Goal: $E[y|x_1,...,x_n] = f(x_1,...,x_n)$ Additive noise implies: $y\ \mathcal{N}(\mathbf{X}\theta, \sigma^2 * \mathbf{I})$

Residual: $\hat\epsilon_i = y_i - x_i^T\hat\theta$;

Partial Residual: $\hat\epsilon_{x_j,i} = y_i - \mathbf{x}_i^T\theta + \theta_j x_{ij}$

Normal equations: $\mathbf{X}^T\mathbf{X}\hat\theta = \mathbf{X}^T y$

Thm: $A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m$. TFAE:
i) Col(A) is lin. indep., ii) $\mathbf{A}^T\mathbf{A}$ invertible

In this case the LS solution is: $\hat\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$\hat{\mathbf{y}} = \mathbf{X}\hat\theta = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y} =: \mathbf{P}\mathbf{y} \Rightarrow P^2=P=P^T$, $Tr(P)=p$; $r=(Id-P)y =: Qy \Rightarrow Q^T = Q^2 = Q, PQ=QP=0, Tr(Q)=n-p$

$\epsilon|X \sim \mathcal{N}(0,\sigma^2 Id) \Rightarrow f(y_1,...,y_n|X) = L_{y,X}(\theta,\sigma^2) = \prod_{i=1}^n \frac{1}{\sigma}\phi(\frac{(y_i-x_i^T\theta)}{\sigma})? \Rightarrow \hat\sigma^2_{MLE} = \frac{\sum_{i=1}^n(y_i-\hat y_i)^2}{n}$ but $\hat\sigma^2 = \frac{\sum_{i=1}^n (y_i-\hat y_i)^2}{n-p}$ unbiased

Props: Ass. $Y=X\theta+\epsilon, \mathbb{E}[\epsilon]=0$, $\text{Cov}(\epsilon) = \mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I_n$. i) $\mathbb{E}[\hat\theta]=\theta$, ii) $\mathbb{E}[\hat\epsilon]=0, \mathbb{E}[\hat y] = \mathbb{E}[y] = X\theta$, iii) $\text{Cov}(\hat\theta) = \sigma^2(X^TX)^{-1}$, iv) $\text{Cov}(\hat y) = \text{Cov}(Py) = \sigma^2 PP^T = \sigma P$, v) $\text{Cov}(\hat\epsilon) = \sigma^2 Q$, vi) $\text{Cov}(\hat\epsilon,\hat y)=0$, vii) $E[\sum_{i=1}^n r_i^2] = \sigma^2(n-p) \Rightarrow \hat\sigma = \frac{||X\hat\theta - y||^2}{n-p}$ still unbiased.

Props: Ass. $Y = X\theta+\epsilon$, $\epsilon\sim\mathcal{N}(0,\sigma^2 I_n)$.
i) $\hat\theta \sim \mathcal{N}_p(\theta, \sigma^2(X^TX)^{-1})$, ii) $\hat y \sim \mathcal{N}_n(X\theta, \sigma^2 P)$, $\hat\epsilon\sim\mathcal{N}(0,\sigma^2 Q)$, iii) $\hat y \perp\!\!\!\perp \hat\epsilon$, iv) $(\sum_{i=1}^n r_i^2)/\sigma^2 \sim \chi^2_{n-p}$, v) $\hat\sigma^2 \perp\!\!\!\perp \hat\theta_{LSE}$

