# Föreläsning 10: Pearsons $\chi^2$-test och linjär regression

Johan Thim (johan.thim@liu.se)

March 22, 2022

## 1 Det grundläggande $\chi^2$-testet

Antag att vi har följande situation

(i) Vi har $n$ stycken oberoende stokastiska variabler $X_j$ med samma fördelning, där $X_j$ har precis $k$ möjliga utfall.

(ii) Numrera utfallen enligt $A_1, \ldots, A_k$ och låt $p_j = P(A_j)$ vara respektive sannolikhet. Då är $p_1 + p_2 \cdots + p_k = 1$.

(iii) Låt $Y_i$, $i = 1, 2, \ldots, k$, vara antalet gånger händelsen $A_i$ inträffar.

För att konkretisera en aning, tänk att vi har $k$ stycken lådor $A_j$ vi kastar bollar i. Experimentet är uppställt så att en kastad boll alltid hamnar i en låda. Vi låter $p_j$ vara sannolikheten att en boll hamnar i låda $A_j$. Vi kastar $n$ bollar (oberoende) och räknar sedan hur många bollar $Y_j$ som det finns i varje låda. Givetvis kommer $Y_j \sim \text{Bin}(n, p_j)$, men variablerna $Y_j$ är *inte* oberoende av varandra (antalet bollar i alla lådorna summerar till $n$).

Vad vi kommer gör är att betrakta uppdelningar av denna typ och ställa upp hypotestest där vi låter nollhypotesen $H_0$ ges av

$$H_0 : P(A_1) = p_1,\ P(A_2) = p_2,\ \ldots,\ P(A_k) = p_k,$$

där $p_1, p_2, \ldots, p_k$ är sannolikheter så att $p_1 + \cdots + p_k = 1$, och testar mot hypotesen

$$H_1 : \text{det finns något } j \text{ så att } P(A_j) \neq p_j.$$

Om $H_0$ är sann, så blir de förväntade frekvenserna $E(Y_j) = n \cdot p_j$, $j = 1, 2, \ldots, k$. Låt oss definiera

$$q = \sum_{j=1}^{k} \frac{(y_j - np_j)^2}{np_j},$$

där $y_j$ är observationen av $Y_j$. Ett stort värde på $q$ borde rimligen indikera att $H_0$ inte gäller (åtminstone något $p_j$ måste skilja sig markant från det förväntade värdet $np_j$).

Storheten $q$ är en observation av den stokastiska variabeln

$$Q = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \overset{\text{appr.}}{\sim} \chi^2(k-1).$$

Att detta blir approximativt $\chi^2$-fördelat följer av följande sats.

**Sats.** Med beteckningarna ovan gäller att $\sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}$ konvergerar till $X \sim \chi^2(k-1)$ i fördelning.

**När duger approximationen?**

Föregående sats gäller alltså *asymptotiskt* (då $n \to \infty$) och säger inget direkt om vad som gäller i det enskilda fallet. En tumregel är att vi vill ha $np_j \geq 5$ för $j = 1, 2, \ldots, k$ för att vara ganska säkra på att approximationen är bra. Har vi lådor med väldigt få "bollar" i kan det hända att testet inte blir bra.

## 2 Test av given diskret fördelning

Låt $X_1, X_2, \ldots, X_n$ vara oberoende diskreta stokastiska variabler med $X_j \in A$ för någon diskret mängd $A$. Vi är intresserade av att testa om $X_j \sim F$ för någon given diskret fördelning med sannolikhetsfunktion $p(j)$, $j \in A$. Vi kommer använda nollhypotesen

$$H_0 : P(X = j) = p(j),\ j \in A,$$

och testar den med mothypotesen

$$H_1 : P(X = j) \neq p(j) \text{ för något } j \in A.$$

**Exempel**

Den stokastiska variabeln $X$ antar värden i mängden $\{0, 1, 2\}$. Vid 1250 observationer fann man att $X = 0$ 783 gånger, $X = 1$ 425 gånger samt $X = 2$ 42 gånger. Testa med signifikansnivån 1% om $X \sim \text{Bin}(2, 1/5)$.

**Lösning**. Vi låter $H_0 : X \sim \text{Bin}(2, 1/5)$. Om vi antar att $H_0$ är sann så gäller att

$$P(X = 0) = \binom{2}{0} \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^2 = \frac{16}{25},$$
$$P(X = 0) = \binom{2}{1} \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^1 = \frac{8}{25},$$
$$P(X = 0) = \binom{2}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^0 = \frac{1}{25}.$$

Kom ihåg att kontrollera att dessa summerar till 1, det är en billig kontroll på tentan. Utifrån detta kan vi beräkna de förväntade frekvenserna vid 1250 försök (om $H_0$ är sann):

$$np_j = \begin{cases} 800, & j = 0, \\ 400, & j = 1, \\ 50, & j = 2. \end{cases}$$

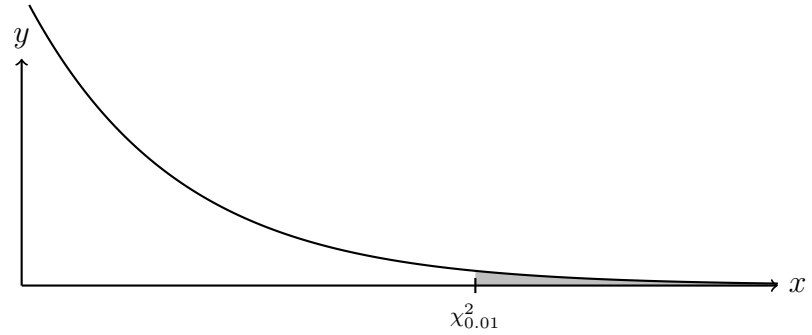
Testvariabeln $q$ ges nu av

$$q = \sum_{j=0}^{2} \frac{(x_j - np_j)^2}{np_j} = \frac{(783-800)^2}{800} + \frac{(425-400)^2}{400} + \frac{(42-50)^2}{50} \approx 3.2038.$$

Eftersom $k = 3$ är $q$ en observation av $Q \overset{\text{appr.}}{\sim} \chi^2(2)$ om $H_0$ är sann. Vi finner att

$$0.01 = P(Q > \chi^2_{0.01}(2)) \quad \Leftrightarrow \quad \chi^2_{0.01}(2) = 9.21$$

ur tabell.




Eftersom $q = 3.2038 < 9.21$ kan vi inte förkasta $H_0$. Fördelningen kan mycket riktigt vara binomialfördelning med $p = 1/5$.

# 3 Test för kontinuerlig fördelning

Om vi istället har en kontinuerlig situation där vi vill testa om mätdata följer en given fördelning $F$ måste vi agera lite annorlunda. Vi skulle önska att ställa upp

$$H_0 : X \sim F$$

mot

$$H_1 : X \text{ har ej fördelningen } F.$$

Men detta blir lite för komplicerat i det generella fallet.

Istället gör vi så att vi diskretiserar det hela på något sätt. Vi gör oftast detta genom att skapa lådor i form av intervall och sedan undersöka hur många observationer som hamnar i varje delintervall. Detta gör att vi inte exakt testar om nollhypotesen ovan utan vi testar en svagare nollhypotes.

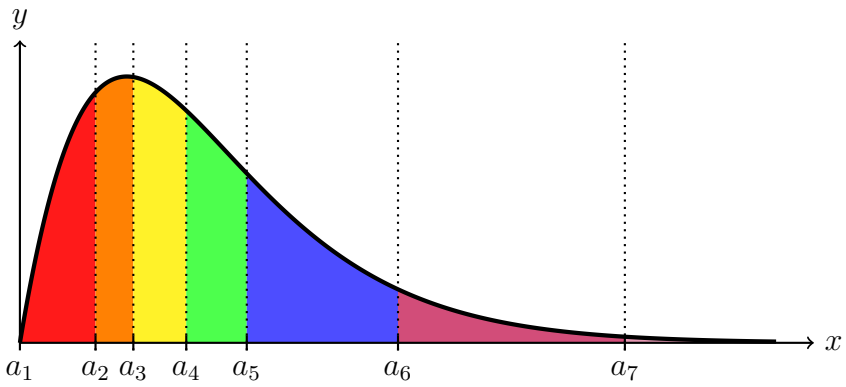
Låt $X_i$, $i = 1, 2, \ldots, n$ vara oberoende och likafördelade variabler med täthetsfunktion $f(x)$. Vi väljer $a_j$, $j = 1, 2, \ldots, k+1$, så att

$$-\infty \leq a_1 < a_2 < \cdots < a_k < a_{k+1} \leq \infty$$

och definierar $A_j = [a_j,\ a_{j+1}[$ för $j = 2, 3, \ldots, k$ och låter typiskt $A_1 = ]-\infty,\ a_2[$. Vi definierar sedan

$$p_j = P(X_i \in A_j) = \int_{a_j}^{a_{j+1}} f(x)\, dx.$$

Om $f$ är en täthetsfunktion så blir nu $p_1 + p_2 + \cdots + p_k = 1$ och vi har täckt alla möjligheter. Om stödet för $f$ inte är hela $\mathbf{R}$ modifierar vi naturligt definitionen (eller låter $f(x) = 0$ utanför sin definition). En tumregel för valet är att vi låter $k \approx n/10$. En annan tumregel är att välja intervallen så stora att alla $p_j$ är ungefär lika stora.




Hypotesen vi kommer testa är

$$H_0 : P(X \in A_j) = p_j, \quad j = 1, 2, \ldots, k,$$

mot

$$H_1 : P(X \in A_j) \neq p_j \text{ för något } j.$$

Skulle $X$ ha rätt fördelning kommer $H_0$ att vara sann med stor sannolikhet, men om vi styrker $H_0$ innebär det inte nödvändigtvis att det är just den fördelning vi utgick från när vi ställde upp $A_j$ som är den sanna (bara någon med motsvarande sannolikheter i uppdelningen). Vill man ha ett starkare resultat krävs andra metoder.

**Exempel**

Säljaren på ELFA hävdar bestämt att livslängden på en komponent är exponentialfördelad med väntevärde 2 år. Uttråkade pensionären Sture tror inte på det utan köper 50 stycken komponenter för att testa. Sture kopplar upp komponenterna och kikar till var 6:e månad för att se hur många som gått sönder.

| Tid (mån) | $< 6$ | $< 12$ | $< 18$ | $< 24$ | $< 30$ | $< 36$ | $< 42$ | $< 48$ | $< 54$ | $< 60$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Antal: | 11 | 19 | 25 | 31 | 36 | 39 | 39 | 40 | 42 | 43 |

Undersök om antagandet är rimligt på approximativt 1% nivån.

