Misc

- $n\in\mathbb{N}^*$，$\forall k\in\mathbb{Z},k\geq 4\log_2 n$，有 $k!\geq n^3$
- $\forall x\geq -1$，$\ln(1+x)\leq x$；$\forall x\leq 1/2$，$\ln(1-x)\geq -x-x^2$
- $\exp(x)\geq x+1$
- $\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)}f(x)\mathrm{d}x=f(b(t))b'(t)-f(a(t))a'(t)$
- Gram-Schmidt 正交化！
- 非负离散：$E(X)=\sum_{x=0}^{+\infty}P(X>x)$，连续 $E(X)=\int_0^{+\infty}P(X>x)\mathrm{d}x$
- $\int_{-\infty}^{+\infty}e^{-t^2}\mathrm{d}t=\sqrt{\pi}$（高斯积分）
- 分部积分：$\int u\mathrm{d}v=uv-\int v\mathrm{d}u$
- 凑微分：$\int \mathrm{d}f(u)=\int f'(u)\mathrm{d}u$
- $\ln(1-x)=-\sum_{k\geq 1}\frac{x^k}{k}$，$(1+x)^\alpha=\sum_{k\geq 0}\binom{\alpha}{k}x^k$
- $\int(ax+b)^n\mathrm{d}x=\frac{(ax+b)^{n+1}}{a(n+1)}+C$，也即 $\int x^n\mathrm{d}x=\frac{x^{n+1}}{n+1}+C$

1 概率论基本概念

概率公理化：S 为样本空间，F 为 S 的某些子集组成的事件域。如果定义在 F 上的实值函数 P 满足。1. $\forall A\in F,P(A)\geq 0$；2. $P(S)=1$；3. $\forall A_1,A_2,\ldots\in F$ 且两两互斥，有 $P(\bigcup_{i=1}^\infty A_i)=\sum_{i=1}^\infty P(A_i)$，则称 P 为概率测度，(S,F,P) 为概率空间。

一般加法公式 $P(\bigcup_{i=1}^n A_i)=\sum_{i=1}^n P(A_i)-\sum_{1\leq i<j\leq n}P(A_iA_j)+\cdots+(-1)^{n-1}P(A_1\cdots A_n)$

贝叶斯公式 $P(A|B)=P(B|A)\cdot\dfrac{P(A)}{P(B)}$

相互独立比两两独立**更强**。

Union Bound $P(\bigcup A_i)\leq\sum P(A_i)$，对 A_i 无要求。

证明存在可以通过证明其概率大于 0

2 离散随机变量

$P(X\geq E(X))>0,P(X\leq E(X))>0,\mathrm{Var}(aX+b)=a^2\mathrm{Var}(X)$

尾不等式/集中不等式：随机变量与期望的偏离。

Markov：对**非负** X，$E(X)>0,a>0$，有

$$P(X\geq a)\leq\frac{E(X)}{a}\quad P(X\geq aE(X))\leq\frac{1}{a}$$

Chebyshev：对 $\sigma(X)>0,c>0$，

$$P(|X-E(X)|\geq c\cdot\sigma(X))\leq 1/c^2$$
$$P(|X-E(X)|\geq a)\leq \mathrm{Var}(X)/a^2$$

分布	分布列 $P(X=k)=$	期望	方差
$B(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
$G(p)$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\pi(\lambda)$	$\lambda^k e^{-\lambda}/k!$	λ	λ
$NB(r,p)$	$\binom{k-1}{r-1}p^r(1-p)^{k-r}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

二项分布：n 次独立重复伯努利实验成功次数。**几何分布**：第一次成功的试验次数。**负二项分布**：第 r 次成功的次数。有 $G(p)=NB(1,p)$。**泊松分布**：单位时间内事件发生次数。几何分布具有**无记忆性**：$P(X>m+n|X>m)=P(X>n)$。$NB(r,p)$ 可以拆成 r 个独立的 $G(p)$ 之和。

3 连续随机变量

正态分布：$X\sim N(\mu,\sigma^2)$，$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$，$E(X)=\mu,\mathrm{Var}(X)=\sigma^2$。标准正态分布 $Z\sim N(0,1)$。