Props: Ass. $Y = X\theta+\epsilon$, $\epsilon\sim\mathcal{N}(0,\sigma^2 I_n)$.
i) $\frac{\hat\theta-\theta}{\hat\sigma\sqrt{(X^TX)^{-1}_{ii}}} \sim t_{n-p}$, ii) $\frac{(\hat\theta-\theta)^T(X^TX)(\hat\theta-\theta)}{p\hat\sigma^2} \sim F_{p,n-p}$, iii) $\vartheta = B\theta$, $B\in\mathbb{R}^{q\times p}$, $V = B(X^TX)^{-1}B^T \Rightarrow \frac{(\hat\vartheta-\vartheta)^TV^{-1}(\hat\vartheta-\vartheta)}{q\hat\sigma^2}\sim F_{p,n-p}$, iv) $\frac{(\hat y_i - \mathbb{E}[y_i])}{\hat\sigma\sqrt{P_{ii}}}\sim t_{n-p}$, v) $\frac{\hat y_0 - \mathbb{E}[y_0]}{\hat\sigma\sqrt{x_0^T(X^TX)^{-1}x_0}} \sim t_{n-p}$, vi) $\frac{y_0-\hat y_0}{\hat\sigma\sqrt{1+x_0^T(X^TX)^{-1}x_0}}\sim t_{n-p}$

Thm: If $(\epsilon_i)_i$ iid, $\mathbb{E}[\epsilon_i]=0$, $\text{Var}(\epsilon_i)=\sigma^2$, $\lambda_{min,n} = \min\{\sigma(X^TX)\}\to\infty$ and $\max_j P_{jj}\max_j x_j^T(\sum_{i=1}^n X_iX_i^T)^{-1}x_j \to 0$, then $\hat\theta_{LSE}$ is consistent (for θ), and $(X^TX)^{1/2}(\hat\theta-\theta)\to\mathcal{N}_p(0,\sigma^2 I_n)$.

R-output: $H_{0,j}:\theta_j = 0 \Rightarrow T_j \frac{\hat\theta_j}{\hat\sigma\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p}, p_j = P(|T_j|\ge|\hat T_j|)$

$H_{0,\texttt{global}}: \theta_1=\dots=\theta_k = 0 \Rightarrow T = \frac{||\hat y-\bar y||^2/(p-1)}{||y-\hat y||^2/(n-p)} \sim F_{p-1,n-p}, p = P(|T|\ge|\hat T|)$, $R^2 = \frac{||\hat y-\bar y||}{y-\bar y}$

Partial F-test: $H_{0,B}: B\theta = b\in\mathbb{R}^p \Rightarrow \frac{(B\hat\theta-b)^T(B(X^TX)^{-1}B^T)^{-1}(B\hat\theta-b)}{(p-q)\hat\sigma^2}\sim F_{p-q,n-p}$

E.g.: $H_{0,B}: B\theta = 0\in\mathbb{R}^p$, $B=(I_{p-q}|0_q)\in\mathbb{R}^{(p-q)\times p}$ (first $p-q$ coeffs $=0$) $\Rightarrow \frac{(SSE_0-SSE)/(p-q)}{SSE/(n-p)} = \frac{||\hat y - \hat y^{(0)}||^2/(p-q)}{||y-\hat y||^2/(n-p)} \sim F_{p-q,n-p}$

E.g.: $H_{0,B}: B\theta=0\in\mathbb{R}^p$, $B=(0|I_{p-1})\in\mathbb{R}^{(p-1)\times p}$ $(H_{0,\texttt{global}}:\theta_1=\dots=\theta_k=0)$, $\hat\theta_{(0)} = (\bar y, 0,.,0)^T$, $\hat y_{(0)} = \bar y \Rightarrow F = \frac{||\hat y-\bar y||^2/(p-1)}{||y-\hat y||^2/(n-p)}\sim F_{p-1,n-p}$

ANOVA:$||y-\hat y^{(0)}||^2 = ||y-\hat y||^2 + ||\hat y - \hat y^{(0)}||^2$

.

.

.

Corr.: $\hat\rho = \frac{\sum(x_i-\bar x)(y_i-\bar y)}{\sqrt{(x_i-\bar x)^2}\sqrt{(y_i-\bar y)^2}}$. $\rho\approx 0\Rightarrow \text{Var}(\rho)$ high. $\rho\approx\pm1 \Rightarrow \text{Var}(\rho)$ low. Stabilize by $z := \tanh^{-1}(\hat\rho)\sim\mathcal{N}(\tanh^{-1}(\rho), \frac{1}{n-3})$. Test $H_0:\rho=0$ by z-trafo, t-/F-test of $\beta = 0$.

