}_{\mathrm{res}}}{n}) - l(\hat{\mu}, \ldots, \hat{\mu}, \frac{\mathrm{SSD}_y}{n})\right) \\
&= \left(n \log \frac{\mathrm{SSD}_y}{n} + n\right) - \left(n \log \frac{\mathrm{SSD}_{\mathrm{res}}}{n} + n\right) \\
&= n \log \frac{\mathrm{SSD}_y}{\mathrm{SSD}_{\mathrm{res}}} = n \log\left(1 + \frac{\mathrm{SSD}_y - \mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_{\mathrm{res}}}\right).
\end{aligned}$$

Kvotienttestørrelsen afhænger således monotont voksende af størrelsen

$$\frac{\mathrm{SSD}_y - \mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_{\mathrm{res}}}$$

eller, ækvivalent hermed, af størrelsen

$$f = \frac{(\mathrm{SSD}_y - \mathrm{SSD}_{\mathrm{res}})/(G-1)}{\mathrm{SSD}_{\mathrm{res}}/(n-G)}.$$

Af generelle resultater som vi skal bevise senere (sætning 10.3) følger, at denne størrelse under hypotesen om homogenitet følger en F–fordeling med $(G-1, n-G)$ frihedsgrader. Kvotienttestet for homogenitet kan altså udføres ved vurdering af, om ovenstående størrelse f er ekstremt stor i denne fordeling.

En intuitiv fortolkning af dette resultat får vi ved hjælp af følgende omskrivning:

$$\begin{aligned}\mathrm{SSD}_y = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi}-\bar{y})^2 &= \sum_{g=1}^{G}\left(\sum_{i=1}^{n_g}(y_{gi}-\bar{y}_{g\cdot})^2 + n_g(\bar{y}_{g\cdot}-\bar{y})^2\right)\\ &= \mathrm{SSD}_{\mathrm{res}} + \sum_{g=1}^{G} n_g(\bar{y}_{g\cdot}-\bar{y})^2.\end{aligned}$$

Bemærk at vi bruger betegnelsen $\bar{y}$ for gennemsnittet af alle observationerne. Vi kan *ikke* bruge betegnelsen $\bar{y}_{\cdot\cdot}$, fordi det ville antyde at dette gennemsnit kan udregnes som gennemsnit af de G gruppegennemsnit, og det er jo kun rigtigt hvis grupperne er lige store. Der gælder en mere kompliceret formel, som udtrykker $\bar{y}$ som et *vægtet gennemsnit* af gruppernes gennemsnit, nemlig

$$\bar{y} = \frac{n_1\bar{y}_{1\cdot} + \cdots + n_G\bar{y}_{G\cdot}}{n}.$$

Sidste led på højre side af ovenstående udtryk for SSD_y kan fortolkes som det bidrag til den totale kvadratafvigelsessum, der skyldes variationen *mellem grupper*, medens det første led $\mathrm{SSD}_{\mathrm{res}}$ jo kan fortolkes som bidraget fra variationen *indenfor grupper.* Bemærk at hvis grupperne er lige store kan sidste led fortolkes som kvadratafvigelsessummen af gruppegennemsnittenes afvigelser fra totalgennemsnittet, ganget med den fælles gruppestørrelse (idet n_g kan sættes udenfor som en konstant). En lignende fortolkning har vi i det generelle tilfælde, blot med den modifikation at det enkelte led er vægtet med gruppestørrelsen.

Af omskrivningen følger at F–størrelsen i testet for homogenitet kan skrives

$$f = \frac{\left(\sum_{g=1}^{G} n_g(\bar{y}_{g\cdot}-\bar{y})^2\right)/(G-1)}{\mathrm{SSD}_{\mathrm{res}}/(n-G)}.$$

Det fremgår heraf, at teststørrelsen udtrykker forholdet mellem variationen mellem grupper og variationen indenfor grupper, idet begge disse kvadratsummer i passende forstand er normerede ved division med deres frihedsgradsantal.