指数分布：$X\sim\mathrm{Exp}(\lambda)$，$f(x)=\lambda e^{-\lambda x}$，$F(x)=1-e^{-\lambda x}$，$E(X)=1/\lambda,\mathrm{Var}(X)=1/\lambda^2$。具有无记忆性 $P(X>s+t|X>s)=P(X>t)$。理解为泊松分布假设下，第一次事件发生的时刻。

伽马分布：$X\sim\Gamma(\alpha,\lambda)$，理解为泊松假设下第 α 次的时刻。

$$f(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},x\geq 0$$

$E(X)=\alpha/\lambda$，$\mathrm{Var}(X)=\alpha/\lambda^2$。其中 $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x,\Gamma(\alpha+1)=\alpha\cdot\Gamma(\alpha),\Gamma(n+1)=n!,\Gamma(n+\frac{1}{2})=\frac{(2n)!}{4^n n!}\sqrt{\pi}$。具有可加性：$\Gamma(\alpha_1+\alpha_2,\lambda)=\Gamma(\alpha_1,\lambda)+\Gamma(\alpha_2,\lambda)$。$\alpha=1$ 时即指数分布。$\alpha=n/2,\lambda=1/2$ 时即 $\chi^2(n)$ 分布。

概率密度变换：$Y=g(X)$，g 单调且反函数 $h(y)$ 有连续导数，则 $f_Y(y)=f_X(h(y))\cdot|h'(y)|$。如果没法直接套这个公式的话可以从**分布函数**的定义出发进行变换。即先算 $P(Y\leq y)$ 再求导。

4 多维离散随机变量

对于**独立**的 X,Y，有 $\mathrm{Var}(X\pm Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$

协方差：$\mathrm{Cov}(X,Y)=E((X-E(X))(Y-E(Y)))=E(XY)-E(X)E(Y)$。
$\mathrm{Var}(X_1+\cdots+X_n)=\sum_i\sum_j\mathrm{Cov}(X_i,X_j)$，
$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$，
$\mathrm{Cov}(aX,bY)=ab\cdot\mathrm{Cov}(X,Y)$，
$\mathrm{Cov}(X_1+X_2,Y)=\mathrm{Cov}(X_1,Y)+\mathrm{Cov}(X_2,Y)$，
$\mathrm{Cov}(X,Y)\leq\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$

条件期望：$E(X|Y=y)=\sum_i x_iP(X=x_i|Y=y)$，是关于 y 的函数。$E(X|Y)$ 是随机变量。

重期望公式：$E(E(X|Y))=E(X)$

5 多维连续随机变量

条件分布函数：$F(x|y)=P(X\leq x|Y=y)=\int_{-\infty}^x\frac{f(u,y)}{f_Y(y)}\mathrm{d}u$

条件密度函数：$f(x|y)=\frac{\partial}{\partial x}F(x|y)=\frac{f(x,y)}{f_Y(y)}$。

二维高斯：$X,Y\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，要求 $|\rho|<1$。

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)\right]$$

边际密度函数与 ρ 无关，即 $X\sim N(\mu_1,\sigma_1^2),Y\sim N(\mu_2,\sigma_2^2)$。$\rho=0$ 时独立。协方差 $\mathrm{Cov}(X,Y)=\rho\sigma_1\sigma_2$。

相关系数 $\mathrm{Corr}(X,Y)=\mathrm{Cov}(X,Y)/\sigma(X)\sigma(Y)$。相关系数/协方差大于 0 则正相关，小于 0 则负相关，等于 0 则不相关（但不一定独立）。相关系数等于 ± 1 代表 X,Y 呈严格线性关系。证明考虑标准化 $\tilde{X},\tilde{Y}$，然后通过 $\mathrm{Var}(\tilde{X}-\tilde{Y})=0$ 推导出 $P(\tilde{X}-\tilde{Y}=c)=1$。

5.1 概率密度变换

卷积公式：若 X,Y 独立，$Z=X+Y$，则

$$f_Z(z)=\int_{-\infty}^\infty f_X(x)f_Y(z-x)\mathrm{d}x=\int_{-\infty}^\infty f_X(z-y)f_Y(y)\mathrm{d}y$$

min/max：若 $X_1,\cdots,X_n$ 独立，则 $Y=\max\{X_i\}$ 的分布函数为 $F_Y(y)=\prod_{i=1}^n F_{X_i}(y)$，$Y=\min\{X_i\}$ 的分布函数为 $F_Y(y)=1-\prod_{i=1}^n(1-F_{X_i}(y))$。