Sp.'s Rank: $r_S = 1 - \frac{6\sum_{i=1}^n D_i^2}{n(n^2-1)}$, $D_i := Rg(X_i)-Rg(Y_i)$.

Kend.'s Rank: $r_K = 2\frac{T_k-T_d}{n(n-1)}$, $T_k = |\{(i,j):(x_i-x_j)(y_i-y_j)>0\}|$, $T_d = |\{(i,j):(x_i-x_j)(y_i-y_j)<0\}|$

$\rho_{XY.Z} := \frac{\rho_{XY}-\rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho^2_{XZ})(1-\rho^2_{YZ})}}$=corr. of x and y after accounting for z

Reg2mean:$y-\bar y = \hat\rho\frac{\hat\sigma_Y}{\hat\sigma_X}(x-\bar x)$

Norm. Plot:$\hat F_n(x) := |\{X_i\le x\}|$.
$H_0: X_1,...,X_n\sim\mathcal{N}(\mu,\sigma^2)\Rightarrow \hat F_n(x)\approx\Phi(\frac{x-\mu}{\sigma})\Rightarrow z:=\Phi^{-1}(\hat F_n(x))\Rightarrow z\approx\frac{x-\mu}{\sigma}$ $\hat F_n(X_{(i)}) = i/n$.

Tukey-Anscombe: $\sum r_i\hat y_i = 0$. Plot non-linear ⇒ model ass. broken.

TS plot: $\epsilon\sim\mathcal{N}_n(0,\Sigma)\Rightarrow \hat\theta\sim\mathcal{N}_p(\theta,(X^TX)^{-1}(X^T\Sigma X)(X^TX)^{-1})$

Durbin-Watson: $T = \frac{\sum_{i=1}^{n-1}(r_{i+1}-r_i)^2}{\sum_{i=1}^n r_i^2}\approx 2(1-\frac{\sum_{i=1}^{n-1}r_ir_{i+1}}{\sum_{i=1}^n r_i^2}) \overset{\epsilon_i\perp\!\!\!\perp\epsilon_{i+1}}{\approx} 2$

GLS:$Y = X\theta+\epsilon$, $\epsilon\sim\mathcal{N}(0,\sigma^2)$, $\Sigma = AA^T$ known and pos. def., A reg. ⇒ $\tilde y := A^{-1}y = A^{-1}(X\theta+\epsilon) = \tilde X\theta+\tilde\epsilon$, $||\tilde y - \tilde X\theta||^2 = (y-X\theta)^T\Sigma^{-1}(y-X\theta)\Rightarrow$
$\hat\theta = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$
$\hat\theta\sim\mathcal{N}_p(\theta,\sigma^2(X^T\Sigma^{-1}X)^{-1})$

Alg: `1.LS`→ $\hat\theta^{(1)}, r^{(1)}$ `2.`$\hat\Sigma^{(i+1)}(\hat\theta^{(i)}, r^{(i)})$ `3.GLS(`$\hat\Sigma$`)` → $\hat\theta^{(i+1)}, r^{(i+1)}$ `4.Repeat 2.and 3. until conv`

E.g.:$\Sigma = \text{diag}(v_1,...,v_n)\Rightarrow ||y-X\theta||^2 = \sum_i r_i^2/v_i$,

Forward greedy: `1.`$X^{(1)} = (1,..,1)^T$ `2.`$X^{(i+1)} = (X^{(i)}|X_j)$, X_j = `most significant F-value of models w. remaining covars (or "which is most significant t-test for new var in new model with that var") 3.Repeat until no F-vals are significant`

