14. JÄMFÖRELSEMÅTT

14.1 Inledning

I många sammanhang har man ett behov av att jämföra modeller med ett datamaterial och att kunna kvantifiera skillnaden mellan en modell och ett datamaterial. I prövningsproblemet, t ex, måste vi förkasta en hypotes om den modell som specificeras av hypotesen ger en dålig beskrivning av verkligheten, dvs om skillnaden mellan modellen och datat är stor. I uppskattningsproblemet kan vi tänka oss att varje värde på modellens parametrar fulständigt specificerar en modell och att det finns en viss skillnad mellan en sådan fullständigt specificerad modell och ett datamaterial. Vi kan då formulera uppskattningsproblemet som problemet att söka det värde på parametrarna som ger den minsta skillnaden mellan modellen och datamaterialet.

I det här kapitlet introducerar vi tre olika sätt att kvantifiera en jämförelse mellan modell och data, dvs vi definierar tre olika mått på skillnaden mellan modell och data. Det är viktigt att påpeka att det finns egentligen hur många sätt som helst att mäta en sådan skillnad. De tre mått vi diskuterar här tillhör de vanligaste måten, men det existerar andra mått som också förekommer i diverse tillämpningar. Det är också viktigt att påpeka att valet av jämförelsemått måste ske med stor omsorg och eftertanke, speciellt som valet kan påverka resultaten av både uppskattningsproblem och prövningsproblem.

14.2 Kvadratisk avvikelse

Antag att vi har n stycken observationer y_1, y_2, ..., y_n och att vi har en modell som predikterar observationerna med $\hat{y}_1$, $\hat{y}_2$, ..., $\hat{y}_n$. Dvs, för den i-te observationen föreslår modellen att vi bör observera värdet $\hat{y}_i$ medan vi faktiskt observerar värdet y_i. Ett sätt att jämföra modell och verklighet är då att bilda prediktionsfelet $y_i - \hat{y}_i$ för varje observation och sedan summera prediktionsfelen vi får för varje observation.

Ett problem med en sådan jämförelse är att ett stort positivt prediktionsfel då helt kompenseras av ett lika stort negativt prediktionsfel, dvs jämförelsen kan indikera att modell och data är nära varandra medan modellen i själva verket inte lyckas prediktera någon observation bra.

För att lösa detta problem måste vi låta jämförelsemåttet få positiva bidrag från alla prediktionsfel, oavsett om felen är positiva eller negativa. Ett sätt att åstadkomma detta är att bilda kvadratiska prediktionsfel $(y_i - \hat{y}_i)^2$ och att summera alla sådana. Detta ger oss det kvadratiska avvikelsemåttet

$$SS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Beteckningen SS står här för förkortningen av det engelska uttrycket "sum of squares".

Antag nu speciellt att vi har en lägesparametermodell, dvs vi önskar beskriva värdet på alla observationer med en (läges-) parameter, μ. Detta innebär att $\hat{y}_1 = \hat{y}_2 = ... = \hat{y}_n = \mu$ och det kvadratiska avvikelsemåttet blir

$$SS = \sum_{i=1}^{n} (y_i - \mu)^2.$$

Observera här att observationerna y_1, y_2, ..., y_n är kända tal och att värdet på SS varierar om värdet på μ varieras. Om μ inte är känd men skall uppskattas, söker vi sålunda det värde på μ som minimerar skillnaden mellan modell och data. Då vi använder det kvadratiska avvikelsemåttet skall vi alltså uppskatta lägesparametern μ med det tal som minimerar SS. Det värde på μ som minimerar SS kallas *minstakvadratskattningen* av μ. Minimeringen av SS går matematiskt till så att man deriverar uttrycket för SS med avseende på μ och sätter derivatan lika med noll. Detta ger ekvationen

$$-2 \cdot \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0,$$

där $\hat{\mu}$ är det värde på μ som minimerar SS. Lösningen till denna ekvation visar sig nu vara