换元：X,Y 的概率密度为 $f(x,y)$，函数 $u=u(x,y)$ 和 $v=v(x,y)$ 偏导连续且 $x=x(u,v),y=y(u,v)$ 为唯一反函数，则 $U=u(X,Y),V=v(X,Y)$ 的联合概率密度为

$$g(u,v)=f(x(u,v),y(u,v))\cdot|J|,J=\left|\frac{\partial(x,y)}{\partial(u,v)}\right|=\begin{vmatrix}\frac{\partial x}{\partial u}&\frac{\partial x}{\partial v}\\\frac{\partial y}{\partial u}&\frac{\partial y}{\partial v}\end{vmatrix}$$

5.2 线性代数

对角化：$\boldsymbol{A}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$。保行列式，平方的行列式和 trace 不变。实对称矩阵可对角化，不同特征值的特征向量正交。

正定：半正定矩阵 $\boldsymbol{A}$ 存在 $\boldsymbol{B}=\boldsymbol{A}^{1/2}$，$\boldsymbol{B}^\top\boldsymbol{B}=\boldsymbol{B}^2=\boldsymbol{A}$，且 B 不唯一。对于正定矩阵，这样的 $\boldsymbol{B}$ 可逆，$(\boldsymbol{B}^{-1})^2=\boldsymbol{A}^{-1}$。

协方差矩阵：对于随机变量 $\boldsymbol{X}=(X_1,\cdots,X_n)$，$\mathrm{Cov}(\boldsymbol{X})=E((\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))^\top)$ 为协方差矩阵，有

$$\mathrm{Cov}(\boldsymbol{X})=\begin{pmatrix}\mathrm{Var}(X_1)&\mathrm{Cov}(X_1,X_2)&\cdots&\mathrm{Cov}(X_1,X_n)\\\mathrm{Cov}(X_2,X_1)&\mathrm{Var}(X_2)&\cdots&\mathrm{Cov}(X_2,X_n)\\\vdots&\vdots&&\vdots\\\mathrm{Cov}(X_n,X_1)&\mathrm{Cov}(X_n,X_2)&\cdots&\mathrm{Var}(X_n)\end{pmatrix}$$

其对称且半正定。

高斯：n 维高斯的联合密度函数，$\boldsymbol{B}$ 为协方差矩阵：

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\boldsymbol{B})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\cdot(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

若 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{B})$，令 $\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}$，且 $\boldsymbol{A}$ 行满秩，则 $Y\sim N(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top)$

5.3 结论

- 对于独立的 $X_i\sim N(\mu_i,\sigma_i^2)$，有
$$\sum a_iX_i\sim N(\sum a_i\mu_i,\sum a_i^2\sigma_i^2)$$
- 对于单个 $X_i\sim N(0,1)$，有 $X_i^2\sim\chi^2(1)=\Gamma(1/2,1/2)$，且 $\sum_{i=1}^n X_i^2\sim\chi^2(n)=\Gamma(\frac{n}{2},\frac{1}{2})$

6 尾不等式、大数定律与中心极限定理

6.1 尾不等式

尾不等式：$P(X\geq k)$ 的上界。集中不等式：$P(|X-E(X)|\geq k)$ 的上界。

矩：$E(X^n)$ 称为 X 的 n 阶矩，$E((X-E(X))^n)$ 称为 X 的 n 阶中心矩。切比雪夫不等式的本质是对二阶中心矩使用 Markov。

矩生成函数：$M_X(t)=E(e^{tX})=\sum_{i\geq 0}\frac{t^i}{i!}E(X^i)$。所以求 k 阶矩可以求其封闭形式的 k 阶导然后令 $t=0$。

Chernoff Bound：求 k 阶中心矩然后用 Markov 得到的尾不等式通常较弱（没有真正用到 n 重伯努利实验的独立性）。对于任意 $t>0$，有

$$P(X\geq a)=P(e^{tX}\geq e^{ta})\leq\frac{E(e^{tX})}{e^{ta}}=e^{-ta}M_X(t)$$

对于任意 $t<0$ 有 $P(X\leq a)\leq M_X(t)\cdot e^{-ta}$ 通过调节 t 可得到更紧的上界。一般而言是**求导**找最小值。但是 Chernoff Bound 不一定是最紧的。

