FORMELSAMLING
MATEMATISK STATISTIK, AK 9HP, FMS012
[UPPDATERAD 2007-09-12]

Sannolikhetsteori

Sannolikhetsteorins grunder

- Följande gäller för sannolikheter:
  - $0 \leq \mathsf{P}(A) \leq 1$
  - $\mathsf{P}(\Omega) = 1$
  - $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B)$, om händelserna A och B är oförenliga (disjunkta).
- Additionssatsen för två händelser: $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.
- Betingad sannolikhet: $\mathsf{P}(B \mid A) = \dfrac{\mathsf{P}(A \cap B)}{\mathsf{P}(A)}$.
- "Satsen om total sannolikhet": $\mathsf{P}(A) = \sum_{i=1}^{n} \mathsf{P}(A \mid H_i)\,\mathsf{P}(H_i)$,

  där händelserna $H_1, \ldots, H_n$ är parvis oförenliga (disjunkta) händelser och $\bigcup_{i=1}^{n} H_i = \Omega$.
- A och B är oberoende $\iff \mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B)$.
- Antalet olika sätt, m, att dra k element ur n är:
  - Med återläggning, med hänsyn till ordning: $m = n^k$
  - Med återläggning, utan hänsyn till ordning: $m = \binom{n+k-1}{k}$
  - Utan återläggning, med hänsyn till ordning: $m = n(n-1)\ldots(n-k+1)$
  - Utan återläggning, utan hänsyn till ordning: $m = \binom{n}{k}$

Endimensionella stokastiska variabler

- Fördelningsfunktion för X: $F_X(x) = \mathsf{P}(X \leq x)$.
- Sannolikhetsfunktion för en diskret s.v. X: $p_X(k) = \mathsf{P}(X = k)$.
- Täthetsfunktionen för en kontinuerlig s.v. X: $f_X(x) = \dfrac{dF_X(x)}{dx}$.
- $$\mathsf{P}(a < X \leq b) = F_X(b) - F_X(a) = \begin{cases} \displaystyle\sum_{k=a+1}^{b} p_X(k) & \text{om } X \text{ är diskret och } a \text{ och } b \text{ är heltal,} \\ \displaystyle\int_a^b f_X(x)\,dx & \text{om } X \text{ är kontinuerlig.} \end{cases}$$

Flerdimensionella stokastiska variabler

- Simultan fördelningsfunktion:

$$F_{X,Y}(x,y) = \mathsf{P}(X \le x,\ Y \le y) = \begin{cases} \displaystyle\sum_{j \le x,\, k \le y} p_{X,Y}(j,k) & \text{om } (X,Y) \text{ är diskret,} \\ \displaystyle\int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(t,u)\,dt\,du & \text{om } (X,Y) \text{ är kontinuerlig.} \end{cases}$$

- $\mathsf{P}(g(X,Y) \in A) = \displaystyle\iint_{g(x,y)\in A} f_{X,Y}(x,y)\,dx\,dy.$

- Marginell täthetsfunktion för X: $f_X(x) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$

- Att X och Y är oberoende är ekvivalent med att $F_{X,Y}(x,y) = F_X(x)\,F_Y(y)$ för alla x och y, samt
  - ∗ $p_{X,Y}(j,k) = p_X(j)\,p_Y(k)$ för alla j och k om X och Y är diskreta s.v..
  - ∗ $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ för alla x och y om X och Y är kontinuerliga s.v..

- Betingad sannolikhetsfunktion för X, givet $Y = k$: $p_{X|Y=k}(j) = \dfrac{p_{X,Y}(j,k)}{p_Y(k)}$

- Betingad täthetsfunktion för X, givet $Y = y$: $f_{X|Y=y}(x) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

Summor av stokastiska variabler

- Om X och Y är oberoende, så gäller för $Z = X + Y$,
  - ∗ $p_Z(k) = \displaystyle\sum_{i=0}^{k} p_X(i)\,p_Y(k-i),$
  - ∗ $f_Z(z) = \displaystyle\int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx.$

Väntevärden

- Väntevärdet av $g(X,Y)$:

$$\mathsf{E}(g(X,Y)) = \begin{cases} \displaystyle\sum_{j,k} g(j,k)\,p_{X,Y}(j,k) & \text{om } (X,Y) \text{ är diskret,} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy & \text{om } (X,Y) \text{ är kontinuerlig.} \end{cases}$$

- Väntevärden är linjära, dvs $\mathsf{E}(ag(X) + bh(Y)) = a\mathsf{E}(g(X)) + b\mathsf{E}(h(Y))$.