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y},$$

dvs, om vi använder det kvadratiska avvikelsemåttet så fås skattningen av lägesparametern som medelvärdet av alla observationer. Vi ser också att det minsta värdet på avvikelsemåttet måste vara

$$SS_A = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)\cdot s^2,$$

där s^2 är variansen i observationsmaterialet

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Om vi har hypotesen

$$H_0 : \mu = \mu_0$$

för något givet värde på μ_0, får vi en fullständigt specificerad modell vars kvadratiska avvikelse till observationerna är

$$SS_0 = \sum_{i=1}^{n}(y_i - \mu_0)^2.$$

I analogi med tidigare beteckningar har vi här lagt till indexet 0 för att beteckna att detta är värdet på SS under förutsättning att hypotesen H_0 är sann. Vi vet att SS_A är det minsta värdet på SS, så att SS_0 kan inte vara mindre än SS_A. Om H_0 är en rimlig hypotes bör modellen som specificeras av H_0 bara vara marginellt sämre än en modell utan att restriktionen i H_0 gäller. Uttryckt i termer av avvikelsemått innebär detta att $SS_0 - SS_A$ bör vara litet om H_0 är en rimlig hypotes. Om däremot $SS_0 - SS_A$ är stort antyder det att modellen som specificeras av H_0 är en dålig modell, vilket ger

oss anledning att förkasta H_0. Vi kan alltså använda skillnaden $SS_A - SS_0$ för att avgöra om en hypotes är rimlig eller ej. Av tekniska skäl som beskrivs i nästa kapitel, är det inte alltid möjligt att arbeta med skillnaden $SS_A - SS_0$ utan vi måste arbeta med den relativa skillnaden

$$F = \frac{SS_0 - SS_A}{SS_A/\left(n-1\right)} = \left(n-1\right)\frac{SS_0 - SS_A}{SS_A}.$$

Storheten F uttrycker alltså något som kan tolkas som $n-1$ gånger den relativa försämring i anpassning till data vi upplever då vi inför den restriktion av modellen som beskrivs av hypotesen H_0. Om vi sätter in uttrycken för SS_A och SS_0 i uttrycket för F och räknar en stund finner vi att F också kan uttryckas som

$$F = \left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}}\right)^2.$$

14.3 χ^2-måttet

Det kvadratiska avvikelsemåttet är intuitivt rimligt då alla prediktioner $\hat{y}_1$, $\hat{y}_2, ...,\ \hat{y}_n$, har samma osäkerhet. I många tillämpningar har dock prediktioner olika stor osäkerhet. Det är då rimligt att en kvadratisk avvikelse $(y_i - \hat{y}_i)^2$ får betyda mer om osäkerheten i prediktionen bedöms som liten av modellen. På samma sätt får den kvadratiska avvikelsen betyda mindre om osäkerheten i prediktionen är stor. Ett sätt att åstadkomma detta är att vikta de kvadratiska avvikelserna $(y_i - \hat{y}_i)^2$ med variansen σ_i^2 för prediktionen. Detta motiverar det så kallade χ^2-måttet

$$Q = \sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n}\frac{\left(y_i - \widehat{y}_i\right)^2}{\sigma_i^2}.$$

En av det viktigaste tillämpningarna av χ^2-måttet uppkommer i samband med analys av multinomialt fördelade stokastiska variabler. Antag först att vi har en binomialfördelad stokastisk variabel, dvs de observerade individerna tillhör en av två möjliga grupper. Om den modellerade sannolikheten att

tillhöra den ena gruppen är π_1 är det predikterade antalet individer som observeras i den gruppen lika med det förväntade antalet,

$$\hat{y}_1 = n\pi_1$$

och antalet individer i gruppen har variansen

$$\sigma_1^2 = n\pi_1(1-\pi_1),$$

där n är totala antalet observerade individer. Detta ger oss χ^2-måttet

$$Q_1 = \frac{(y_1 - n\pi_1)^2}{n\pi_1\,(1-\pi_1)},$$

med y_1 som antalet individer som faktiskt tillhör den ena gruppen. Antalet som tillhör den andra gruppen är uppenbarligen $y_2 = n - y_1$. Sannolikheten att observera en individ som tillhör den andra gruppen är $\pi_2 = 1 - \pi_1$ och det predikterade antalet individer är $\hat{y}_2 = n\pi_2 = n - n\pi_1 = n - \hat{y}_1$. Det följer därför att

$$\begin{aligned} Q_1 &= \frac{(y_1 - n\pi_1)^2}{n\pi_1\pi_2} = \frac{(y_1 - n\pi_1)^2}{n\pi_1} + \frac{(y_1 - n\pi_1)^2}{n\pi_2} \\ &= \frac{(y_1 - n\pi_1)^2}{n\pi_1} + \frac{(y_2 - n\pi_2)^2}{n\pi_2} \\ &= \frac{(y_1 - n\pi_1)^2}{\widehat{y}_1} + \frac{(y_2 - n\pi_2)^2}{\widehat{y}_2}. \end{aligned}$$

Detta resonemang kan nu generaliseras till fallet med ν stycken möjliga utfall. Låt y_i vara antalet individer i den i-te gruppen och $\hat{y}_i$ vara det predikterade antalet individer i den i-te gruppen. Då blir, på samma sätt som ovan, χ^2-måttet