Hoeffding 引理：若实数随机变量 $a\leq X\leq b$，则对任意实数 t 有

$$E(e^{t(X-E(X))})\leq\exp\left(\frac{t^2(b-a)^2}{8}\right)$$

Chernoff-Hoeffding 不等式：$X_1,\cdots,X_n$ 独立，且 $a_i\leq X_i\leq b_i$，令 $X=\sum_{i=1}^n X_i$，则对任意 $t>0$ 有

$$P(X\geq E(X)+t)\leq\exp\left(\frac{-2t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right)$$
$$P(X\leq E(X)-t)\leq\exp\left(\frac{-2t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right)$$

对于 $a\leq X_i\leq b$ 的情况，分母就是 $n(b-a)^2$。

6.2 尾不等式结论

矩生成函数的应用：对于要算 e^X 的期望，可以先算矩生成函数然后令 $t=1$。

矩生成函数：

- $X\sim B(n,p)$，$M_X(t)=(1-p+e^tp)^n$
- $X\sim\pi(\lambda)$，$M_X(t)=\exp(\lambda(e^t-1))$
- $X\sim N(\mu,\sigma^2)$，$M_X(t)=\exp(\mu t+\frac{\sigma^2t^2}{2})$

常见 Chernoff-Hoeffding 界：

- $X\sim\pi(\lambda)$，$P(X\geq x)\leq\frac{e^{-\lambda}(e\lambda)^x}{x^x}$
- $X\sim N(\mu,\sigma^2)$，$P(X\geq E(X)+k\sigma)\leq\exp(-k^2/2)$
- $X\sim B(n,p)$，$P(|X-E(X)|\geq n\epsilon)\leq 2\cdot\exp(-2n\epsilon^2)$

对期望分段放缩：对于 $E(Y)=\sum_{k=1}^n P(Y=k)\cdot k$，若我们知道对于 $k>k'$ 有 $P(Y=k)\leq c$，那么就可以分段放缩：

$$E(Y)=\sum_{k=1}^{k'}P(Y=k)\cdot k+\sum_{k=k'+1}^{n}P(Y=k)\cdot k$$

左边的 k 放成 k'，右边的放成 n，然后和 c 消掉。

6.3 大数定律

大数定律的一般形式：对于随机变量 $\{X_n\}$，对于任意 $\epsilon>0$，若

$$\lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\epsilon\right)=1$$

Markov 大数定律：$\frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^n X_i)\to 0$，则 $\{X_n\}$ 满足大数定律的一般形式。

辛钦大数定律：$\{X_n\}$ 独立同分布，且 $E(X_i)=\mu$，则 $\{X_n\}$ 满足大数定律的一般形式。对比 Markov，需要 iid，但不需要方差。

依概率收敛：随机变量序列 $\{X_n\}$ 依概率收敛于 X，记作 $X_n\xrightarrow{P}X$，如果对任意 $\epsilon>0$，有

$$\lim_{n\to\infty}P(|X_n-X|\geq\epsilon)=0$$

依分布收敛：随机变量序列 $\{X_n\}$ 依分布收敛于 X，记作 $X_n\xrightarrow{d}X$，如果对任意 x，$F_{X_n}(x)\to F_X(x)$。**依概率收敛可以推出依分布收敛，反之不亦然**

几乎必然收敛：随机变量序列 $\{X_n\}$ 几乎必然收敛于 X，记作 $X_n\xrightarrow{a.s.}X$，如果 $\forall\epsilon>0$

$$\lim_{n\to\infty}P\left(\bigcup_{m=n}^{\infty}|X_m-X|\geq\epsilon\right)=1$$

6.4 应用

6.4.1 随机快排

算法：随机一个 pivot x，将其余元素排在两侧 L,R，然后递归 L,R。如何计算 $E(T)$？$T=O(\sum_{i<j}1_{C_{i,j}})$，$C_{i,j}$ 表示 i,j 是否比较过。

发现算法比较过 i,j iff i 或 j 是 $[i,j]$ 中第一个被选为 pivot 的元素。因为每个元素被选为 pivot 的概率相等，$P(C_{i,j})=\frac{2}{j-i+1}$。所以 $E(T)=\sum_{i<j}\frac{2}{j-i+1}=O(n\log n)$。