Backward greedy: `1.`$X^{(1)}=X$ `2.Def` $X^{(i+1)}$ `as` $X^{(i)}$ `without the covar "whose F-value in the comparative test is smallest" (or has highest p-value in t-test 3.Repeat until all F-(or t-)stats are significant`

Model quality:M correct⇒
$\mathbb{E}[\hat y^M] = X^M((X^M)^TX^M)^{-1}(X^M)^T\mu$,
$\text{Cov}(\hat y^M) = \sigma^2X^M((X^M)^TX^M)^{-1}(X^M)^T$,
$\sum_{i=1}^n\text{Var}(\hat y_i^M) = Tr(\text{Cov}(\hat y^M)) = |M|\sigma^2$.
$SMSE = SMSE(M) = \mathbb{E}[||\hat y^M-\mu||^2] = \sum_{i=1}^n\mathbb{E}[(\hat y_i^M-\mu_i)^2] = \sum_{i=1}^n \text{Var}(\hat y_i^M) + \sum_{i=1}^n(\mathbb{E}[\hat y_i^M]-\mu_i)^2 = |M|\sigma^2 + \sum_{i=1}^n(\mathbb{E}[\hat y_i^M]-\mu_i)^2$
$\Gamma_p(M) = \frac{SMSE}{\sigma^2}\ge|M|$ and $=|M|$ if unbiased
$SPSE = \sum_{i=1}^n\mathbb{E}[(Y_{n+i}-\hat y_i^M)^2] = \sum_{i=1}^n\mathbb{E}[(Y_{n+i}-\mu_i)^2] + \sum_{i=1}^n\mathbb{E}[(\hat y_i^M-\mu_i)^2] = n\sigma^2+SMSE$
$SSE(M) = ||y-\hat y^M||^2 = \sum_{i=1}^n(y_i-\hat y_i^M)^2$
$\mathbb{E}[||y-\hat y^M||^2] = \sum_{i=1}^n\text{Var}(y_i-\hat y_i^M) + \sum_{i=1}^n(\mathbb{E}[y_i]-\mathbb{E}[\hat y_i^M])^2 = (n-|M|)\sigma^2 + \sum_{i=1}^n(\mathbb{E}[y_i]-\mu_i)^2 = SPSE - 2|M|\sigma^2$
$\hat{SPSE} = SSE + 2|M|\sigma^2$
$C_p(M) := \frac{SSE(M)}{\hat\sigma^2} - n + 2|M| = \hat\Gamma_p$
$AIC(\alpha) = -2\hat l_{|M|} + \alpha|M|$
Model select by min. AIC.

Gauss-Markov:$Y=X\theta+\epsilon$, $\mathbb{E}[\epsilon]=0$, $\text{Cov}[\epsilon]=\sigma^2I_n, rank[X]=p$, $c\in\mathbb{R}^p$ arb.. Then $c^T\hat\theta_{MLE}$ has minimal variance among all **linear** unbiased estimators of $c^T\theta$.

Cor: Let furthermore ϵ be normally distributed. Then $c^T\hat\theta$ has minimal variance amongst all unbiased estimators of $c^T\theta$.

Cramer-Rao:?

$\theta^{(-i)}$:=LSE without i-th obsv.
$\hat\theta^{(-i)} - \hat\theta = -\frac{r_i}{1-P_{ii}}(X^TX)^{-1}x_i$

Cook's d: $D_i := \frac{(\hat\theta^{(-i)}-\hat\theta)^T(X^TX)(\hat\theta^{(-i)}-\hat\theta)}{p\sigma^2} = \frac{1}{p}\frac{r_i^2}{\hat\sigma^2(1-P_{ii})}\frac{P_{ii}}{1-P_{ii}}$

(x,y) new, then $\Delta\hat\theta_{LSE} = \frac{(X^TX)^{-1}x(y-x^T\hat\theta)}{1+x^T(X^TX)^{-1}x}$
$\Rightarrow \Delta\hat\theta_{LSE}\sim\frac1n(\mathbb{E}[x_ix_i^T])^{-1}x(y-x^T\theta)$ (as $n\to\infty$).