- Varians: $\mathsf{V}(X) = \mathsf{E}[(X - \mathsf{E}(X))^2] = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2$.

- Standardavvikelse: $\mathsf{D}(X) = \sqrt{\mathsf{V}(X)}$.

- Kovarians: $\mathsf{C}(X,Y) = \mathsf{E}[(X - \mathsf{E}(X))(Y - \mathsf{E}(Y))] = \mathsf{E}(XY) - \mathsf{E}(X)\,\mathsf{E}(Y)$.

- Korrelationskoefficient: $\rho(X,Y) = \dfrac{\mathsf{C}(X,Y)}{\mathsf{D}(X)\,\mathsf{D}(Y)}$.

- Kovariansen är bilinjär, dvs $C(\sum_j a_j X_j,\ \sum_k b_k Y_k) = \sum_j \sum_k a_j b_k C(X_j,\ Y_k)$.
- Väntevärde av linjärkombination $E(\sum_i a_i X_i + b) = \sum_i a_i E(X_i) + b$.
- Varians av linjärkombination $V(\sum_i a_i X_i + b) = \sum_i a_i^2 V(X_i) + 2\sum_{i<j} a_i a_j C(X_i, X_j)$.
- X, Y oberoende $\Rightarrow X$, Y okorrelerade, dvs $C(X, Y) = 0$.
- Betingat väntevärde för X, givet $Y = k$: $E(X \mid Y = k) = \sum_j j\, p_{X|Y=k}(j)$.
- Betingat väntevärde för X, givet $Y = y$: $E(X \mid Y = y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\, dx$.
- För betingade väntevärden gäller

$$E(X) = \begin{cases} \sum_k E(X \mid Y = k)\, p_Y(k), \\ \int_{-\infty}^{\infty} E(X \mid Y = y) f_Y(y)\, dy. \end{cases}$$

- Gauss approximationsformler:

$$E(g(X_1, \ldots, X_n)) \approx g(E(X_1), \ldots, E(X_n)).$$

$$V(g(X_1, \ldots, X_n)) \approx \sum_{i=1}^{n} c_i^2\, V(X_i) + 2\sum_{i<j} c_i\, c_j\, C(X_i,\ X_j),$$

där $c_i = \left. \dfrac{\partial g}{\partial x_i}(x_1, \ldots, x_n) \right|_{x_k = E(X_k),\ \forall k}$.

Normalfördelning och centrala gränsvärdessatsen

- Om $X_1, \ldots, X_n$ är oberoende och $N\left(\mu_1,\ \sigma_1\right), \ldots, N\left(\mu_n,\ \sigma_n\right)$ och $c_1, \ldots, c_n \in \mathbb{R}$, så gäller att

$$\sum_{i=1}^{n} c_i X_i \in N\left(\sum_{i=1}^{n} c_i \mu_i,\ \sqrt{\sum_{i=1}^{n} c_i^2 \sigma_i^2}\right).$$

- Centrala gränsvärdessatsen (CGS):
  Om $X_1, X_2, \ldots, X_n$ är oberoende och likafördelade med $E(X_i) = \mu$ och $D(X_i) = \sigma$, så gäller att

$$Y_n = X_1 + \cdots + X_n \in AsN(E(Y_n),\ D(Y_n)),$$

  då $n \to \infty$.

- Med utnyttjande av, bland annat, CGS gäller följande approximationer

Hypergeometrisk	$\to$	Binomial	om	$n/N \leq 0.1$.
Hypergeometrisk	$\to$	Poisson	om	$p + n/N \leq 0.1$ och $n \geq 10$.
Hypergeometrisk	$\to$	Normal	om	$\dfrac{N-n}{N-1} npq \geq 10$.
Binomial	$\to$	Poisson	om	$p \leq 0.1$ och $n \geq 10$.
Binomial	$\to$	Normal	om	$npq \geq 10$.
Poisson	$\to$	Normal	om	$\mu \geq 15$.