**Lösning**. Vi kan organisera om datan mer användbart enligt hur många enheter som gick sönder under en viss tidsenhet. För att få ungefär jämnstora klasser så buntar vi ihop enligt följande.

| Tid | Hur många dog |
|---|---|
| $I_1 = [0,\ 6)$ | 11 |
| $I_2 = [6,\ 12)$ | 8 |
| $I_3 = [12,\ 24)$ | 12 |
| $I_4 = [24,\ 36)$ | 8 |
| $I_5 = [36,\ \infty)$ | 11 |

Om vi antar $H_0$ så gäller att täthetsfunktionen för livslängden hos en komponent $X$ ges av $f(x) = \mu^{-1} \exp(-\mu^{-1} x)$, så

$$P(a \leq X < b) = \int_a^b \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) dx = \exp\left(-\frac{a}{\mu}\right) - \exp\left(-\frac{b}{\mu}\right).$$

Med siffrorna ovan ser vi att

$$P(X \in I_k) = \begin{cases} p_1 = 0.2212, & k = 1, \\ p_2 = 0.1723, & k = 2, \\ p_3 = 0.2387, & k = 3, \\ p_4 = 0.1447, & k = 4, \\ p_5 = 0.2231, & k = 5. \end{cases}$$

Teststorheten vi använder kommer nu ges av

$$q = \sum_{j=1}^{5} \frac{(x_j - np_j)^2}{np_j} = \frac{(11 - 50 \cdot 0.2212)^2}{50 \cdot 0.2212} + \cdots + \frac{(11 - 50 \cdot 0.2231)^2}{50 \cdot 0.2231} = 0.1276.$$

Om $H_0$ är sann så kommer $q$ vara en observation av $Q \overset{\text{appr.}}{\sim} \chi^2(5-1) = \chi^2(4)$, så med det kritiska området $C = (0,\ c)$ där $c = 13.28$, ser vi att vi inte kan förkasta $H_0$. Säljaren kan mycket väl ha rätt.

# 4 Skattade storheter

Normalt sätt kanske vi inte får exakt väntevärde (eller andra parametrar i fördelningen) utan dessa måste skattas innan vi kan utföra testet. Hur påverkar det fördelningen för teststorheten $Q$? Svaret är enkelt: för varje skattning vi gör tappar vi en frihetsgrad, under förutsättningen att skattningen är vettig (ML-skattningar brukar bete sig bra). Bevis är däremot lite bökigare (å andra sidan får vi det första $\chi^2$-testet mer eller mindre på köpet). Får jag tid över kommer jag skriva ned det och uppdatera anteckningarna. Om vi antar att sannolikheterna $p_j$ beror på okända $\boldsymbol{\theta} = (\theta_1\ \theta_2\ \cdots\ \theta_r)^T$, så gäller alltså att

$$Q = \sum_{j=1}^{k} \frac{(Y_j - n\widehat{p}_j(\boldsymbol{\theta}))^2}{n\widehat{p}_j(\boldsymbol{\theta})} \overset{\text{appr.}}{\sim} \chi^2(k - r - 1),$$

under förutsättning att skattningarna som används beter sig tillräckligt bra.

**Exempel**

Linnea gör en signalbehandlingslaboration i matlab men hennes algoritm fungerar inte som planerat. Givetvis tycker Linnea att felet måste ligga i matlabs sätt att generera normalfördelade slumptal. För att testa hypotesen att slumptalen inte är normalfördelade genererar Linnea 1000 slumptal och sorterar dessa i storleksordning följt av en klassindelning så det är precis 100 element i varje klass. Gränserna kan ses nedan.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Undre gräns | 1.57 | 12.47 | 15.00 | 17.04 | 18.80 | 20.33 | 21.76 | 23.26 | 25.00 | 27.43 |
| Övre gräns | 12.46 | 14.98 | 17.03 | 18.77 | 20.32 | 21.75 | 23.25 | 24.99 | 27.42 | 40.80 |

Det beräknade medelvärdet är $\overline{x} = 20.14$ och stickprovsvariansen är $s^2 = 35.25$. Testa på nivån 5% om värdena är normalfördelade.

**Lösning**. Låt $H_0$ : datan kommer från $N(\mu, \sigma)$ och $H_1$ : datan är inte normalfördelad. Om vi använder $\overline{x} = 20.14$ som skattning för väntevärdet och $s = \sqrt{35.25} = 5.94$ som skattning för standardavvikelsen, så kan vi (om vi antar att $H_0$ är sann) beräkna sannolikheterna för en normalfördelad variabel $Z$ att hamna i de olika klasserna enligt

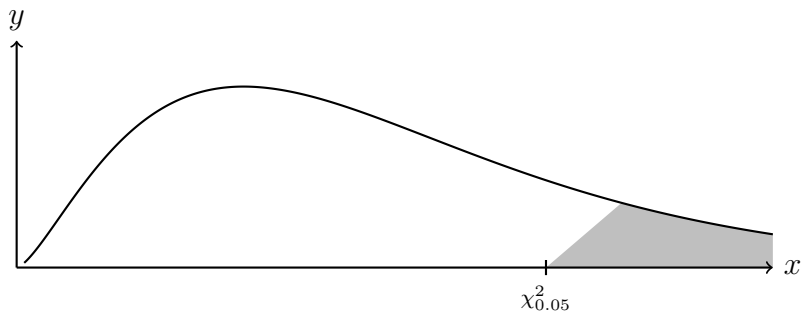
$$P(a \leq Z < b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{Z-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$
$$\approx \Phi\left(\frac{b-20.14}{5.94}\right) - \Phi\left(\frac{a-20.14}{5.94}\right).$$

Resultatet kan beskådas nedan.

| Intervall | Sannolikhet |
|---|---|
| $I_1 = (-\infty,\ 12.46)$ | 0.10 |
| $I_2 = [12.47,\ 14.99)$ | 0.09 |
| $I_3 = [15.00,\ 17.03)$ | 0.11 |
| $I_4 = [17.04,\ 18.77)$ | 0.11 |
| $I_5 = [18.78,\ 20.32)$ | 0.10 |
| $I_6 = [20.33,\ 21.75)$ | 0.09 |
| $I_7 = [21.76,\ 23.25)$ | 0.09 |
| $I_8 = [23.26,\ 24.99)$ | 0.09 |
| $I_9 = [25.00,\ 27.42)$ | 0.10 |
| $I_{10} = [27.43,\ \infty)$ | 0.11 |

Vi ser redan nu att sannolikheterna väldigt nära hamnar runt 10-delar (vilket borde ske om normalfördelning gäller med tanke på konstruktionen). Men låt oss ställa upp teststorheten och se:

$$q = \sum_{j=1}^{10} \frac{(x_j - n\widehat{p}_j)^2}{n\widehat{p}_j} = \frac{(100 - 1000 \cdot 0.10)^2}{1000 \cdot 0.10} + \cdots + \frac{(100 - 1000 \cdot 0.11)^2}{1000 \cdot 0.11} = 4.34.$$




Om $H_0$ är sann så är $q$ en observation av $\chi^2(10-2-1) = \chi^2(7)$ eftersom vi skattar två parametrar. På nivån 0.1% så gäller att $P(Q > 14.07) = 0.05$, och då $4.34 < 14.07$ så kan vi inte förkasta nollhypotesen. Linnea har antagligen implementerat sin algoritm fel.

Att testa normalfördelning på detta sätt är inte helt lämpligt. Det finns betydligt bättre metoder som till exempel Kolmogorov-Smirnovs metod som istället baserar sig på den empiriska fördelningsfunktionen. Test av denna typ ger bättre resultat i allmänhet.

Låt oss avrunda med att betrakta ett annat diskret exempel som exemplifierar hur vi hanterar potentiellt oändligt många utfall och skattade parametrar.

## Exempel

Astrid har slaktat 100 stycken brandvarnare och försöker mäta hur radioaktivt materialet är. I lämplig enhet finner hon följande frekvensdata från den 100 enheterna.

| $X = i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| antal ($y_i$) | 8 | 8 | 30 | 17 | 21 | 9 | 5 | 2 |

Testa $H_0 : X \sim \text{Po}(\mu)$ mot $H_1 : X$ är inte Poissonfördelad. på nivån 5%.

**Lösning**. Eftersom väntevärdet $\mu$ är okänt måste vi skatta detta och det gör vi enklast med

$$\widehat{\mu} = \frac{1}{100}\sum_{i=0}^{7} iy_i = \frac{0 \cdot 8 + 1 \cdot 8 + \cdots + 7 \cdot 2}{100} = 2.92.$$

Detta är medelvärdet och det är en rimlig skattning av väntevärdet (undersök om det är ML-skattningen!). Om $H_0$ är sann kan vi räkna ut sannolikheterna för de olika observationerna med hjälp av sannolikhetsfunktionen för Poissonfördelning:

$$p_i = P(X = i) = e^{-\mu}\frac{\mu^i}{i!} \approx e^{-\widehat{\mu}}\frac{\widehat{\mu}^i}{i!} = \widehat{p}_i$$

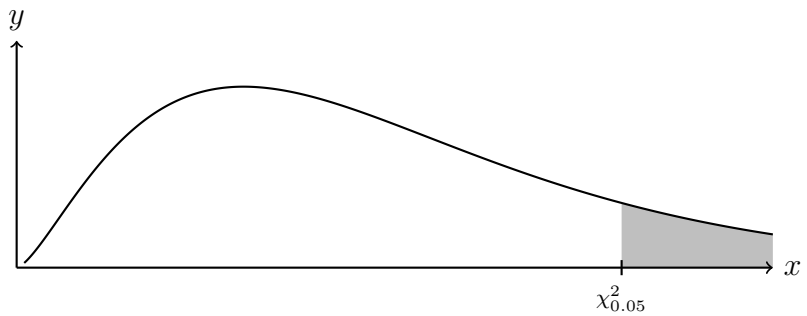
för $i = 1, 2, 3, \ldots$ (oändligt många möjligheter). Vi måste slå ihop utfallen i svansen på fördelningen då sannolikheterna där blir för små ($np_i < 5$ och dessutom har vi inget test för oändligt många möjligheter) så vi testar att låta $i \geq 6$ vara den sista klassen och gör om tabellen:

| $X = i$ | 0 | 1 | 2 | 3 | 4 | 5 | $\geq 6$ |
|---|---|---|---|---|---|---|---|
| antal ($y_i$) | 8 | 8 | 30 | 17 | 21 | 9 | 5+2=7 |
| $100\widehat{p}_i$ | 5.4 | 15.8 | 23.0 | 22.4 | 16.3 | 9.5 | 7.61 |

Vi ser här att samtliga $n\widehat{p}_i \geq 5$ (även om den första är ganska nära) så kraven är uppfyllda. Vi räknar ut teststorheten

$$q = \sum_{i=0}^{6} \frac{(y_i - 100\widehat{p}_i)^2}{100\widehat{p}_i} = \frac{(8 - 5.4)^2}{5.4} + \cdots + \frac{(7 - 7.61)^2}{7.61} = 9.91.$$

Om $H_0$ är sann så är $q$ en observation av $Q \overset{\text{appr.}}{\sim} \chi^2(7 - 1 - 1) = \chi^2(5)$ eftersom vi har 7 klasser och skattar en parameter (med en vettig skattning). Ur tabell finner vi att $P(Q > \chi^2_{0.05}(5) = 11.07) = 0.05$ och eftersom $q < 11.07$ så kan vi *inte* förkasta $H_0$. Det kan mycket väl vara så att observationerna kommer ifrån en Poissonfördelning.