Huber reg: $\hat\theta = \arg\min_\theta\sum_{i=1}^n\rho_c(y_i-x_i^T\theta)$,
$\rho_c(u) = \frac12u^21\{|u|\le c\} + c(|u|-\frac c2)1\{|u|\ge c\}$
$\overset{d/d\theta=0}{\Rightarrow}\sum_{i=1}^n\psi_c(y_i-x_i^T\theta)x_i = 0$, where $\psi_c(u) = \rho_c(u)' = \text{sgn}(u)\min(|u|,c)$.

H's proposal 2: $\psi_c(u) = \rho_c(u)^2-\beta$
Set $\sum_{i=1}^n\psi_c(\frac{y_i-x_i^T\hat\theta}{\hat\sigma})x_i = 0 \to \sum_{i=1}^n\chi(y_i-x_i^T\theta)x_i=0$ with $\chi(u)$ =Huber's proposal 2 or $=\text{sgn}(|u|-\beta)$ s.t. $\int\chi(u)\exp(-u^2/2)du = 0\Rightarrow\hat\sigma$ consistent.
No closed form for Huber and L1. **Alg.**

Huber: Iterate diag WLS
$\frac{1}{v_i} = w_i \propto \frac{\psi_c((y_i-x_i^T\hat\theta)/\hat\sigma)}{y_i-x_i^T\hat\theta}\propto\min(1,\frac{c\hat\sigma}{|y_i-x_i^T\hat\theta|}$ until stabilizes.⇒ $\sqrt n(\hat\theta_{GLS}-\theta)\overset{d}{\to}\mathcal{N}(0,\frac{\mathbb{E}[\psi_c(\epsilon/\sigma)^2}{P(|\epsilon_i|\le c\sigma)^2}\sigma^2\mathbb{E}[x_ix_i^T]^{-1}$
$\Delta\hat\theta\sim\frac{(\mathbb{E}[x_ix_i^T])^{-1}x\psi_c((y-x^T\theta)/\sigma)\sigma}{nP(|\epsilon_i|\le c\sigma)}$

Alts:$\sum_{i=1}^n\eta(x_i,\frac{y_i-x_i^T theta}{\hat\sigma})x_i = 0$
Mallows:$\eta(x,r) = \min(1,\frac{a}{||Ax||}\psi_c(r)$ (lower w_i's 4 deviant x)
Schweppe:$\eta(x,r) = \frac{1}{||Ax||}\psi_c(||Ax||r)$ (lowers corner in ψ_c, lets w_i 4 deviant x be large if $r\approx0$)
A chosen s.t. $||Ax||^2 = C\cdot x^T(X^TX)^{-1})x$
$\Delta\hat\theta\sim\frac1n(\mathbb{E}[\frac{\partial}{\partial r}\eta(x_i,\frac{\epsilon_i}{\sigma})x_ix_i^T])^{-1}x\eta(x,\frac{y-x^T\theta}{\sigma})\sigma$
$\arg\min_\theta\arg\text{median}((y_i-x_i^T\theta)^2)$ sandwiches 50% of D (beakdown?), less eff. than H. & S.

Nonlin: Sps. $f(x_i,\theta)\approx f(x_i,\theta_0)+a(\theta_0)_i^T(\theta-\theta_0)$ w. $a(\theta)_i = (\frac{\partial}{\partial\theta_j}f(x_i,\theta);j=1,..,p)^T$
$\hat\theta\overset{d}{\to}\mathcal{N}(\theta_0,\sigma^2(A(\theta_0)^TA(\theta_0))^{-1})$, where $A(\theta) = (a(\theta)_1,..,a(\theta)_p)^T$

Conf Int: $\hat\theta_k\pm F^{-1}_{t_{n-p}}(1-\alpha/2)se(\hat\theta_k)$ w. $se(\hat\theta_k) = \hat\sigma\sqrt{((A(\hat\theta)^TA(\hat\theta))^{-1})_{kk}}$.