Stokastiska processer med diskret tid

- Övergångssannolikhet:

  $$p_{ij} = P(X_{n+1} = j \mid X_n = i)$$

  $\mathbf{P} = \{p_{ij}\}$ är övergångsmatrisen.

- Övergångssannolikhet av ordning m:

  $$p_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i)$$

  $\mathbf{P}^{(m)} = \{p_{ij}^{(m)}\}$ är övergångsmatrisen av ordning m.

- Chapman-Kolmogorovs sats:

  $$\mathbf{P}^{(m)} = \mathbf{P}^m$$

- Absoluta sannolikheter: $p_i(n) = P(X_n = i)$. $\mathbf{p}(n)$ är radvektorn $\{p_i(n)\}$. $\mathbf{p}(0)$ kallas initialfördelningen.

  $$\mathbf{p}(n) = \mathbf{p}(n-1)\mathbf{P} = \mathbf{p}(0)\mathbf{P}^n$$

- Stationär fördelning: $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$

- Beständighet: tillstånd i är beständigt om

  $$P(X_n = i \text{ för något } n > 0 \mid X_0 = i) = 1$$

  annars transient.

- Kommunikation: tillstånd i kommunicerar ensidigt med j om $p_{ij}^{(m)} > 0$ för något $m > 0$. Om i kommunicerar ensidigt med j och vice versa kommunicerar i och j tvåsidigt.

- Irreducibilitet: alla tillstånd kommunicerar tvåsidigt. Alla tillstånd är då antingen transienta, positivt beständiga eller nollbeständiga, och de har samma period.

- Existens av stationär fördelning: om $\{X_n\}$ är irreducibel så existerar en (unik) stationär fördelning om och endast om tillstånden är positivt beständiga.

Poissonprocessen

- En homogen Poissonprocess $\{X(t),\ t \geq 0\}$ har oberoende och stationära ökningar och

  $$X(t+s) - X(s) \in Po(\lambda t).$$

- Avstånden mellan konsekutiva händelser är oberoende och $Exp(\lambda)$-fördelade.

- En inhomogen Poissonprocess $\{X(t),\ t \geq 0\}$ har oberoende ökningar och

  $$X(t+s) - X(s) \in Po\left(\int_s^{s+t} \lambda(u)\, du\right).$$

Tabell 1: Vanliga fördelningar

Fördelning			Väntevärde	Varians
Binomialfördelning, $Bin(n,\ p)$	$p(k) = \binom{n}{k} p^k q^{n-k}$	$k = 0, 1, \ldots, n$	np	npq
Hypergeometrisk fördelning	$p(k) = \dfrac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}}$	$k \leq Np$, $n - k \leq Nq$	np	$\dfrac{N-n}{N-1}\, npq$
Poissonfördelning, $Po(\mu)$	$p(k) = e^{-\mu}\dfrac{\mu^k}{k!}$	$k = 0, 1, 2, \ldots$	μ	μ
Geometrisk fördelning	$p(k) = pq^k$	$k = 0, 1, 2, \ldots$	q/p	q/p^2
ffg-fördelning	$p(k) = pq^{k-1}$	$k = 1, 2, \ldots$	$1/p$	q/p^2
Normalfördelning, $N\left(\mu,\ \sigma\right)$	$f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$x \in \mathbb{R}$	μ	σ^2
Gammafördelning, $\Gamma(p, \lambda)$	$f(x) = \dfrac{\lambda^p}{\Gamma(p)}\, x^{p-1} e^{-\lambda x}$ $(*)$	$x \geq 0$	p/λ	p/λ^2
Exponential-fördelning, $Exp(\lambda)$	$f(x) = \lambda\, e^{-\lambda x}$	$x \geq 0$	$1/\lambda$	$1/\lambda^2$
Rektangel-fördelning, $R(a,\ b)$	$f(x) = \dfrac{1}{b-a}$	$a \leq x \leq b$	$\dfrac{a+b}{2}$	$\dfrac{(a-b)^2}{12}$
Dubbel exponential-fördelning	$F(x) = e^{-e^{-(x-b)/a}}$ (OBS! fördelningsfunktion)	$x \in \mathbb{R}$, $a > 0$	$b + \gamma a$ $(**)$	$\dfrac{a^2\pi^2}{6}$
Weibull-fördelning	$F(x) = 1 - e^{-\left(\frac{x-b}{a}\right)^c}$ (OBS! fördelningsfunktion)	$x \geq b$, $a, c > 0$	$b + a\Gamma(1 + 1/c)$	$a^2\left[\Gamma(1+\frac{2}{c}) - \Gamma^2(1+\frac{1}{c})\right]$
Lognormal-fördelning $\ln X \in N\left(b,\ a\right)$	$f(x) = \dfrac{1}{\sqrt{x^2 a^2 2\pi}}\, e^{-\frac{(\ln x - b)^2}{2a^2}}$	$x \geq 0$	$e^{b+a^2/2}$	$e^{2b+2a^2} - e^{2b+a^2}$