# 5 Homogenitetstest

Det kan ofta vara intressant att avgöra om egenskaper skiljer sig åt mellan olika grupper. Låt oss sätta upp följande scenario. Vi har $s$ stycken grupper eller serier av försök som vi är nyfikna på om de uppvisar samma sorts fördelning med avseende på en mängd egenskaper $A_1, A_2, \ldots, A_r$.

Vi kan då ställa upp datan enligt följande där siffrorna är absoluta frekvenser.

| | Egenskap 1 | Egenskap 2 | $\cdots$ | Egenskap $r$ | Summa |
|---|---|---|---|---|---|
| Grupp 1 | $N_{11}$ | $N_{12}$ | $\cdots$ | $N_{1r}$ | $G_1$ |
| Grupp 2 | $N_{21}$ | $N_{22}$ | $\cdots$ | $N_{2r}$ | $G_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| Grupp $s$ | $N_{n1}$ | $N_{n2}$ | $\cdots$ | $N_{nr}$ | $G_s$ |
| Summa | $E_1$ | $E_2$ | $\cdots$ | $E_r$ | $N$ |

Om vi antar att grupperna är *homogena*, dvs att de uppvisar samma fördelning för egenskaperna, så är en bra skattning för sannolikheten $p_j$ att ett objekt har egenskap $j$ helt enkelt

$$\widehat{p}_j = E_j/N.$$

Vi formar samma sorts teststorhet som vi gjort innan

$$Q = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{(N_{ij} - G_i \cdot \widehat{P}_j)^2}{G_i \widehat{P}_j} \overset{\text{appr.}}{\sim} \chi^2((r-1)(s-1)).$$

Att det blir just $(r-1)(s-1)$ kommer från de linjära restriktioner som trillar ut ur tabellen ovan. Vi kan se att

$$\sum_{j=1}^{r} \frac{(N_{ij} - G_i \cdot p_j)^2}{G_i p_j} \overset{\text{appr.}}{\sim} \chi^2(r-1)$$

enligt tidigare argument, men då antar vi att $p_j$ är kända. Sedan summerar vi $s$ sådana oberoende variabler, så resultatet blir $\chi^2(s(r-1))$-fördelat. Men nu vill vi skatta $p_j$ och för varje skattning tappar vi en frihetsgrad. Märk dock att vi inte skattar alla $r$ stycken $p_j$, utan bara $r-1$ stycken då den sista ges av att summan måste bli ett. Vi tappar alltså $r-1$ frihetsgrader. Totalt sett har vi alltså $s(r-1) - (r-1) = (s-1)(r-1)$ frihetsgrader.

Så när gäller approximationen? Den gäller under förutsättning att $n_i\widehat{p}_j$ är stora. En rimlig tumregel är att $n_i\widehat{p}_j \geq 5$.

**Exempel**

Från en stor population frågar vi två grupper om de tycker det borde vara lagligt att kasta tallkottar på hundägare som inte håller sina hundar kopplade.

| Grupp | Kottkastning är OK! | Nej man får inte kasta tallkottar på folk. |
|---|---|---|
| $G_1$ | 59 | 41 |
| $G_2$ | 145 | 55 |

Testa på signifikansnivån 1% (approximativt) om det finns någon skillnad mellan vad grupperna tycker.

**Lösning**. Vi utför ett homogenitetstest.

| Grupp | Kottkastning är OK! | Nej man får inte kasta tallkottar på folk. | Summa |
|---|---|---|---|
| $G_1$ | 59 | 41 | 100 |
| $G_2$ | 145 | 55 | 200 |
| $G_1 + G_2$ | 204 | 96 | 300 |
| $\widehat{p_j}$ | 0.68 | 0.32 | 1.0 |

Låt $H_0$ : Grupperna tycker likadant mot $H_1$ : Grupperna tycker olika. Vår observation av teststorheten ges av

$$q = \frac{(59-68)^2}{68} + \frac{(41-32)^2}{32} + \frac{(145-136)^2}{136} + \frac{(55-64)^2}{64} \approx 5.58.$$

Detta är en observation av $Q \overset{\text{appr.}}{\approx} \chi^2(1 \cdot 1)$. Ur tabell finner vi att $P(Q > 6.6349) = 0.01$. Eftersom $q < 6.6349$ så kan vi inte förkasta hypotesen att grupperna tycker lika.

**Exempel**

Alla som lyssnar på hårdrock i någon form har säkert funderat över vilken av Slayer-låtarna *Angel of Death* och *Raining Blood* som är bäst[a]. Examinator funderade över om resultaten är homogena över några olika grupper och samlade in följande siffror på internet:

| | *Angel of Death* | *Raining Blood* |
|---|---|---|
| `Returntothepit.com` | 199 | 173 |
| `MetalStorm.net` | 47 | 43 |
| `RockBand.com` | 21 | 16 |
| `MetalRules.com` | 23 | 3 |

Utför ett homogenitetstest på nivån 5% för att se om man kan förkasta hypotesen att åsikterna är likafördelade i de fyra olika grupperna.

[a]Självklart är *Angel of Death* den bästa av dessa två, men det är inte poängen!

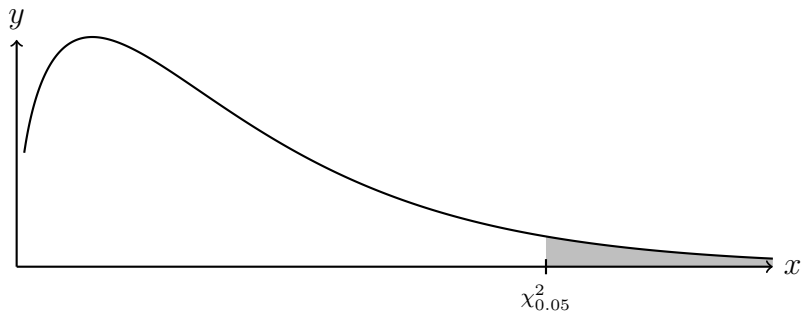
**Lösning**. Först kompletterar vi tabellen med all information som behövs:

| | *Angel of Death* | *Raining Blood* | Summa ($n_i$) |
|---|---|---|---|
| `Returntothepit.com` | 199 | 173 | 372 |
| `MetalStorm.net` | 47 | 43 | 90 |
| `RockBand.com` | 21 | 16 | 37 |
| `MetalRules.com` | 23 | 3 | 26 |
| Summa | 290 | 235 | 525 |
| $\widehat{p_j}$ (skattat $p_j$) | $\widehat{p_1} = 0.552$ | $\widehat{p_2} = 0.448$ | 1.00 |

Vi beräknar observationen $q$

$$\begin{aligned} q &= \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{(x_{ij} - n_i\widehat{p_j})^2}{n_i\widehat{p_j}} \\ &= \frac{(199-372\widehat{p_1})^2}{372\widehat{p_1}} + \frac{(173-372\widehat{p_2})^2}{372\widehat{p_2}} + \frac{(47-90\widehat{p_1})^2}{90\widehat{p_1}} + \frac{(43-90\widehat{p_2})^2}{90\widehat{p_2}} \\ &\quad + \frac{(21-37\widehat{p_1})^2}{37\widehat{p_1}} + \frac{(16-37\widehat{p_2})^2}{37\widehat{p_2}} + \frac{(23-26\widehat{p_1})^2}{26\widehat{p_1}} + \frac{(3-26\widehat{p_2})^2}{26\widehat{p_2}} \\ &= 12.43 \end{aligned}$$

Låt $H_0$ vara utsagan att favoriten bland de två låtarna är likadant fördelad i alla fyra serier. Det vill säga, att $P(\text{AoD favorit}) = p_1$ och $P(\text{RB favorit}) = p_2$ gäller i alla fyra serierna med samma sannolikheter $p_j$. Antag att $H_0$ är sann.




Vi förkastar $H_0$ om $Q > \chi^2_\alpha(3)$, d v s om den observerade testvariabeln hamnar utanför det skuggade området i figuren ovan. Med $\alpha = 0.05$ finner vi att $\chi^2_{0.05}(3) = 7.81$ (ur en tabell eller med matlab), så $Q > \chi^2_\alpha(3)$. Vi kan alltså förkasta hypotesen att alla grupperna tycker likadant (ganska tydligt från siffrorna att den fjärde raden skiljer sig markant från de andra).

**Svar**: Vi kan förkasta hypotesen om homogenitet på nivån 5%.

# 6 Skattningar för kovarians och korrelation

Om vi har ett stickprov $(x_k, y_k)$, $k = 1, 2, \ldots, n$, där $(X_k, Y_k)$ är stokastiska variabler med samma fördelning, så skattar vi kovariansen $C$ med

$$\widehat{c} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

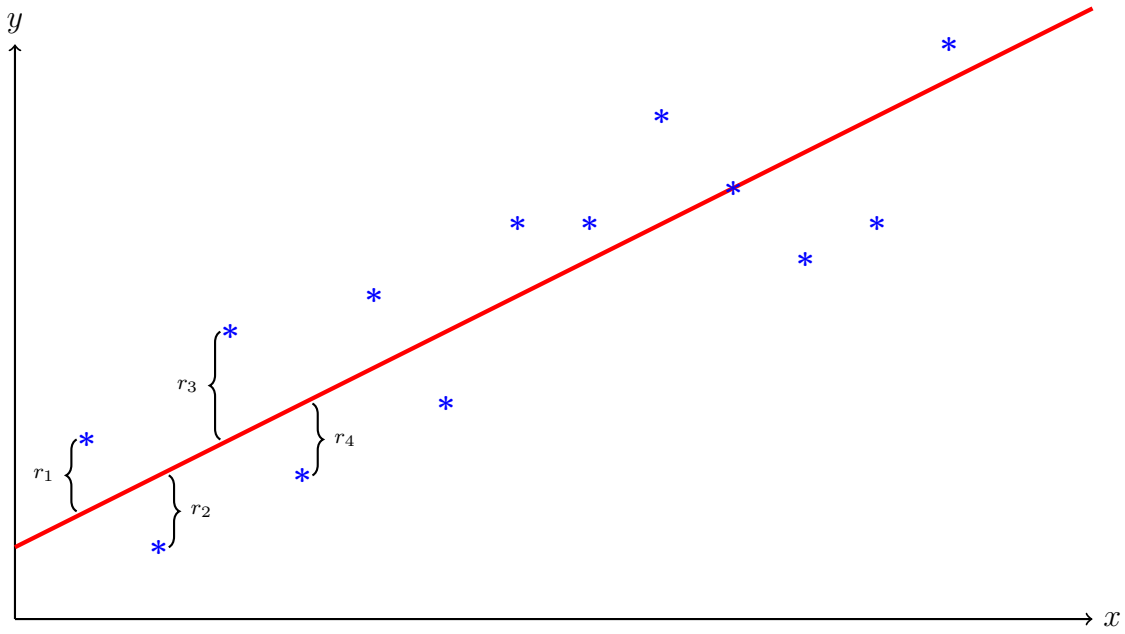
och korrelationen med

$$\widehat{\rho} = \frac{\widehat{c}}{s_x\, s_y} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{\left(\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2\right)^{1/2}\left(\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y})^2\right)^{1/2}}.$$

Av tradition betecknar man ofta $\widehat{\rho} = r$. En naturlig fråga i detta skede är om vi kan säga något om fördelningen för den skattade korrelationen under något lämpligt antagande om det slumpmässiga stickprovet. I vissa fall kan man det, men det är inget vi ger oss in på i denna kurs.