9.4. Parvise sammenligninger af gruppemiddelværdier.

Undertiden er man interesseret i at foretage parvise sammenligninger af gruppernes middelværdier, dvs. for givne faktorniveauer g' og g'' at teste hypotesen $\mu_{g'} = \mu_{g''}$. Det kan man gøre ved kvotienttestning. Men som så ofte før viser det sig at kvotienttestet er ækvivalent med et tosidet T–test, hvor T–størrelsen udregnes som estimatet for den parameter der ifølge hypotesen er lig med 0, divideret med dens estimerede standardafvigelse. I dette tilfælde er hypotesen

$$\mu_{g'} - \mu_{g''} = 0.$$

Da ML–estimatet for $\mu_{g'} - \mu_{g''}$ er $\bar{y}_{g'}. - \bar{y}_{g''}.$ og

$$\operatorname{var}\left(\bar{Y}_{g'}. - \bar{Y}_{g''}.\right) = \left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)\sigma^2,$$

får vi teststørrelsen

$$T = \frac{\bar{y}_{g'}. - \bar{y}_{g''}.}{\sqrt{\left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)\hat{\sigma}^2}},$$

som skal vurderes (tosidet) i en T–fordeling med $n - G$ frihedsgrader.

I en ensidet variansanalyse med mange grupper skal man her være opmærksom på det fænomen der kaldes *massesignifikans.* Hvis man foretager tests svarende til samtlige parvise sammenligninger mellem gruppemiddelværdier vil man med stor sandsynlighed finde nogle signifikante forskelle, også selvom middelværdierne i virkeligheden alle sammen er ens. Forkastelse på 5%–niveauet får man jo ca. hver tyvende gang, så hvis man foretager væsentlig mere end 20 tests skal man bestemt ikke forkaste på dette niveau.

Omvendt kan man sige, at selv om middelværdierne alle sammen er forskellige, vil der nok være nogle af dem der ligger så tæt ved hinanden at testet alligevel giver godkendelse. Herved kan man få slået grupper sammen på en mere eller mindre arbitrær måde, som ikke nødvendigvis gør én klogere på den underliggende faglige problemstilling. Parvise sammenligninger er ikke noget man bør sætte sin computer til at generere pr. automatik. Det er noget man kan foretage for udvalgte par af faktorniveauer, hvis der i selve problemstillingen er en rimelig begrundelse for det. Hvis man får godkendt en hypotese af formen $\mu_{g'} = \mu_{g''}$ slår man de to grupper sammen til én, og går videre med en ensidet variansanalysemodel, der har én gruppe mindre.

I tilfældet $G = 2$ er hypotesen $\mu_1 = \mu_2$ ækvivalent med homogenitetshypotesen. Derfor er testet for $\mu_1 = \mu_2$ ækvivalent med testet for homogenitet, i den forstand at kvadratet på T–størrelsen ved test for $\mu_1 = \mu_2$ netop er den F–størrelse der benyttes ved testet for homogenitet.

Eksempel 9.1 fortsat. Vi gennemgår udregningerne i forbindelse med eksemplet vedrørende villaerne i kystområderne N og S. Vi begynder med inden for hver af de to grupper at udregne hjælpestørrelser ganske som i kapitel 7. Med en notation, som gerne skulle give sig selv, får vi

$$\begin{aligned}
\mathrm{S}_y^{\mathrm{N}} &= 65.24 + \cdots + 98.99 = 3018.78 \\
\mathrm{SS}_y^{\mathrm{N}} &= 65.24^2 + \cdots + 98.99^2 = 232919.50 \\
\mathrm{SSD}_y^{\mathrm{N}} &= 232919.50 - \frac{1}{42} 3018.78^2 = 15942.53 \\
\mathrm{S}_y^{\mathrm{S}} &= 81.73 + \cdots + 75.24 = 1682.92 \\
\mathrm{SS}_y^{\mathrm{S}} &= 81.73^2 + \cdots + 75.24^2 = 114698.48 \\
\mathrm{SSD}_y^{\mathrm{S}} &= 114698.48 - \frac{1}{26} 1682.92^2 = 5766.96
\end{aligned}$$

Vi får også brug for de tilsvarende størrelser vedrørende hele datasættet:

$$\begin{aligned}
\mathrm{S}_y &= 3018.78 + 1682.92 = 4701.70 \\
\mathrm{SS}_y &= 232919.50 + 114698.48 = 347617.98 \\
\mathrm{SSD}_y &= 347617.98 - \frac{1}{68} 4701.70^2 = 22529.99
\end{aligned}$$

og endelig skal vi bruge

$$\mathrm{SSD}_{\mathrm{res}} = 15942.53 + 5766.96 = 21709.49.$$

F–størrelsen i testet for homogenitet bliver så

$$f = \frac{(22529.99 - 21709.49)/(2-1)}{21709.49/(68-2)} = 2.494$$

som skal vurderes i en F–fordeling med (1,66) frihedsgrader. Den er helt klart insignifikant, så vores snobbede fordomme falder brat til jorden. Den relevante model er herefter (hvis man stadig vil interessere sig for disse tal) modellen for fuldstændig homogenitet, som i kapitel 7.