$(*)$ $\quad \Gamma(p) = \int_0^\infty x^{p-1} e^{-x}\, dx, \quad p > 0 \qquad \Gamma(p) = (p-1)\cdot\Gamma(p-1)$

$\Gamma(p) = (p-1)!$ om p heltal $\qquad \Gamma(\frac{1}{2}) = \sqrt{\pi}$

$(**)$ $\quad \gamma \approx 0.57722.$

Tvådimensionell normalfördelning

(X, Y) är tvådimensionellt normalfördelad med väntevärdesvektor $\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ och kovariansmatris $\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$ om $\quad f_{X,Y}(x,y) = \dfrac{1}{2\pi\sqrt{\det(\mathbf{\Sigma})}}\, e^{-Q/2}, \quad (x,y) \in \mathbb{R}^2,$

där $\det(\mathbf{\Sigma}) = \sigma_X^2\sigma_Y^2(1-\rho^2)$ och $Q = \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}^{\mathsf{T}} \mathbf{\Sigma}^{-1} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}.$

Den betingade fördelningen för X givet att $Y = y$ är en endimensionell normalfördelning med

$$\mathsf{E}(X \mid Y = y) = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y),$$
$$\mathsf{V}(X \mid Y = y) = \sigma_X^2(1-\rho^2).$$

Fördelningar besläktade med normalfördelningar

χ^2-fördelning	$X_1, \ldots, X_n \in N(0,\ 1)$,	oberoende	$\Rightarrow \sum_{i=1}^{n} X_i^2 \in \chi^2(n).$
$\chi^2(f) = \Gamma(f/2, 1/2)$			
t-fördelning, $t(f)$	$X \in N(0,\ 1)$, $Y \in \chi^2(f)$,	oberoende	$\Rightarrow \dfrac{X}{\sqrt{Y/f}} \in t(f).$
F-fördelning, $F(f_1, f_2)$	$X \in \chi^2(f_1)$, $Y \in \chi^2(f_2)$,	oberoende	$\Rightarrow \dfrac{X/f_1}{Y/f_2} \in F(f_1, f_2).$

Additionsformler

Om X_1 och X_2 oberoende så gäller:

$$X_1 \in Bin(n_1, p),\ X_2 \in Bin(n_2, p) \quad \Rightarrow \quad X_1 + X_2 \in Bin(n_1 + n_2,\ p).$$

$$X_1 \in Po(\mu_1),\ X_2 \in Po(\mu_2) \quad \Rightarrow \quad X_1 + X_2 \in Po(\mu_1 + \mu_2).$$

$$X_1 \in \Gamma(p_1, a),\ X_2 \in \Gamma(p_2, a) \quad \Rightarrow \quad X_1 + X_2 \in \Gamma(p_1 + p_2, a).$$

$$X_1 \in \chi^2(f_1),\ X_2 \in \chi^2(f_2) \quad \Rightarrow \quad X_1 + X_2 \in \chi^2(f_1 + f_2).$$

Statistikteori

Punktskattningar vid normalfördelning och helt okänd fördelning

Ett stickprov

Låt $x_1, \ldots, x_n$ vara observationer av oberoende och likafördelade s.v. $X_1, \ldots, X_n$ med väntevärde μ och standardavvikelse σ. Väntevärdesriktiga skattningar av μ och σ^2 är då

- $\mu^* = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ Om $X_i \in N(\mu, \sigma)$ så $\mu^* \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- $(\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ då μ känd. Om $X_i \in N(\mu, \sigma)$ så $\frac{n(\sigma^2)^*}{\sigma^2} \in \chi^2(n)$
- $(\sigma^2)^* = S^2 = \frac{Q}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ då μ okänd. Om $X_i \in N(\mu, \sigma)$ så $\frac{Q}{\sigma^2} \in \chi^2(n-1)$