# 7 Enkel linjär regression

Vi återgår nu till ett exempel vi stött på redan vid ett flertal tillfällen (om inte annat så i tidigare kurser som linjär algebra), nämligen att anpassa en rät linje $y = \beta_0 + \beta_1 x$ efter mätdata $(x_i, y_i)$, $i = 1, 2, \ldots, n$. Grafiskt illustrerat enligt nedan.




Målsättningen är att – givet en mätserie – hitta den linje som approximerar denna serie på lämpligt sätt. Det resulterar i ett par naturliga funderingar.

(i) Hur hittar man en approximativ linje systematiskt?

(ii) Om man upprepar försöket, får man samma linje?

(iii) I vilken mening är linjen optimal? På vilket sätt mäter vi avvikelserna mellan linjen och mätserien?

Vi ska försöka svara på dessa frågor och för att göra det behöver vi ställa upp en modell.

## Enkel linjär regression

**Definition.** Vi kommer betrakta följande modell: givet $(x_j, y_j)$, $j = 1, 2, \ldots, n$, där vi betraktar $x_j$ som fixerade och $y_j$ som observationer av stokastiska variabler

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

där $\epsilon_j \sim N(0, \sigma)$ antas oberoende (och likafördelade). Den räta linjen $y = \beta_0 + \beta_1 x$ kallas **regressionslinjen**.

Vi använder beteckningen $\mu_j = \beta_0 + \beta_1 x_j = E(Y_j)$.

Är det givet att denna modell är sann? Nej, det är inte självklart utan hänger på vilka förutsättningar datan kommer från. Däremot tenderar modellen att fungera bra i de flesta fall om vi har en rimlig mängd observationer. Med notationen från figuren ser vi att

$$r_j = y_j - \mu_j$$

om vi tar hänsyn till tecknet (positivt tecken om $y_j$ ligger ovanför regressionslinjen). Ett mått på hur väl linjen approximerar mätserien ges av kvadratsumman

$$\sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - \mu_j)^2.$$

Vi skulle kunna minimera denna summa med avseende på $(\beta_0, \beta_1)$, vilket ger MK-skattningar för $\beta_0$ och $\beta_1$. Detta var det som hände i exemplet från föreläsning 7. Istället för att upprepa argumentet går vi över till den generella modellen för linjär regression.

Dubbelindexeringen är lite jobbig att arbeta med, så låt oss gå över till matrisnotation:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 x_1 \\ \beta_0 & \beta_1 x_2 \\ \vdots & \vdots \\ \beta_0 & \beta_1 x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Vi låter

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \text{samt } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

vilket leder till sambandet

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Således gäller att

$$E(\boldsymbol{Y}) = X\boldsymbol{\beta} \quad \text{och} \quad C_{\boldsymbol{Y}} = \sigma^2 I_2,$$

där $I_2$ är den $n$-dimensionella enhetsmatrisen och $C_{\boldsymbol{Y}}$ är kovariansmatrisen för $\boldsymbol{Y}$. Vi söker MK-skattningen $\widehat{\boldsymbol{\beta}}$ för $\boldsymbol{\beta}$, vilket vi kan erhålla genom att minimera den kvadratiska formen

$$Q(\beta_0, \beta_1) = \sum_{j=1}^{n} (y_j - \mu_j)^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}),$$

där $\boldsymbol{y} = (y_1 \ y_2 \ \cdots \ y_n)^T$. En variant är att sätta igång och derivera, men lite mer elegant gäller följande sats (välkänd från linjär algebra).

## Normalekvationerna

**Sats.** Om det $X^T X \neq 0$ så ges MK-skattningen av $\boldsymbol{\beta}$ enligt

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Med förutsättningarna ovan gäller att

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta},\ \sigma^2 (X^T X)^{-1}\right).$$

Detta vektorvärda samband innebär att $\widehat{\beta}_0 \sim N(\beta_0,\ \sigma\sqrt{h_{00}})$ och $\widehat{\beta}_1 \sim N(\beta_1,\ \sigma\sqrt{h_{11}})$, där

$$(X^T X)^{-1} = \begin{pmatrix} h_{00} & h_{01} \\ h_{10} & h_{11} \end{pmatrix}.$$

När det gäller $\widehat{\boldsymbol{\beta}}$ och dess komponenter kommer vi använda samma beteckningar för observationer av punktskattningen och den stokastiska variabeln. Skulle vi vara konsekventa borde den stokastiska variabeln betecknas $\widehat{\boldsymbol{B}}$, men av tradition görs inte så. Var observant!

# 8 Variansanalys

Vi låter

$$\widehat{\mu_j} = \widehat{\beta_0} + \widehat{\beta_1} x_j, \quad j = 1, 2, \ldots, n.$$

Ibland betecknas vektorn $\widehat{\boldsymbol{\mu}}$ som $\widehat{\boldsymbol{y}}$ eftersom det i någon mening är en skattningen av $\boldsymbol{y}$, men det blir lite olyckligt för det är inte $\boldsymbol{y}$ vi skattar. Vi ska nu studera hur bra den skattning vi tagit fram är.

**Regressionsanalysens kvadratsummor**

**Definition.** Vi definierar tre kvadratsummor som beskriver variationen hos $y$-värdena:

(i) Den **totala variationen** definieras enligt $\mathrm{SS_{TOT}} = \sum_{j=1}^{n} (y_j - \overline{y})^2$.

(ii) Variationen som förklaras av $x_1, x_2, \ldots, x_k$, definieras enligt $\mathrm{SS_R} = \sum_{j=1}^{n} (\widehat{\mu_j} - \overline{y})^2$.

(iii) Variationen som inte förklaras av regressionsmodellen är $\mathrm{SS_E} = \sum_{j=1}^{n} (y_j - \widehat{\mu_j})^2$.

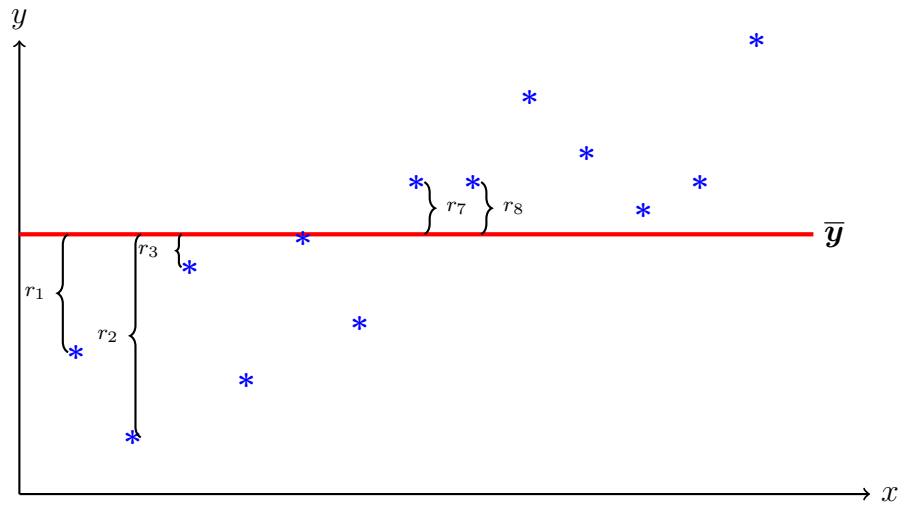
Ibland används beteckningarna $Q_{\mathrm{TOT}}$ för $\mathrm{SS_{TOT}}$, $Q_{\mathrm{REGR}}$ för $\mathrm{SS_R}$ och $Q_{\mathrm{RES}}$ för $\mathrm{SS_E}$. Hur hänger dessa summor ihop? Som tur är finns ett enkelt svar.

**Sats.** Den totala variationen $\mathrm{SS_{TOT}}$ kan delas upp enligt

$$\mathrm{SS_{TOT}} = \mathrm{SS_R} + \mathrm{SS_E}.$$

## 8.1 $\mathrm{SS_{TOT}}$

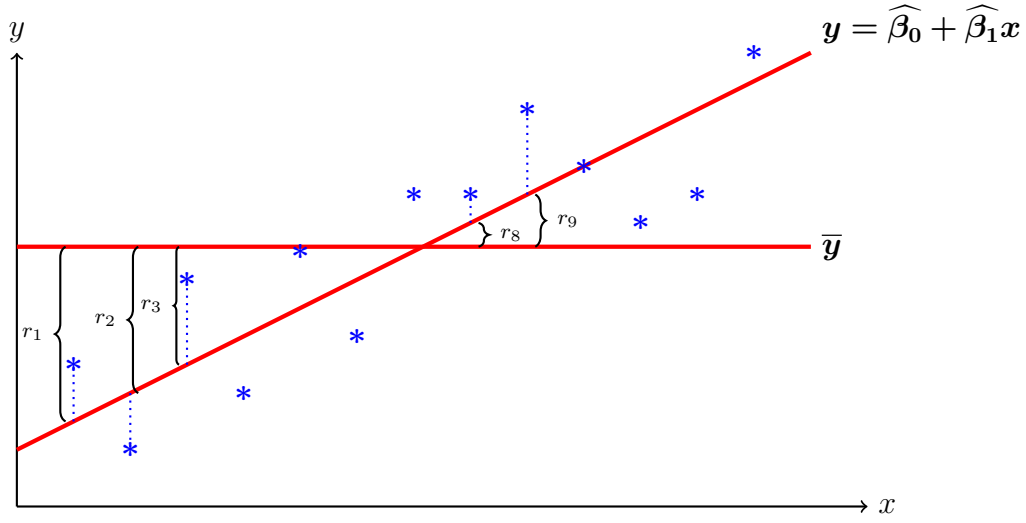
Storheten $\mathrm{SS_{TOT}}$ mäter den totala variation av mätvärden jämfört med mätvärdenas medelvärde. I fallet då vi använder modellen $Y = \beta_0 + \epsilon$ ger således $\mathrm{SS_{TOT}}$ hela felet i regressionen.