令 D_i 表示 i 被比较的次数，给出尾不等式。发现若 pivot 落在 $[n/4,3n/4]$ 则 i 所在数组大小至少减小 1/4，前者概率为 1/2。若至少有 $3\log n$ 次，则完成排序。即 $-(3/4)^{3\log n}\leq 1/n$。所以 $P(D_i>20\log n)\leq P(X_i<3\log n)$，$X_i\sim B(20\log n,1/2)$。$P(X_i<3\log n)\leq\exp(-4\log n)\leq 1/n^4$。Union Bound 一下 $P(T>20n\log n)\leq 1/n^3$。

6.4.2 JL 降维

结论：给定 $x_i\in\mathbb{R}^n$，存在线性映射 $F:\mathbb{R}^n\to\mathbb{R}^k$，其中 $m=O(\epsilon^{-2}\log N)$，$\geq 1/2$ 概率 $\forall i,j$，$(1-\epsilon)\|x_i-x_j\|_2^2\leq\|F(x_i)-F(x_j)\|_2^2\leq(1+\epsilon)\|x_i-x_j\|_2^2$。$F=\frac{1}{\sqrt{k}}\cdot Ax$，其中 A 的每个元素独立服从 $N(0,1)$。

思路：对所有可能的 $x=x_i-x_j$ 使用引理然后 Union Bound。

7 参数估计

估计量：样本的函数，用于估计未知参数。

偏差：$\mathrm{Bias}(\hat\theta)=E(\hat\theta)-\theta$，$\mathrm{Bias}(\hat\theta)=0$ 称为无偏估计量。若 $\lim_{n\to\infty}E(\hat\theta)=\theta$，称为**渐近无偏**估计量。

均方误差：$\mathrm{MSE}(\hat\theta)=E((\hat\theta-\theta)^2)=\mathrm{Var}(\hat\theta)+(\mathrm{Bias}(\hat\theta))^2$。若无偏则 $\mathrm{MSE}(\hat\theta)=\mathrm{Var}(\hat\theta)$。

一致估计量：若估计量 $\hat\theta_n\xrightarrow{P}\theta$，则称 $\hat\theta_n$ 为参数 θ 的一致估计量。等价于 $\mathrm{MSE}\to 0$。

k 阶矩：$A_k=\frac{1}{n}\sum_{i=1}^n X_i^k$，$A_k$ 是 $\mu_k=E(X^k)$ 的无偏估计量，且一致。k 阶中心矩：$B_k=\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})^k$。$B_2$ 是 σ^2 的渐近无偏估计量，且一致，但不是无偏估计量。因为 $E(B_2)=E(X^2)-E(\overline{X}^2)$，而 $E(\overline{X}^2)=(E(X))^2+\mathrm{Var}(\overline{X})$（平方的期望减期望的平方），然后 $\mathrm{Var}(\overline{X})=\mathrm{Var}(X)/n$，所以 $E(B_2)=\frac{n-1}{n}\mathrm{Var}(X)$。**样本方差** $S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X})^2$ 是 $\mathrm{Var}(X)$ 的无偏估计量。

矩法：用样本矩替换总体矩。**MLE**：最大化似然函数 $L(\theta)=P(X_1=x_1,\cdots,X_n=x_n)$。

区间估计：设计统计量 $\hat\theta_L(X_1,\cdots,X_n)$ 和 $\hat\theta_U(X_1,\cdots,X_n)$，使得 $P(\hat\theta_L\leq\theta\leq\hat\theta_U)\geq 1-\alpha$。

方法：**枢轴量法**。设计枢轴量 G 使得 G 的分布与未知参数无关，然后选择 c,d 使得 $P(c\leq G\leq d)=1-\alpha$，从而得到不等式 $c\leq G(X_1,\cdots,X_n,\theta)\leq d$，解出 θ 的区间估计。

7.1 技术

说明正态总体下 $\overline{X}$ 和 S^2 独立。核心思路：**通过线性变换将缠在一起的变量 $X_1,\cdots,X_n$（各自包含了均值和方差信息）解耦开来**。构造正交矩阵 U，第一行全为 $1/\sqrt{n}$，其他随意。令随机变量 $\boldsymbol{X}=(X_1,\cdots,X_n)$，令 $\boldsymbol{Y}=U\boldsymbol{X}$，显然 $\boldsymbol{Y}$ 服从高斯分布。注意到 $E(\boldsymbol{Y})=(\sqrt{n}\mu,0,\cdots,0)$，且 $\mathrm{Cov}(\boldsymbol{Y})=\sigma^2I$。因此 $Y_1,\cdots,Y_n$ 独立，且 $Y_1\sim N(\sqrt{n}\mu,\sigma^2)$，$Y_2,\cdots,Y_n\sim N(0,\sigma^2)$。注意到 $\overline{X}=\frac{Y_1}{\sqrt{n}}$，且 $\sum(X_i-\overline{X})^2=\sum_{i=2}^n Y_i^2$，因此 $\overline{X}$ 和 S^2 独立。