Flera stickprov

Låt $x_{i1}, \ldots, x_{in_i}$ vara ober. obs. från $N(\mu_i, \sigma)$ då $i = 1, \ldots, k$. Då är

- $(\sigma^2)^* = S^2 = \frac{Q}{f} = \frac{(n_1 - 1)\, S_1^2 + \cdots + (n_k - 1)\, S_k^2}{(n_1 - 1) + \cdots + (n_k - 1)}$ Eftersom $X_{ij} \in N(\mu_i, \sigma)$ så $\frac{Q}{\sigma^2} \in \chi^2(f)$

Vanliga skattningsmetoder

- ML-skattning:
  Låt $x_1, \ldots, x_n$ vara observationer av $X_1, \ldots, X_n$, som är oberoende s.v. med täthets- (sannolikhets-) funktion $f(x_i; \theta)$, $i = 1, \ldots, n$ ($p(x_i; \theta)$, $i = 1, \ldots, n$).
  ML-skattningen av parametern θ är det θ^*_{ML} som maximerar likelihood-funktionen

$$L(\theta;\, x_1, \ldots, x_n) = \begin{cases} p(x_1; \theta) \cdot p(x_2; \theta) \cdot \cdots \cdot p(x_n; \theta), \\ \\ f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta). \end{cases}$$

- MK-skattning:
  Låt $x_1, \ldots, x_n$ vara oberoende observationer av stokastiska variabler med $\mathsf{E}(X_i) = \mu_i(\theta)$, där funktionerna μ_i är kända och parametern θ okänd.
  MK-skattningen av parametern θ är det θ^*_{MK} som minimerar förlustfunktionen

$$Q(\theta;\, x_1, \ldots, x_n) = \sum_{i=1}^{n}(x_i - \mu_i(\theta))^2.$$

- Viktad MK-skattning:
  (Förutsättningar enligt MK-skattning ovan.) Den viktade MK-skattningen av θ är det θ^*_{MK} som minimerar förlustfunktionen

$$Q(\theta;\, x_1, \ldots, x_n) = \sum_{i=1}^{n}\lambda_i(x_i - \mu_i(\theta))^2,$$

  där λ_i är en följd av vikter, t.ex. $\lambda_i = 1/\sigma_i^2$, där $\sigma_i^2 = \mathsf{V}(X_i)$.

Konfidensintervall

- Konfidensintervall med konfidensgrad $1-\alpha$ för väntevärdet av en normalfördelad skattning:
  Om $\theta^*(X_1,\ldots,X_n) \in N(\theta,\ \mathsf{D}(\theta^*))$ så
  - $I_\theta = (\theta^* \pm \lambda_{\alpha/2} \cdot \mathsf{D}(\theta^*))$ om $\mathsf{D}(\theta^*)$ är känd
  - $I_\theta = (\theta^* \pm t_{\alpha/2}(f) \cdot \mathsf{d}(\theta^*))$ om $\mathsf{D}(\theta^*) = c \cdot \sigma$ där $\sigma = \mathsf{D}(X_i)$, c är en konstant och $(\sigma^2)^* = S^2 = \dfrac{Q}{f}$ med $\dfrac{Q}{\sigma^2} \in \chi^2(f)$

- Konfidensintervall med konfidensgrad $1-\alpha$ för väntevärdet i en normalapproximation:
  Om $\theta^*(X_1,\ldots,X_n) \underset{\sim}{\in} N(\theta,\ \mathsf{D}(\theta^*))$ enligt CGS (el. dyl.) så
  - $I_\theta \approx (\theta^* \pm \lambda_{\alpha/2} \cdot \mathsf{D}(\theta^*))$ om $\mathsf{D}(\theta^*)$ är känd
  - $I_\theta \approx (\theta^* \pm \lambda_{\alpha/2} \cdot \mathsf{d}(\theta^*))$ om $\mathsf{D}(\theta^*)$ skattas med $\mathsf{d}(\theta^*)$
  - $I_\theta \approx (\theta^* \pm t_{\alpha/2}(f) \cdot \mathsf{d}(\theta^*))$ om $\mathsf{D}(\theta^*)$ skattas med $\mathsf{d}(\theta^*)$ där $\mathsf{D}(\theta^*) = c \cdot \sigma$ med $\sigma = \mathsf{D}(X_i)$, c är en konstant och $(\sigma^2)^* = S^2 = \dfrac{Q}{f}$