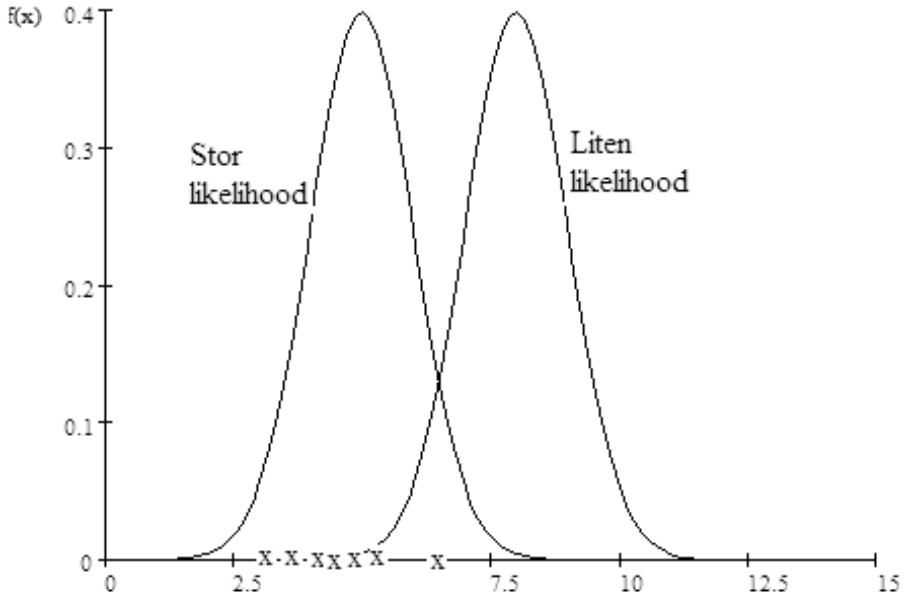
$$Q_{\nu-1} = \sum_{i=1}^{\nu} \frac{(y_i - \widehat{y}_i)^2}{\widehat{y}_i}.$$

Denna version av χ^2-måttet används mycket ofta i statistisk analys och förekommer på listor över de tio viktigaste vetenskapliga framstegen under 1900-talet. Små värden på $Q_{\nu-1}$ svarar mot en god överensstämmelse mellan modell och data medan stora värden antyder att modellen inte lyckas beskriva observationsmaterialet på ett bra sätt. Kapitel 18 ägnas speciellt åt prövningar av hypoteser med hjälp av χ^2-måttet.

14.4 Likelihood

De jämförelsemått som presenteras i de föregående avsnitten bygger på jämförelser mellan observerade och predikterade värden. Likelihoodmåttet för en modell, däremot, bygger på hur troligt det är att ett givet datamaterial har blivit genererat på det sätt som beskrivs av modellen. För att avgöra troligheten använder vi sannolikheter. Om vi t ex vet att Y är en binomialfördelad stokastisk variabel med parametrarna $n = 10$ och $\pi = 0.8$, $Y \sim bin(n = 10, \pi = 0.8)$, så inträffar utfallet $Y = 8$ med sannolikheten 0.302. Om Y, å andra sidan, är binomialfördelad med $n = 10$ och $\pi = 0.5$, $Y \sim bin(n = 10, \pi = 0.5)$, inträffar samma utfall, $Y = 8$, med sannolikheten 0.044. Om vi har observerat utfallet $y = 8$ från en binomialmodell med $n = 10$ men med okänt π, säger vi därför att det är troligare att utfallet är genererat av modellen $bin(n = 10, \pi = 0.8)$ än av modellen $bin(n = 10, \pi = 0.5)$. Vi säger att modellernas likelihood är $L(\pi = 0.8) = 0.302$ respektive $L(\pi = 0.5) = 0.044$ eller att den förra modellen har $L(\pi = 0.8)/L(\pi = 0.5) = 0.302/0.044 = 6.86$ gånger så stor likelihood som den senare modellen. Figur 14.1 ger en grafisk illustration av likelihood för två modeller av kontinuerliga stokastiska variabler.

Det faktum att vi arbetar med observerade utfall medför en komplikation i tolkningen av likelihoodmåttet. Det som har inträffat, det har inträffat och det är meningslöst att prata om sannolikheter i det fallet. Likelihood får därför inte blandas ihop med sannolikhet, även om vi använder sannolikhetsmodeller för att bestämma likelihoodvärden.



Figur 14.1 Två modeller för kontinuerliga stokastiska variabler. "x" markerar observationer.