Vi ser att $\mathrm{SS_{TOT}} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - \overline{y})^2$.

## 8.2 $\mathrm{SS_R}$ och $\mathrm{SS_E}$

Om vi istället använder modellen $y = \beta_0 + \beta_1 x + \epsilon$ blir summorna $\mathrm{SS_E}$ och $\mathrm{SS_R}$ relevanta ($\mathrm{SS_{TOT}}$ ser fortfarande ut som ovan). Låt oss först illustrera $\mathrm{SS_R}$.

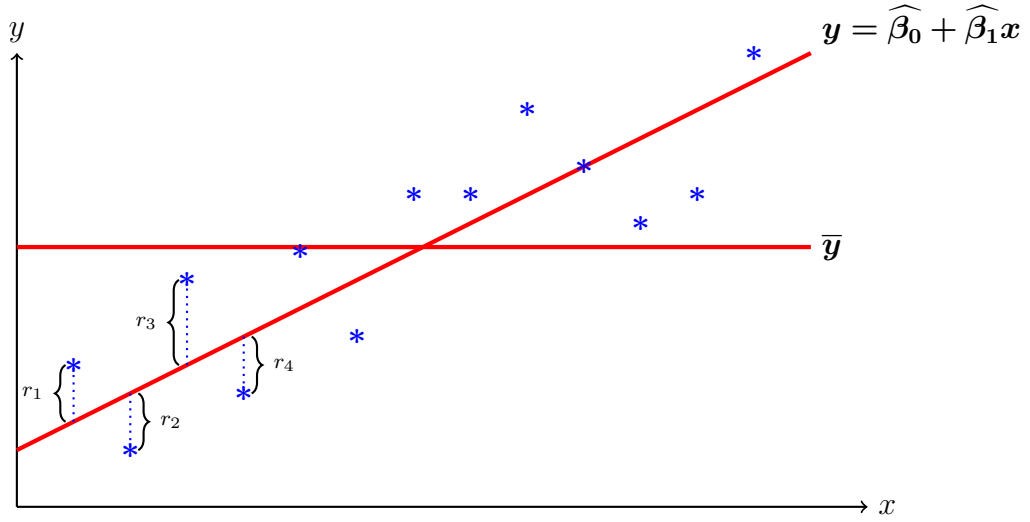



Vi ser att

$$\mathrm{SS_R} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 x_j - \overline{y})^2$$

och $\mathrm{SS_R}$ mäter alltså hur mycket den skattade regressionslinjen (i mätpunkterna $x_j$) skiljer sig från medelvärdet av $y$-värdena.

När det gäller $\mathrm{SS_E}$ är det istället skillnaden mellan den skattade regressionslinjen och mätvärdena vi betraktar.




Kvadratsumman av dessa avvikelser blir

$$\mathrm{SS_E} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - (\widehat{\beta}_0 + \widehat{\beta}_1 x_j))^2.$$

# 9 Hypotestester och konfidensintervall

Vi har nu samlat på oss en ordentlig verktygslåda, så det kanske är dags att se hur vi använder de olika delarna.

## 9.1 Skattning av $\sigma^2$

Variansen $\sigma^2$ skattar vi med

$$s^2 = \frac{\mathrm{SS_E}}{n-2}.$$

Denna skattning är väntevärdesriktig:

$$E(S^2) = \frac{\sigma^2}{n-2} E\left(\frac{\mathrm{SS_E}}{\sigma^2}\right) = \sigma^2 \frac{n-2}{n-2} = \sigma^2$$

ty $\frac{1}{\sigma^2}\mathrm{SS_E} \sim \chi^2(n-2)$. Det faktum att vi får en $\chi^2(n-2)$-fördelning är inte självklart; se avsnitt 12.

## 9.2 Enskilda koefficienter

Med beteckningen

$$(X^T X)^{-1} = \begin{pmatrix} h_{00} & h_{01} \\ h_{10} & h_{11} \end{pmatrix}$$

så är $\widehat{\beta}_i \sim N(\beta_i, \sigma\sqrt{h_{ii}})$. Om vi känner $\sigma^2$ kan vi använda att

$$Z = \frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{h_{ii}}} \sim N(0,1)$$

för att testa hypotesen $H_0 : \beta_i = 0$ eller för att ställa upp konfidensintervall $I_{\beta_i}$.

Nu är det extremt sällan vi känner $\sigma^2$ exakt, men vi vet att $\widehat{\boldsymbol{\beta}}$ är oberoende av $\mathrm{SS_E}$ enligt huvudsatsen (åter igen refererar vi till avsnitt 12), så $\widehat{\beta}_i$ är oberoende av $\mathrm{SS_E}$. Vidare är $s^2 = \mathrm{SS_E}/(n-2)$ en skattning av $\sigma^2$ vi känner väl, så

$$\frac{(\widehat{\beta}_i - \beta_i)/(\sigma\sqrt{h_{ii}})}{\sqrt{((n-2)S^2/\sigma^2)/(n-2)}} = \frac{\widehat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-2)$$

enligt Gossets sats.

## 9.3 Hypotestest: $H_0 : \beta_i = 0$

Vi kan även utföra hypotestester för en enskild koefficient $\beta_i$ för att se om den förklaringsvariabeln tillför något signifikant (givetvis går det att ställa upp konfidensintervall också för mer kvantitativt innehåll).

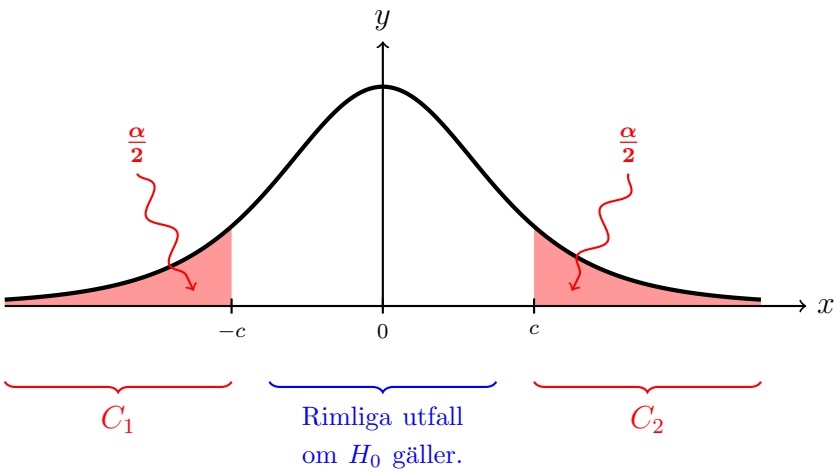
Låt $H_0 : \beta_i = 0$ och $H_1 : \beta_i \neq 0$. Vi vet enligt ovan att

$$T = \frac{\widehat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-2),$$

så det kritiska området finner vi enligt

$$C = \{t \in \mathbf{R} : |t| > c\}$$

för lämpligt tal $c > 0$ beroende på signifikansnivån och antalet frihetsgrader.




Gränsen hittar vi i tabell genom att leta reda på ett tal $c = F_T^{-1}(1-\alpha/2)$ (sitter du med MATLAB kan du använda `c = -tinv(alpha/2)`).

## 9.4 Formler utan matriser

Vi diskuterade enkel linjär regression tidigare i samband med MK-skattningar (föreläsning 7). Låt oss visa att vi får samma resultat med den metod vi nu tagit fram (och fördelningar för ingående storheter på ett enkelt sätt).

Vi låter $x_1, x_2, \ldots, x_n$ vara fixerade tal och $y_1, y_2, \ldots, y_n$ vara observationer från

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

där $\epsilon_i \sim N(0,\sigma)$ är oberoende. Alltså precis den modell vi använt tidigare. Syntesmatrisen $X$ ges av

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

så

$$X^T X = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \Rightarrow \quad (X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}$$

om $\det(X^T X) \neq 0$. Således blir

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} &= (X^T X)^{-1} X^T \boldsymbol{y} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix} \begin{pmatrix} n\overline{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} n\overline{y}\sum_{i=1}^n x_i^2 - n\overline{x}\sum_{i=1}^n x_i y_i \\ -n^2\overline{x}\overline{y} + n\sum_{i=1}^n x_i y_i \end{pmatrix}, \end{aligned}$$

vilket ger att

$$\widehat{\beta}_1 = \frac{\sum_{j=1}^n x_j y_j - n\overline{x}\,\overline{y}}{\sum_{j=1}^n x_j^2 - n\overline{x}^2} = \frac{\sum_{j=1}^n y_j(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} = \frac{\sum_{j=1}^n (y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} \quad \text{och} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Det följer nu att

$$\widehat{\beta}_0 \sim N\left(\beta_0,\, \sqrt{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}}\right) \quad \text{och} \quad \widehat{\beta}_1 \sim N\left(\beta_1,\, \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right).$$

Någonstans nu inser vi vilket kraftig syntaktiskt verktyg matriser och linjär algebra är...

Eftersom $\sigma$ är okänd skattar vi $\sigma^2$ med

$$s^2 = \frac{\mathrm{SS_E}}{n-2} = \frac{\sum_{j=1}^n (y_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)^2}{n-2}$$

där $S^2 \sim \chi^2(n-2)$. Vi kan skriva om $\mathrm{SS_E}$ genom att utnyttja att

$$\begin{aligned} \mathrm{SS_R} = \sum_{j=1}^n (\widehat{\mu}_j - \overline{y})^2 = {\widehat{\beta}_1}^2 \sum_{j=1}^n (x_j - \overline{x})^2 &= \left(\frac{\sum_{j=1}^n (y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2}\right)^2 \sum_{j=1}^n (x_j - \overline{x})^2 \\ &= r^2 \sum_{j=1}^n (y_i - \overline{y})^2 \end{aligned}$$

så

$$\mathrm{SS_E} = \mathrm{SS_{TOT}} - \mathrm{SS_R} = (1 - r^2) \sum_{j=1}^n (y_i - \overline{y})^2.$$

Vi kan här se att $r = 1$ ger perfekt matchning så alla $(x_j, y_j)$ ligger på en rät linje.