For kontrollens skyld kan vi her også forsøge at foretage testet for homogenitet som et T–test for hypotesen $\mu_{\mathrm{N}} = \mu_{\mathrm{S}}$. Vi får

$$T = \frac{\frac{3018.78}{42} - \frac{1682.92}{26}}{\sqrt{\frac{21709.49}{68-2}\left(\frac{1}{42} + \frac{1}{26}\right)}} = 1.579.$$

Vi kan så glæde os over at der indenfor sædvanlig afrundingsnøjagtighed gælder $1.579^2 = 2.494$.

I øvrigt kan vi nu også konstatere, at det ikke gjorde så meget at vi lod den afvigende observation fra område N blive i datamaterialet. Tendensen (som altså ikke var signifikant) gik jo — som det fremgår af T–størrelsens fortegn — i retning af at husene i område N var dyrere end husene i område S. Hvis vi havde fjernet denne observation ville vi have gjort denne tendens svagere. Faktisk falder forskellen i månedlig kvadratmeterydelse mellem område N og S herved fra 7 til 5 kr., og selvom estimatet for variansen også falder lidt, bliver testet for homogenitet herved endnu mindre signifikant.

9.5. Bartlett's test.

Før man går i gang med selve analysen af modellen, herunder testet for homogenitet, bør man, så godt det nu lader sig gøre, kontrollere at modellen holder. I indledningen af dette kapitel har vi antydet hvordan det kan gøres grafisk ved hjælp af "parallelle histogrammer", ét for hver gruppe, hvorved man også får et første indtryk af hvordan gruppernes middelværdier ligger i forhold til hinanden. Modellens antagelse om, at varianserne er ens i de G grupper, kan yderligere undersøges ved et simpelt test, kaldet *Bartlett's test.* Opstil, som alternativ, en model hvor grupperne har hver sin varians, altså en model hvor Y_{gi} er normalfordelt med middelværdi μ_g og varians σ_g^2. Estimation i denne model er enkel, fordi grupperne er beskrevet ved modeller som i kapitel 7 med hvert sit sæt af parametre. Likelihoodfunktionen bliver et produkt af "kapitel 7"–likelihooder, log–likelihoodfunktionen tilsvarende en sum, så vi får åbenbart som maksimum for log–likelihooden

$$-\frac{1}{2}\sum_g \left(n_g \log \frac{\mathrm{SSD}_{\mathrm{res}}^g}{n_g} + n_g \right),$$

hvor $\mathrm{SSD}_{\mathrm{res}}^g = \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{g\cdot})^2$ betegner bidraget til residualkvadratsummen fra gruppe g. Kvotientteststørrelsen ved test for hypotesen $\sigma_1^2 = \cdots = \sigma_G^2$ i denne model bliver således

$$-2\log q = 2\left(\frac{1}{2}\left(n \log \frac{\mathrm{SSD}_{\mathrm{res}}}{n} + n \right) - \frac{1}{2}\sum_g \left(n_g \log \frac{\mathrm{SSD}_{\mathrm{res}}^g}{n_g} + n_g \right)\right)$$

$$= n \log \frac{\mathrm{SSD}_{\mathrm{res}}}{n} - \sum_g n_g \log \frac{\mathrm{SSD}_{\mathrm{res}}^g}{n_g}.$$

Den eksakte fordeling af denne størrelse (eller en passende monoton funktion af den) er ikke umiddelbart tilgængelig. Men det følger af generel asymptotisk teori, at den er approksimativt χ^2–fordelt med $G-1$ frihedsgrader.

Det er nu ikke helt denne teststørrelse man plejer at bruge. Det kan vises, at følgende størrelse (som for store gruppestørrelser n_g er approksimativt lig med $-2\log q$ ovenfor), kaldet *Bartlett's korrigerede teststørrelse*, har en fordeling der er bedre approksimeret ved en χ^2–fordeling med $G-1$ frihedsgrader:

$$b = \frac{(n-G)\log\frac{\mathrm{SSD_{res}}}{n-G} - \sum_{g=1}^{G}(n_g-1)\log\frac{\mathrm{SSD}^g_{\mathrm{res}}}{n_g-1}}{1+\frac{1}{3(G-1)}\left(\left(\sum_{g=1}^{G}\frac{1}{n_g-1}\right)-\frac{1}{n-G}\right)}.$$

Testet for varianshomogenitet foretages altså ved vurdering af, om denne størrelse er ekstremt stor i en χ^2–fordeling med $G-1$ frihedsgrader.