- Konfidensintervall med konfidensgrad $1-\alpha$ för variansen i en normalfördelning:
  Om $X_1,\ldots,X_n \in N\left(\mu,\ \sigma\right)$ med $(\sigma^2)^* = S^2 = \dfrac{Q}{f}$ och $\dfrac{Q}{\sigma^2} \in \chi^2(f)$ så

$$I_{\sigma^2} = \left(\frac{f \cdot s^2}{\chi^2_{\alpha/2}(f)},\ \frac{f \cdot s^2}{\chi^2_{1-\alpha/2}(f)}\right)$$

Hypotestest

Styrkefunktion: $h(\theta) = \mathsf{P}(H_0 \text{ förkastas} \parallel \theta \text{ är det rätta parametervärdet})$
Speciellt: Signifikansnivån, $\alpha = \mathsf{P}(H_0 \text{ förkastas} \parallel H_0 \text{ sann})$

Regression

Enkel linjär regression

- Modell: $y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \in N(0,\ \sigma)$.
- Parameterskattningar:

$$\alpha^* = \bar{y} - \beta^* \bar{x}, \qquad \beta^* = \frac{S_{xy}}{S_{xx}}, \qquad (\sigma^2)^* = s^2 = \frac{Q_0}{n-2}, \qquad Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

$$\alpha^* \in N\left(\alpha,\ \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right), \qquad \beta^* \in N\left(\beta,\ \frac{\sigma}{\sqrt{S_{xx}}}\right), \qquad \frac{Q_0}{\sigma^2} \in \chi^2(n-2),$$

$$\mathsf{C}(\alpha^*,\ \beta^*) = -\bar{x} \cdot \frac{\sigma^2}{S_{xx}}, \qquad \mu_0^* = \alpha^* + \beta^* x_0 \in N\left(\alpha + \beta x_0,\ \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right).$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- Ett prediktionsintervall med konfidensgrad $1-p$ för $y(x_0) = \alpha + \beta x_0 + \varepsilon_0$ ges av

$$I_{y(x_0)} = \left(\alpha^* + \beta^* x_0 \pm t_{p/2}(n-2) \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

- Ett kalibreringsintervall med konfidensgrad $1-p$ för $x_0 = \dfrac{y_0 - \alpha}{\beta}$ ges av

$$I_{x_0} = \left(x_0^* \pm t_{p/2}(n-2) \cdot \frac{s}{\beta^*} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0^* - \bar{x})^2}{S_{xx}}} \right) \qquad \text{där} \qquad x_0^* = \frac{y_0 - \alpha^*}{\beta^*}$$

Multipel linjär regression

- Modell: $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i, \quad \varepsilon_i \in N(0, \sigma)$.
- Med matrisrepresentation kan modellen skrivas $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$ med

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_1^{(p)} \\ 1 & x_2^{(1)} & \ldots & x_2^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & \ldots & x_n^{(p)} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Parameterskattningar:

  $\boldsymbol{\beta}^* = (X^T X)^{-1} X^T \mathbf{y}, \qquad \mathsf{V}(\boldsymbol{\beta}^*) = \sigma^2 (X^T X)^{-1},$

  $(\sigma^2)^* = s^2 = \dfrac{Q_0}{n-(p+1)}, \qquad Q_0 = \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^{*T} X^T \mathbf{y}, \qquad \dfrac{Q_0}{\sigma^2} \in \chi^2(n-(p+1)),$

  $\beta_i^* \in N\left(\beta_i,\ \sigma\sqrt{\text{element}(ii) \text{ i } (X^T X)^{-1}}\right),$

  $\mu_0^* = \mathbf{x}_0 \boldsymbol{\beta}^* \in N\left(\mu_0,\ \sigma\sqrt{\mathbf{x}_0 (X^T X)^{-1} \mathbf{x}_0^T}\right) \quad \text{där} \quad \mathbf{x}_0 = \begin{pmatrix} 1 & x_0^{(1)} & \ldots & x_0^{(p)} \end{pmatrix}$