# 10 Ett löst exempel

**Exempel**

Vid ett experiment där man mäter hårdheten på stål som funktion av kolhalten finner man följande data.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Kolhalt (%) $x_i$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | 1.2 | 1.4 |
| Hårdhet (DPH) $y_i$ | 66 | 116 | 129 | 175 | 209 | 238 | 269 | 301 |

Använd modellen att $y_i$ är oberoende observationer av $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma)$.

1. Vad blir ekvationen för den skattade regressionslinjen?
2. Skatta $\sigma^2$ med en väntevärdesriktig skattning.
3. Utför hypotestestet $H_0 : \beta_1 = 0$ med mothypotesen $H_1 : \beta_1 > 0$ på nivån 95%.
4. Beräkna ett 99% konfidensintervall för $\beta_1$.
5. Beräkna ett 99% prediktionsintervall för hårdheten när kolhalten är 1.3%.

Typiskt när en liknande uppgift dyker upp på tentan så får man lite räknehjälp:

$$\overline{x} = 0.7, \qquad \overline{y} = 187.875,$$

$$\sum_{i=1}^{8}(x_i - \overline{x})^2 = 1.680, \qquad \sum_{i=1}^{8}(y_i - \overline{y})^2 = 45988.9, \qquad \sum_{i=1}^{8}(x_i - \overline{x})(y_i - \overline{y}) = 277.1,$$

| $i$ | $\widehat{\beta}_i$ | $d(\widehat{\beta}_i)$ |
|---|---|---|
| 0 | 72.42 | 4.44 |
| 1 | 164.94 | 5.31 |

Variansanalys

| | Frihetsgrader | Kvadratsumma |
|---|---|---|
| REGR | 1 | 45705.01 |
| RES | 6 | 283.87 |
| TOT | 7 | 45988.88 |

$$(X^T X)^{-1} = \begin{pmatrix} 0.4167 & -0.4167 \\ -0.4167 & 0.5952 \end{pmatrix} = \begin{pmatrix} h_{00} & h_{01} \\ h_{10} & h_{11} \end{pmatrix}.$$

Ett par formler som kanske är användbara:

$$\widehat{\beta}_i \sim N(\beta_i, \sigma\sqrt{h_{ii}}), \quad \frac{\widehat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-2), \quad S^2 = \frac{\mathrm{SS_E}}{n-2}.$$

**Lösning.** Syntesmatrisen för vår modell finner vi enligt

$$X = \begin{pmatrix} 1.00 & 0 \\ 1.00 & 0.20 \\ 1.00 & 0.40 \\ 1.00 & 0.60 \\ 1.00 & 0.80 \\ 1.00 & 1.00 \\ 1.00 & 1.20 \\ 1.00 & 1.40 \end{pmatrix}$$

Vi kommer inte behöva använda denna matris direkt, utan kan använda räknehjälpen där vi fått

$$(X^T X)^{-1} = \begin{pmatrix} 0.4167 & -0.4167 \\ -0.4167 & 0.5952 \end{pmatrix}.$$

Om man inte tycker om den framställningen kan man direkt använda formlerna från formelsamlingen (som utgår från $S_{xy}$ etc.). Ni kommer få siffror för att använda båda varianterna.

1. Vi har $\widehat{\boldsymbol{\beta}}$ direkt i tabell ovan, så

   $$y = \beta_0 + \beta_1 x = 72.42 + 164.94x.$$

   Alternativt får vi räkna ut koefficienterna med formelsamlingen:

   $$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{277.1}{1.68} = 164.94$$

   och

   $$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 72.42.$$

2. Den naturliga skattningen för $\sigma^2$ finner vi med

   $$s^2 = \frac{\mathrm{SS_E}}{n-2} = \frac{283.87}{6} = 47.31,$$

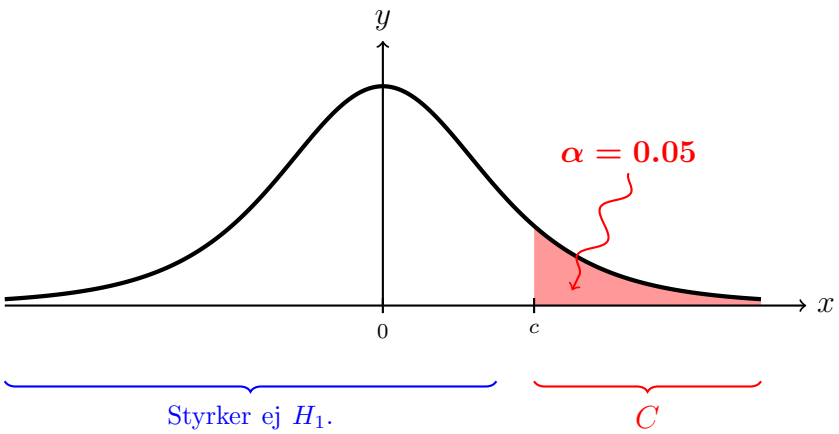
   alternativt ur formelsamlingen (med $S_{yy} = \mathrm{SS_{TOT}} = 45988.9$, $S_{xx} = 1.68$ och $S_{xy} = 277.1$)

   $$s^2 = \frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2} = \frac{45988.9 - 277.1^2/1.680}{6} = 47.32.$$

3. Antag att $H_0$ är sann. Enligt ovan så gäller då att

   $$T = \frac{\widehat{\beta}_1 - 0}{S\sqrt{h_{11}}} \sim t(8-2) = t(6).$$

   Det kritiska området väljer vi som $C = [c, \infty[$ där $c$ är ett tal så att $P(T > c) = 0.05$.



Ur tabell finner vi $c = 1.94$. Vi har som observation av $T$

$$t = \frac{164.94}{\sqrt{47.31}\sqrt{0.5952}} = 31.08 \in C$$

så observationen ligger i det kritiska området. Vi förkastar således $H_0$ och anser att $H_1$ är styrkt (dvs att $\beta_1 > 0$).

Alternativt så använder vi teststorheten ur formelsamlingen:

$$T = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}/\sqrt{S_{xx}}} \sim t(6)$$

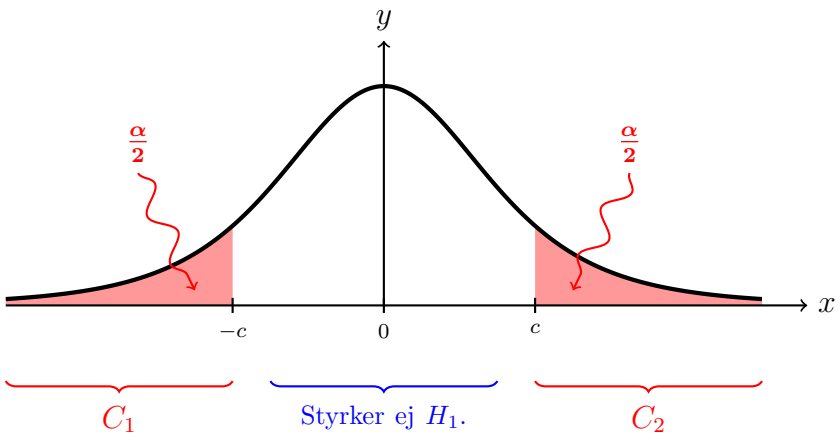
där vi som observation finner att

$$t = \frac{\widehat{\beta}_1}{\sqrt{47.32}/\sqrt{1.68}} = \frac{164.94}{\sqrt{47.32}/\sqrt{1.68}} = 31.08.$$

4. För att ställa upp ett konfidensintervall för $\beta_1$ så använder vi samma teststorhet som ovan, så

$$T = \frac{\widehat{\beta}_1 - \beta_1}{S\sqrt{h_{11}}} \sim t(8-2) = t(6) \qquad \text{eller} \qquad T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/\sqrt{S_{xx}}} \sim t(8-2) = t(6).$$

Vi stänger in $T$ så att $P(-c < T < c) = 0.99$ genom att lägga $\alpha/2 = 0.01/2 = 0.005$ i varje svans.




Genom att lösa ut $\widehat{\beta}_1$ ur olikheten $-c < T < c$ så finner vi endera att

$$\widehat{\beta}_1 - S \cdot c\sqrt{h_{11}} < \beta_1 < \widehat{\beta}_1 + S \cdot c\sqrt{h_{11}}$$

eller att

$$\widehat{\beta}_1 - S \cdot c/\sqrt{S_{xx}} < \beta_1 < \widehat{\beta}_1 + S \cdot c/\sqrt{S_{xx}},$$

beroende på vilken framställning vi använder. Med $c = 3.71$ och $s = \sqrt{47.31} = 6.88$ så får vi intervallet

$$I_{\beta_1} = ]145.3, 184.6[.$$

5. Prediktionsintervall (i samband med linjär regression) kommer inte att dyka upp på tentan, men för fullständighetens skull så gör man som i föreläsning 8 men vi behöver hantera lite kovariansproblem som ställer till det. Detta leder till en formel av följande typ:

$$\widehat{\beta}_9 + \widehat{\beta}_1 \cdot x \pm t_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

där $s$ skattas som tidigare och $x = 1.3$ i detta fall.

# 11 ($\star$)Bevis för vissa resultat i linjär regression

**Sats.** Med förutsättningarna ovan gäller att

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta},\ \sigma^2(X^TX)^{-1}\right).$$

**Bevis**. Det faktum att $\widehat{\boldsymbol{\beta}}$ är normalfördelad följer direkt från faktumet att elementen i $\widehat{\boldsymbol{\beta}}$ är linjärkombinationer av normalfördelade variabler. Återstår att visa väntevärde och kovarians:

$$E(\widehat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^TE(\boldsymbol{Y}) = (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}$$

och

$$\begin{aligned} C_{\widehat{\boldsymbol{\beta}}} &= (X^TX)^{-1}X^TC_{\boldsymbol{Y}}((X^TX)^{-1}X^T)^T = (X^TX)^{-1}X^T\sigma^2I_n((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\sigma^2I_n(X^T)^T((X^TX)^{-1})^T = \sigma^2(X^TX)^{-1}X^TX((X^TX)^T)^{-1} = \sigma^2(X^TX)^{-1}. \end{aligned}$$

Således blir fördelningen precis som beskriven i satsen. □

**Sats.** Den totala variationen $\mathrm{SS_{TOT}}$ kan delas upp enligt

$$\mathrm{SS_{TOT}} = \mathrm{SS_R} + \mathrm{SS_E}.$$

**Bevis**. Vi vill uttrycka $\mathrm{SS_{TOT}}$ i termer av $\mathrm{SS_R}$ och $\mathrm{SS_E}$, så

$$\begin{aligned} \mathrm{SS_{TOT}} &= \sum_{j=1}^n (y_j - \overline{y})^2 = \sum_{j=1}^n (y_j - \widehat{\mu}_j + \widehat{\mu}_j - \overline{y})^2 \\ &= \sum_{j=1}^n (y_j - \widehat{\mu}_j)^2 + 2\sum_{j=1}^n (y_j - \widehat{\mu}_j)(\widehat{\mu}_j - \overline{y}) + \sum_{j=1}^n (\widehat{\mu}_j - \overline{y})^2 \\ &= \mathrm{SS_R} + 2\sum_{j=1}^n (y_j - \widehat{\mu}_j)\widehat{\mu}_j - 2\overline{y}\sum_{j=1}^n (y_j - \widehat{\mu}_j) + \mathrm{SS_E}. \end{aligned}$$

Vi visar att de båda summorna i mitten summerar till noll. Eftersom $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$ gäller det att

$$\begin{aligned} \sum_{j=1}^n (y_j - \widehat{\mu}_j)\widehat{\mu}_j &= \widehat{\boldsymbol{\mu}}^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = (X\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}}^TX^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\boldsymbol{\beta}}^T(X^T\boldsymbol{y} - X^TX(X^TX)^{-1}X^T\boldsymbol{y}) = \widehat{\boldsymbol{\beta}}^T(X^T\boldsymbol{y} - X^T\boldsymbol{y}) = 0. \end{aligned}$$

Med andra ord är $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ vinkelrät mot $\widehat{\boldsymbol{\mu}}$.