Konsekvenserne af en forkastelse af hypotesen om varianshomogenitet er ubehagelige. I princippet er man tvunget til at arbejde videre med en model, hvor grupperne har hver sin varians, og den viser sig at være ret besværlig at have med at gøre. I praksis anlægger man derfor en ret liberal holdning, som for eksempel indebærer, at hvis man i et stort datamateriale får forkastelse af hypotesen om varianshomogenitet, uden at dette afspejler væsentlige relative forskelle mellem gruppernes standardafvigelser, så tillader man sig alligevel at gå videre med den simple ensidede variansanalysemodel. Det betyder så, at man må tage et vist forbehold vedrørende nøjagtigheden af de tests (parvise sammenligninger, test for homogenitet) som senere foretages.

I øvrigt skal man være opmærksom på, at *heteroskedasticitet*, som (tro det eller ej) er betegnelsen for det modsatte af varianshomogenitet (eller *homoskedasticitet*) undertiden kan fjernes ved transformation. Hvis f.eks. alle tallene er positive, og det netop er grupper med stor middelværdi der også har stor spredning, så kan man i heldige tilfælde opnå varianshomogenitet (og endda måske få fordelingerne inden for grupperne til at se mere symmetriske ud) ved at transformere alle tallene med logaritmefunktionen eller en anden voksende konkav funktion.

For $G=2$ plejer man at erstatte Bartlett's test med et tosidet F–test, der kan forklares som følger. Under hypotesen $\sigma_1^2=\sigma_2^2$ er størrelsen

$$f = \frac{\hat\sigma_1^2}{\hat\sigma_2^2} = \frac{\mathrm{SSD}_y^1/(n_1-1)}{\mathrm{SSD}_y^2/(n_2-1)}$$

åbenbart F–fordelt med (n_1-1, n_2-1) frihedsgrader. Hvis derimod de to varianser er meget forskellige må man forvente, at denne størrelse bliver enten meget lille eller meget stor. Man vælger derfor at foretage testet ved tosidet vurdering af denne størrelse i sin F–fordeling under hypotesen, idet man som P–værdi angiver den dobbelte halesandsynlighed, enten til venstre (for $f<1$) eller højre (for $f>1$). I praksis vil man, da det typisk er den højre hale man finder tabelleret, udføre dette ved at udregne forholdet mellem den største og den mindste af de to varianser (altså enten f eller $1/f$), og ved opslag i den relevante tabel (altså en F–fordelingstabel med enten (n_1-1, n_2-1) eller (n_2-1, n_1-1) frihedsgrader) vurdere, om denne størrelse er for stor.

EKSEMPEL 9.1 fortsat. I villa–eksemplet får vi

$$b = \frac{(68-2)\log\frac{21709.49}{68-2} - (42-1)\log\frac{15942.53}{42-1} - (26-1)\log\frac{5766.96}{26-1}}{1 + \frac{1}{3(2-1)}\left(\frac{1}{42-1} + \frac{1}{26-1} - \frac{1}{68-2}\right)}$$

$= 1.978$. Denne størrelse er væsentligt mindre end 95%–fraktilen 3.841 i χ^2–fordelingen med 1 frihedsgrad, så der er ikke noget problem med varianshomogeniteten i dette eksempel.

Her kunne vi også have benyttet et tosidet F–test:

$$f = \frac{15942.53}{42-1} \Bigg/ \frac{5766.96}{26-1} = 1.686,$$

hvilket fører til en P–værdi på $2 \times 0.084 \approx 0.17$; altså igen godkendelse af hypotesen om varianshomogenitet.

OPGAVE 9.5.1. Følgende datasæt findes i Neter, Wasserman and Whitmore: Applied Statistics, Allyn and Bacon inc., 1988. Vi citerer fra beskrivelsen:

Soft Drink. A firm developing a new citrus–flavored soft drink conducted an experiment to study consumer preferences for the color of the drink. Four colors were under consideration: colorless, pink, orange and lime green. Twenty test localitites were selected that were similar in sales potential and representative of the target market for this product. Each color was then randomly assigned to five of these localitites for test marketing. The dependent variable was number of cases sold during the test period per 1000 population, and the independent variable was color. Other factors, such as price, flavor, degree of carbonation, sweetness, and calorie content, were held fixed in all the localities.

Colorless	Pink	Orange	Green
26.5	31.2	27.9	30.8
28.7	28.3	25.1	29.6
25.1	30.8	28.5	32.4
29.1	27.9	24.2	31.7
27.2	29.6	26.5	32.8

Hvad viser dette forsøg?