知道 $\overline{X}\sim N(\mu,\sigma^2/n)$，$(n-1)S^2/\sigma^2\sim\chi^2(n-1)$（因为是 $n-1$ 个独立的 $N(0,1)$ 的平方和）。

8 回归分析

8.1 一元线性回归

回归分析：$y=\alpha+\beta x+\epsilon$，其中 ϵ 为误差项，$E(\epsilon)=0$, $\mathrm{Var}(\epsilon)=\sigma^2$。

最小二乘：$Q(\alpha,\beta)=\sum_{i=1}^n(y_i-\beta x_i-\alpha)^2$，使得 Q 最小的 $\hat\alpha,\hat\beta$ 称为最小二乘估计。求偏导然后令为 0 可得

$$\hat\beta=\frac{S_{xy}}{S_{xx}},\hat\alpha=\overline{y}-\hat\beta\overline{x}$$

其中 $S_{xy}=\sum_{i=1}^n(x_i-\overline{x})(y_i-\overline{y})$，$S_{xx}=\sum_{i=1}^n(x_i-\overline{x})^2$。

一个很关键的技巧：$\sum(x_i-\overline{x})=0$，因此可以对比如 $\sum(x_i-\overline{x})(x_i)$ 的式子进行处理成 s_{xx} 的形式。

无偏性：$\hat\beta=\beta+\sum\epsilon_i(x_i-\overline{x})/s_{xx}$，$\hat\alpha=\alpha+\sum\epsilon_i\left(\frac{1}{n}-\frac{(x_i-\overline{x})}{s_{xx}}\cdot\overline{x}\right)$

估计量的方差与协方差：$\mathrm{MSE}(\hat\beta)=\mathrm{Var}(\hat\beta)=\sigma^2/s_{xx}$，$\mathrm{MSE}(\hat\alpha)=\mathrm{Var}(\hat\alpha)=\sigma^2\left(\frac{1}{n}+\frac{(\overline{x})^2}{s_{xx}}\right)$，算协方差的时候同样考虑只有交叉项有贡献，$\mathrm{Cov}(\hat\alpha,\hat\beta)=-\sigma^2\cdot\frac{\overline{x}}{s_{xx}}$。

预测值的无偏性与方差：预测值 $\hat{y}_i=\hat\alpha+\hat\beta x_i$，$E(\hat{y}_i)=\alpha+\beta x_i$，所以无偏。$\mathrm{Var}(\hat{y}_i)=\sigma^2\left[\frac{1}{n}+\frac{(x_i-\overline{x})^2}{s_{xx}}\right]$，通过 $\mathrm{Var}(\hat{y}_i)=\mathrm{Var}(\hat\alpha)+x_i^2\mathrm{Var}(\hat\beta)+2x_i\mathrm{Cov}(\hat\alpha,\hat\beta)$ 计算。

残差的方差：残差 $e_i=y_i-\hat{y}_i$，$E(e_i)=0$，展开方差的公式来计算 $\mathrm{Var}(e_i)=\sigma^2\left[1-\frac{1}{n}-\frac{(x_i-\overline{x})^2}{s_{xx}}\right]$。

σ^2 的无偏估计：$E(\sum(\hat{y}_i-y_i)^2)=\sum\mathrm{Var}(\hat{y}_i-y_i)=(n-2)\sigma^2$，所以 $s^2=\frac{1}{n-2}\sum(\hat{y}_i-y_i)^2$ 为无偏估计量。

最大似然：需要 $\epsilon_i\sim N(0,\sigma^2)$ 且相互独立，对于 α,β 等价于最小二乘，但是 $\hat{\sigma^2}_{\mathrm{MLE}}=\frac{1}{n}\sum(y_i-\hat{y}_i)^2$，是有偏的。$\hat\alpha$ 和 $\hat\beta$ 服从正态分布（方差我们之前计算过）。若 $s^2=\frac{1}{n-2}\sum(\hat{y}_i-y_i)^2$，则 $\frac{(n-2)s^2}{\sigma^2}\sim\chi^2(n-2)$，且与 $\hat\alpha,\hat\beta$ 独立。