Vidare vet vi att $\widehat{\boldsymbol{\beta}}$ minimerar $Q(\boldsymbol{\beta}) = \sum_{j=1}^n (y_j - \mu_j)^2$, så

$$0 = \frac{\partial}{\partial_{\beta_0}} Q(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = -2\sum_{j=1}^n (y_j - \widehat{\mu}_j) \quad \Leftrightarrow \quad \sum_{j=1}^n y_j = \sum_{j=1}^n \widehat{\mu}_j,$$

vilket var precis det vi behövde. □

# 12 ($\star$)Regressionsanlysens huvudsats

Eftersom vi arbetar med matriser och MK-lösningen i princip är projektionen på ett underrum av $\mathbf{R}^n$ så är följande resultat inte allt för förvånande.

**Hatt-matrisen**

**Definition.** Vi definierar $H = X(X^TX)^{-1}X^T$. Vi definierar även $J$ som en matris vars samtliga element är 1. Vi skriver $J_{nm}$ om vi vill markera dimensionen.

Varför kallar vi $H$ för hatt-matrisen? Ganska enkelt:

$$H\boldsymbol{y} = X(X^TX)^{-1}X^T\boldsymbol{y} = X\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{y}}.$$

Avbildningen sätter alltså hatten på!

**Sats.** Matriserna $H$, $I-H$ och $H-\frac{1}{n}J$ är projektionsmatriser $P$ som uppfyller $P^2 = P^T = P$.

**Bevis**. Direkt från definitionen av $H$ blir

$$H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}X^T = H$$

och

$$H^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-T}X^T = X((X^TX)^T)^{-1}X^T = H.$$

Därav följer det att $(I-H)^2 = I^2 - 2H + H^2 = I - H$ och att $(I-H)^T = I^T - H^T = I - H$. För den sista operatorn skriver vi

$$\left(H - \frac{1}{n}J\right)^2 = H^2 - \frac{1}{n}HJ - \frac{1}{n}JH + \frac{1}{n^2}J^2 = H - \frac{1}{n}HJ - \frac{1}{n}JH + \frac{1}{n}J$$

ty $J^2$ är matrisen med samtliga element lika med $n$. Vidare gäller att $J$ kommuterar med alla kvadratiska matriser, så $JH = HJ$. Här kan vi se att $H\mathbf{1} = \mathbf{1}$ (där $\mathbf{1}$ är en vektor med samtliga element lika med 1) eftersom lösningen till regressionsproblemet om $\boldsymbol{y}$ är konstant är just den konstanten (lösningen är exakt). Alltså blir

$$\left(H - \frac{1}{n}J\right)^2 = H - \frac{1}{n}J.$$

Givetvis gäller även att $\left(H - \frac{1}{n}J\right)^T = H - \frac{1}{n}J.$ □

En följdsats av detta är att vi kan skriva kvadratsummorna som kvadratiska former.

**Sats.** $\mathrm{SS_{TOT}} = \boldsymbol{y}^T\left(I - \frac{1}{n}J\right)\boldsymbol{y}$, $\mathrm{SS_R} = \boldsymbol{y}^T\left(H - \frac{1}{n}J\right)\boldsymbol{y}$, samt $\mathrm{SS_E} = \boldsymbol{y}^T(I-H)\,\boldsymbol{y}$.

Innan vi ger oss in på huvudsatsen visar vi en hjälpsats från linjär algebra.

**Sats.** Låt $\boldsymbol{x} = (x_1\ x_2\ \cdots\ x_n)$ vara sådan att

$$\boldsymbol{x}^T\boldsymbol{x} = \sum_{j=1}^{n} x_j^2 = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T B\boldsymbol{x}$$

för $A, B \in \mathbf{R}^{n\times n}$ positivt semi-definita och symmetriska med $\operatorname{rank}(A) = r$ och $\operatorname{rank}(B) = n-r$ för något heltal $r$ så att $0 < r < n$. Då finns en ON-matris $C$ så att med $\boldsymbol{x} = C\boldsymbol{y}$ gäller att

$$\boldsymbol{x}^T A\boldsymbol{x} = \sum_{j=1}^{r} y_j^2 \quad \text{och} \quad \boldsymbol{x}^T B\boldsymbol{x} = \sum_{j=r+1}^{n} y_j^2.$$

**Bevis**. Eftersom $A$ är positivt semi-definit har $A$ endast icke-negativa egenvärden som vi ordnar enligt $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Vi vet även att $\operatorname{rank}(A) = r$, så $\lambda_{r+1} = \cdots = \lambda_n = 0$. Vidare är $A$ diagonaliserbar eftersom $A$ är symmetrisk, så det finns en ON-matris $C$ så att

$$C^T AC = D = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \vdots & \ddots & \vdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Med $\boldsymbol{x} = C\boldsymbol{y}$ gäller då att

$$\boldsymbol{x}^T\boldsymbol{x} = (C\boldsymbol{y})^T(C\boldsymbol{y}) = \boldsymbol{y}^T C^T C\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y}$$

och

$$\boldsymbol{x}^T A\boldsymbol{x} = (C\boldsymbol{y})^T AC\boldsymbol{y} = \boldsymbol{y}^T C^T AC\boldsymbol{y} = \boldsymbol{y}^T D\boldsymbol{y} = \sum_{j=1}^{r} \lambda_j y_j^2.$$

Vi har även

$$\boldsymbol{x}^T B\boldsymbol{x} = (C\boldsymbol{y})^T BC\boldsymbol{y} = \boldsymbol{y}^T C^T BC\boldsymbol{y}$$

och då blir

$$\sum_{j=1}^{n} y_j^2 = \boldsymbol{y}^T\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T B\boldsymbol{x} = \sum_{j=1}^{r} \lambda_j y_j^2 + \boldsymbol{y}^T C^T BC\boldsymbol{y}$$

så

$$\boldsymbol{y}^T C^T BC\boldsymbol{y} = \sum_{j=1}^{r} (1-\lambda_j) y_j^2 + \sum_{j=r+1}^{n} y_j^2.$$

Det faktum att $\operatorname{rank}(B) = n - r$ visar att $\lambda_1 = \lambda_2 = \cdots \lambda_r = 1$ och vi erhåller identiteten

$$\boldsymbol{y}^T C^T BC\boldsymbol{y} = \sum_{j=r}^{n} y_j^2.$$

Alla beteckningar är nu ur vägen och vi är framme vid huvudresultatet.

## Regressionsanalysens huvudsats

**Sats.** Med förutsättningarna ovan gäller följande för $\mathrm{SS_E}$ och $\mathrm{SS_R}$ sedda som stokastiska variabler.

(i) $\dfrac{\mathrm{SS_E}}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n}(Y_j - \widehat{\mu}_j)^2 \sim \chi^2(n-2).$

(ii) Givet att $\beta_1 = 0$ är $\dfrac{\mathrm{SS_R}}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n}(\widehat{\mu}_j - \overline{Y})^2 \sim \chi^2(1).$

(iii) Både $\mathrm{SS_R}$ och $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{Y}$ är oberoende av $\mathrm{SS_E}$.

**Bevis**. Eftersom $H$ är en projektionsmatris har $H$ endast egenvärdena $\lambda = 0$ och $\lambda = 1$. Således gäller det finurliga att matrisens rang är lika med summan av egenvärdena, vilka kan beräknas genom att ta spåret[1] av matrisen:

$$\mathrm{rank}(H) = \mathrm{tr}(H) = \mathrm{tr}(X(X^TX)^{-1}X^T) = \mathrm{tr}(X^TX(X^TX)^{-1}) = \mathrm{tr}(I_2) = 2,$$

Det följer sedan att

$$\mathrm{rank}(I - H) = n - 2.$$

Eftersom $HX = X$ så gäller att

$$\boldsymbol{Y}^T(I - H)\boldsymbol{Y} = \boldsymbol{\epsilon}^T(I - H)\boldsymbol{\epsilon}.$$

Detta är användbart eftersom $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Enligt föregående hjälpsats gäller att sambandet

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T(I - H)\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^TH\boldsymbol{\epsilon}$$

medför att det finns en ON-matris $C$ så att $\boldsymbol{Z} = C\boldsymbol{\epsilon}$ reducerar likheten till

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sum_{j=1}^{n-2} Z_j^2 + \sum_{j=n-1}^{n} Z_j^2$$

där

$$\boldsymbol{\epsilon}^T(I - H)\boldsymbol{\epsilon} = \boldsymbol{Y}^T(I - H)\boldsymbol{Y} = \mathrm{SS_E} = \sum_{j=1}^{n-2} Z_j^2.$$

Eftersom komponenterna i $\boldsymbol{\epsilon}$ är oberoende och $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$ följer det att

$$E(\boldsymbol{Z}) = C\mathbf{0} = \mathbf{0} \quad \text{och} \quad C_{\boldsymbol{Z}} = C_{C\boldsymbol{\epsilon}} = \sigma^2 C^T I C = \sigma^2 I$$

så $\boldsymbol{Z} \sim N\left(\mathbf{0}, \sigma^2 I\right)$. Komponenterna i $\boldsymbol{Z}$ är alltså oberoende och $Z_j \sim N(0, \sigma^2)$. Detta medför att

$$\frac{\mathrm{SS_E}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n-2} Z_j^2 \sim \chi^2(n-2).$$

[1] se sista (bonus)avsnittet för lite detaljer kring spår av matriser.