Likelihoodfunktionen definieras som frekvensfunktionen för ett utfall av diskreta stokastiska variabler och som täthetsfunktionen för ett utfall av kontinuerliga stokastiska variabler, men betraktad som en funktion av okända parametrar och med utfallen som fixa konstanter. Likelihoodfunktionen är således en funktion med mängden av alla tänkbara parametervärden (även kallad parameterrummet) som definitionsområde och betraktar utfallen som fixa. Frekvens- och täthetsfunktionerna, å andra sidan, är funktioner med mängden av alla utfall (utfallsrummet) som definitionsområde och betraktar parametrarna som fixa. I exemplet ovan med binomialfördelningarna har vi alltså likelihoodfunktionen

$$L(\pi) = f(y = 8; \pi) = \binom{10}{8} \pi^8 (1 - \pi)^2.$$

Uppskattningsproblemet behandlas genom att det parametervärde som svarar mot den modell med det största likelihoodvärdet väljs som skattning. Den

skattningen kallas därför för maximum likelihoodskattningen. Om vi t ex har n oberoende observationer, $y_1, y_2, \ldots, y_n$, från en $N(\mu,\, \sigma^2)$ fördelning är likelihoodfunktionen lika med den simultana täthetsfunktionen

$$L(\mu, \sigma^2) = f(y_1, y_2, ..., y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(y_i-\mu)^2/(2\sigma^2)}$$

Om problemet är att uppskatta μ ser vi att $L(\mu,\, \sigma^2)$ maximeras då $SS = \sum_{i=1}^{n}(y_i - \mu)^2$ minimeras, dvs då vi gör oberoende observationer på en normalfördelning sammanfaller maximum likelihood skattningen av μ med minsta kvadrat skattningen av μ. Därav följer också att maximum likelihood skattningen av μ i denna modell är $\hat{\mu} = \bar{y}$, dvs observationernas medelvärde.

Prövningsproblemet behandlas genom att modellernas likelihoodvärden jämförs. Om vi har n oberoende observationer från en $N(\mu,\, \sigma^2)$ fördelning och vill pröva hypotesen

$$H_0 : \mu = \mu_0$$

för något specificerat värde på μ_0 jämför vi likelihoodvärdet $L(\mu_0,\, \sigma^2)$ med det största tänkbara likelihoodvärde vi kan få för det givna materialet, dvs $L(\hat{\mu},\, \sigma^2)$, där $\hat{\mu}$ är maximum likelihoodskattningen av μ. Jämförelsen görs med hjälp av kvoten

$$\lambda = \frac{L\left(\mu_0, \sigma^2\right)}{L\left(\widehat{\mu}, \sigma^2\right)} = e^{\{\sum_{i=1}^{n}(y_i-\hat{\mu})^2 - \sum_{i=1}^{n}(y_i-\mu_0)^2\}/(2\sigma^2)}.$$

Om $L(\hat{\mu},\, \sigma^2)$ är mycket större än $L(\mu_0,\, \sigma^2)$ indikerar detta att H_0 ger en dålig beskrivning av datamaterialet. Dvs, om kvoten λ är liten måste hypotesen H_0 förkastas. Om, å andra sidan, $L(\mu_0, \sigma^2)$ och $L(\hat{\mu},\, \sigma^2)$ är ungefär lika stora medför inte hypotesen H_0 någon svårare begränsning i modellens möjlighet att beskriva datamaterialet. I detta fall är kvoten λ närmare ett, vilket skall tolkas som att H_0 inte kan förkastas.

I stället för att arbeta med kvoten λ brukar man, av bland annat beräkningstekniska skäl, arbeta med

$$D = -2\ln\lambda = -2\ \ln\frac{L\left(\mu_0,\sigma^2\right)}{L\left(\widehat{\mu},\sigma^2\right)} = 2\{\ln\ L(\hat{\mu},\sigma^2) - \ln\ L(\mu_0,\sigma^2)\}.$$

I fallet med oberoende observationer från en normalfördelning är

$$\begin{aligned} D &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(y_i-\mu_0)^2 - \sum_{i=1}^{n}(y_i-\hat{\mu})^2\right\} = \\ &\quad \frac{1}{\sigma^2}\left\{SS_0 - SS_A\right\}. \end{aligned}$$

Om populationsvariansen σ^2 är okänd och ersätts med urvalsvariansen s^2, är vi tillbaka i samma problem som vi studerade i avsnitt 14.2 med, naturligtvis, samma lösning. Hypotesen H_0 förkastas alltså för stora värden på

$$F = \left(\frac{\overline{y}-\mu_0}{s/\sqrt{n}}\right)^2.$$