8.2 多元线性回归

$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\epsilon$ 写成向量的形式，发现 $y_i=\boldsymbol{x_i}^\top\boldsymbol{\beta}+\epsilon_i$，其中 $\boldsymbol{x}=(1,x_{i,1},\cdots,x_{i,k})$。定义 $Q(\boldsymbol{\beta})=\sum(y_i-\boldsymbol{\beta}^\top\boldsymbol{x_i})^2$，最小化之，最小二乘估计 $\hat{\boldsymbol{\beta}}$。经验回归函数为 $\hat{\boldsymbol{y}}=\boldsymbol{x}^\top\hat{\boldsymbol{\beta}}$

矩阵形式即为 $\boldsymbol{X}\in\mathbb{R}^{n\times(k+1)}$，这个时候 $Q(\boldsymbol{\beta})=|\boldsymbol{y}-\boldsymbol{X\beta}|^2$。正规方程为 $\nabla Q=-2\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X\beta})=0$，若 $\boldsymbol{X}$ 列满秩，则 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$。

发现 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X\beta}+\boldsymbol{\epsilon})=\beta+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}$，所以 $E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$，无偏。$\mathrm{Cov}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$。

这里接下来处理一元的情况，注意到 $\boldsymbol{X}^\top\boldsymbol{X}=\begin{pmatrix}n & \sum x_i\\ \sum x_i & \sum x_i^2\end{pmatrix}$ 所以行列式为 $n\cdot s_{xx}$，于是 $\mathrm{Cov}(\hat{\boldsymbol{\beta}})=\frac{\sigma^2}{n\cdot s_{xx}}\begin{pmatrix}\sum x_i^2 & -n\overline{x}\\ -n\overline{x} & n\end{pmatrix}$，可以和之前的一元结果对应上。

$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$，令 $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$。性质：$\mathrm{tr}(\boldsymbol{H})=\mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)=\mathrm{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X})=\mathrm{tr}(\boldsymbol{I})=k+1$（trace trick），$\boldsymbol{H}^2=\boldsymbol{H}$，$(\boldsymbol{I}-\boldsymbol{H})^2=\boldsymbol{I}-\boldsymbol{H}$，$\boldsymbol{HX}=\boldsymbol{X}$，$\boldsymbol{H}$ 对称且半正定。其本质是**投影矩阵**，将任意向量投影到 $\boldsymbol{X}$ 的列空间上，所以 $\boldsymbol{y}-\hat{\boldsymbol{y}}$ 垂直于该列空间。

所以 $\hat{\boldsymbol{y}}-\boldsymbol{y}=(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{y}=(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{\epsilon}$，于是 $\mathrm{Cov}(\hat{\boldsymbol{y}}-\boldsymbol{y})=\sigma^2(\boldsymbol{H}-\boldsymbol{I})(\boldsymbol{H}-\boldsymbol{I})^\top=\sigma^2(\boldsymbol{I}-\boldsymbol{H})$，$E(|\hat{\boldsymbol{y}}-\boldsymbol{y}|^2)=\sigma^2\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H})=\sigma^2(n-k-1)$，所以 $\frac{1}{n-k-1}|\hat{\boldsymbol{y}}-\boldsymbol{y}|^2$ 为 σ^2 的无偏估计量。对于一元的情况，$k=1$。

$\mathrm{SST}=\sum(y_i-\overline{y})^2$（总平方和），$\mathrm{SSE}=\sum(\hat{y}_i-y_i)^2$（残差平方和），$\mathrm{SSR}=\sum(\hat{y}_i-\overline{y})^2$（回归平方和）。有 $\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}$。（证明：知道 $\boldsymbol{y}-\hat{\boldsymbol{y}}$ 垂直于 $C(X)$，而显然 $\overline{\boldsymbol{y}}$ 和 $\hat{\boldsymbol{y}}$ 在列空间内，勾股定理）定义 $R^2=\mathrm{SSR}/\mathrm{SST}$，表示回归模型对总变异的解释比例。有 $R^2=1-\mathrm{SSE}/\mathrm{SST}$ 且 $R^2\in[0,1]$。