Vi erhåller även att $\mathrm{SS_E}$ och $H\boldsymbol{\epsilon}$ är oberoende (de delar inga variabler $Z_j$). Detta medför att $\mathrm{SS_E}$ och $\widehat{\boldsymbol{\beta}}$ är oberoende. Det faktum att $\mathrm{SS_E}$ och $\mathrm{SS_R}$ är oberoende följer av att kovariansen

$$\begin{aligned} C\left(\left(H-\frac{1}{n}J\right)\boldsymbol{Y},(I-H)\boldsymbol{Y}\right) &= \left(H-\frac{1}{n}J\right)C(\boldsymbol{Y},\boldsymbol{Y})(I-H)^T = \sigma^2\left(H-\frac{1}{n}J\right)(I-H) \\ &= \sigma^2\left(H-H^2-\frac{1}{n}J+\frac{1}{n}JH\right) = \sigma^2\left(-\frac{1}{n}J+\frac{1}{n}HJ\right) \\ &= \sigma^2\left(-\frac{1}{n}J+\frac{1}{n}J\right) = 0, \end{aligned}$$

så dessa vektorer är okorrelerade och normalfördelade, så oberoende. Det följer direkt att $\mathrm{SS_E}$ och $\mathrm{SS_R}$ är oberoende (som funktioner av oberoende variabler).

Om vi fokuserar på fördelningen för $\mathrm{SS_R}$ så kan vi på samma sätt som ovan utnyttja att $\left(H-\frac{1}{n}J\right)$ är en projektionsmatris, så

$$\operatorname{rank}\left(H-\frac{1}{n}J\right) = \operatorname{tr}\left(H-\frac{1}{n}J\right) = \operatorname{tr}(H)-\operatorname{tr}\left(\frac{1}{n}J\right) = 1+1-1 = 2.$$

Matrisen $J$ är synnerligen rangdefekt med $\operatorname{rank}(J)=1$ (eftersom alla kolonner är lika spänner vi bara upp ett en-dimensionellt rum).

Nu stöter vi på lite problem. Vi ser att

$$\left(H-\frac{1}{n}J\right)\boldsymbol{Y} = \left(I-\frac{1}{n}J\right)X\boldsymbol{\beta}+\left(H-\frac{1}{n}J\right)\boldsymbol{\epsilon}$$

där den första termen *inte* försvinner såvida inte $\beta_1=0$. Men under detta antagande har vi åter igen en situation där vi kan ”byta ut” $\boldsymbol{Y}$ mot $\boldsymbol{\epsilon}$. Föregående hjälpsats visar – analogt med föregående argument – att

$$\boldsymbol{Y}^T\left(H-\frac{1}{n}J\right)\boldsymbol{Y} = \sum_{j=1}^{1} Z_j^2,$$

där $Z_j$ är oberoende och $Z_j \sim N(0,\sigma^2)$, vilket medför att $\frac{1}{\sigma^2}\mathrm{SS_R} \sim \chi^2(k)$, under förutsättningen att $\beta_1=\beta_2=\cdots=\beta_k=0$. □

Kommentar: utan villkoret $\beta_1=\beta_2=\cdots=\beta_k=0$ kan man fortfarande genomföra argumentet, men resultatet blir en icke-centrerad $\chi^2(k)$-fördelning (något som inte ingår i kursen).

# 13 ($\star$) Bevis av $\chi^2$-testet

**Sats.** Med beteckningarna ovan gäller att $\displaystyle\sum_{j=1}^{k}\frac{(Y_j-np_j)^2}{np_j} \xrightarrow{\mathrm{D}} X$, där $X \sim \chi^2(k-1)$. Konvergensen är alltså i fördelning.

**Bevis**. Eftersom $Y_j$ är binomialfördelad vet vi att $E(Y_j) = np_j$ och $V(Y_j) = np_j(1-p_j)$, så de standardiserade variablerna

$$\frac{Y_j - np_j}{\sqrt{np_j(1-p_j)}} \xrightarrow{D} \widetilde{Z_j} \sim N(0,1),$$

för något $\widetilde{Z_j}$ enligt centrala gränsvärdessatsen (CGS). Konvergensen är i meningen att fördelningsfunktionen $F_{n,j}(y) \to \Phi(y)$ för alla $y \in \mathbf{R}$. En följd av detta är att

$$\frac{Y_j - np_j}{\sqrt{np_j}} \xrightarrow{D} Z_j \sim N(0, 1-p_j),$$

eftersom om $U_n \xrightarrow{D} U$ så gäller att $h(U_n) \xrightarrow{D} h(U)$ för alla kontinuerliga funktioner $h$ (brukar kallas sannolikhetsteorins open mapping theorem). Anledningen till den sista manövern är att vi ska få det lite lättare att analysera beroendestrukturen hos $Z_j$, $j = 1, 2, \ldots, k$. Eftersom väntevärdet är $E(Y_j) = np_j$ kommer

$$\begin{aligned} C\left(\frac{Y_i - np_i}{\sqrt{np_i}}, \frac{Y_j - np_j}{\sqrt{np_j}}\right) &= E\left(\frac{Y_i - np_i}{\sqrt{np_i}}\frac{Y_j - np_j}{\sqrt{np_j}}\right) = \frac{1}{n\sqrt{p_ip_j}}\left(E(Y_iY_j) - 2n^2p_ip_j + n^2p_ip_j\right) \\ &= \frac{1}{n\sqrt{p_ip_j}}\left(E(Y_iY_j) - n^2p_ip_j\right) \end{aligned}$$

För att beräkna $E(Y_iY_j)$ går vi tillbaka till variablerna $X_i$, $i = 1, 2, \ldots, n$. Låt $I_A$ beteckna *indikatorfunktionen* för mängden $A$. Detta innebär att

$$I_{A_j}(X_i) = \begin{cases} 1 \text{ om } X_i \in A_j, \\ 0 \text{ om } X_i \notin A_j. \end{cases}$$

Vi kan då skriva $Y_j = \sum_{i=1}^{n} I_{A_j}(X_i)$ och eftersom $X_i$ är Bernoullifördelade (2-punktsfördelade) följer det att $E(I_{A_j}(X_i)) = p_j$. Vi har nu, för $i \neq j$,

$$\begin{aligned} E(Y_iY_j) &= E\left(\left(\sum_{l=1}^{n} I_{A_i}(X_l)\right)\left(\sum_{m=1}^{n} I_{A_j}(X_m)\right)\right) = E\left(\sum_{l=1}^{n}\sum_{m=1}^{n} I_{A_i}(X_l)\, I_{A_j}(X_m)\right) \\ &= E\left(\sum_{l=1}^{n} I_{A_i}(X_l)\, I_{A_j}(X_l)\right) + E\left(\sum_{l=1}^{n}\sum_{\substack{m=1 \\ m\neq l}}^{n} I_{A_i}(X_l)\, I_{A_j}(X_m)\right) \\ &= 0 + \sum_{l=1}^{n}\sum_{\substack{m=1 \\ m\neq l}}^{n} E(I_{A_i}(X_l))\, E(I_{A_j}(X_m)) = \sum_{l=1}^{n}\sum_{\substack{m=1 \\ m\neq l}}^{n} p_ip_j = n(n-1)p_ip_j, \end{aligned}$$

eftersom $I_{A_i}(X_l)\, I_{A_j}(X_l) = 0$ (samma boll kan inte hamna i två lådor) samt att $I_{A_i}(X_l)$ och $I_{A_j}(X_m)$ är oberoende om $l \neq m$. Således blir

$$C\left(\frac{Y_i - np_i}{\sqrt{np_i}}, \frac{Y_j - np_j}{\sqrt{np_j}}\right) = -\sqrt{p_ip_j},$$

för $i \neq j$. Följaktligen måste således kovariansmatrisen för $\boldsymbol{Z} = (Z_1\ Z_2\ \cdots\ Z_k)^T$ ha utseendet

$$C_{\boldsymbol{Z}} = \begin{pmatrix} 1-p_1 & -\sqrt{p_1p_2} & -\sqrt{p_1p_3} & \cdots & -\sqrt{p_1p_k} \\ -\sqrt{p_2p_1} & 1-p_2 & -\sqrt{p_2p_3} & \cdots & -\sqrt{p_2p_k} \\ -\sqrt{p_3p_1} & -\sqrt{p_3p_2} & 1-p_3 & \cdots & -\sqrt{p_3p_k} \\ \vdots & \vdots & & \ddots & \vdots \\ -\sqrt{p_kp_1} & -\sqrt{p_kp_2} & -\sqrt{p_kp_3} & \cdots & 1-p_k \end{pmatrix}.$$

vilket kan skrivas lite mer kompakt som $C_{\boldsymbol{Z}} = I - \boldsymbol{p}\boldsymbol{p}^T$, där $\boldsymbol{p} = (\sqrt{p_1}\ \sqrt{p_2}\ \cdots\ \sqrt{p_k})^T$. Denna omskrivning gör att vi enkelt kan se att

$$(I - \boldsymbol{p}\boldsymbol{p}^T)^2 = I - \boldsymbol{p}\boldsymbol{p}^T \quad \text{och} \quad (I - \boldsymbol{p}\boldsymbol{p}^T)^T = I - \boldsymbol{p}\boldsymbol{p}^T,$$

så $I - \boldsymbol{p}\boldsymbol{p}^T$ är en projektionsmatris och har därför egenvärdena $\lambda = 0$ och $\lambda = 1$. Vi har nu att $\boldsymbol{Z} \sim N(\boldsymbol{0}, C_{\boldsymbol{Z}})$. På samma sätt som i beviset av regressionsanalysens huvudsats ser vi att

$$\operatorname{rank}(I - \boldsymbol{p}\boldsymbol{p}^T) = \operatorname{tr}(I - \boldsymbol{p}\boldsymbol{p}^T) = k - 1,$$

så $\lambda = 0$ är ett enkelt egenvärde. Matrisen är symmetrisk och positivt semidefinit, så det finns en ON-matris $C$ så att $C^T C_{\boldsymbol{Z}} C = \operatorname{diag}(1, 1, \ldots, 1, 0)$ blir en diagonalmatris. Om vi låter $\boldsymbol{W} = C\boldsymbol{Z}$ ser vi att $\boldsymbol{W} \sim N(\boldsymbol{0}, \operatorname{diag}(1, 1, \ldots, 1, 0))$ och att

$$\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{W}^T\boldsymbol{W} = \sum_{j=1}^{k-1} W_j^2,$$

där $W_j \sim N(0, 1)$ är oberoende. Denna summa är som bekant $\chi^2(k-1)$-fördelad! □