$\hat{\theta}_0 = \arg\min_{\theta; B\theta=b} S(\theta)$

$T = \frac{(S(\hat{\theta}_0)-S(\hat{\theta}))/q}{S(\hat{\theta})/(n-p)} \approx F_{q,n-p}$

E.g.:$T_k(\theta_k^*) = \frac{S(\hat{\theta}^{(-k)})-S(\hat{\theta})}{S(\hat{\theta})/(n-p)} = \frac{S(\hat{\theta}^{(-k)})-S(\hat{\theta})}{\hat{\sigma}^2}$, where $\hat{\theta}^{(-k)} := \arg\min_{\theta;\theta_k=\theta_k^*} S(\theta)$. **skipped somin here**

GLM: $p_{\beta_i}(y_i) = \exp(y_i\beta_i + c(\beta_i))h(y_i) \to \mathbb{E}[y_i] = -c'(\beta_i) = \mu(\beta_i) \to g(\mu(\beta_i)) = x_i^T\theta$, $g: D \to \mathbb{R}$ w. D suitable and g bij

Eg: Gauss $p(y) = \exp(y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2})$. If σ known, then $\beta = \frac{\mu}{\sigma^2}$, $c(\beta) = -\frac{1}{2}\sigma^2\beta^2$

Eg: Binom $p(y) = \binom{n}{y}p^y(1-p)^{n-y}$ w. $\beta = \log\frac{p}{1-p}$, $c(\beta) = -n\log(1+\exp(\beta))$

Eg: Poisson $p(y) = \frac{\exp(y\log\lambda-\lambda)}{y!}$ w. $\beta = \log\lambda$, $c(\beta) = -\exp(\beta)$

E.g.:log reg: $\log(\frac{P_\theta(Y_i=1)}{P_\theta(Y_i=0)}) = \sum_{j=1}^p x_{ij}\theta_j = x_i^T\theta \Rightarrow$ $P_\theta(Y_i=1) = \frac{\exp(x_i^T\theta)}{1+\exp(x_i^T\theta)} = P(U \geq -x_i^T\theta)$ w. log. distr. $P(U \leq u) = P(U \geq -u) = \frac{\exp(u)}{1+\exp(u)} = \int_{-\infty}^u \frac{\exp(t)}{(1+\exp(t))^2}dt$.

$Z_i = x_i^T\theta + \epsilon_i$, $Y_i = 1\{Z_i \geq\}$

$P_\theta(y) = p^y(1-p)^{1-y} = \exp(yx_i^T\theta - \log(1+\exp(x_i^T\theta))$

$l(\theta) = \sum P_\theta(Y_i = y_i) \overset{d/d\theta}{\Rightarrow} \sum_i (y_i - P_{\hat{\theta}}(Y_i=1))x_i = 0$

$\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta, V(\theta))$ w. $V(\theta)^{-1} = I(\theta) = \sum_i x_ix_i^T\mathbb{E}[(y_i - P_\theta(Y_i=1))^2 = \sum_i x_ix_i^T\frac{\exp(x_i^T\theta)}{1+\exp(x_i^T\theta)}$

$2(l(\hat{\theta}^{(p)}) - l(\hat{\theta}^{(q)})) \xrightarrow{d} \chi^2_{p-q}$

Alg:`1. Initialize`

$\hat{p}_i = 0.99 \cdot 1\{y_i = 1\} + 0.01 \cdot 1\{y_i = 0\}$

`2. Taylor exp. of logit:`

$Z_i := logit(\hat{p}_i) + logit'(\hat{p}_i)(Y_i - \hat{p}_i) \approx x_i^T\hat{\theta} + \frac{1}{\hat{p}_i(1-\hat{p}_i)}(Y_i - \hat{p}_i)$

`3. Do weighted least squares` ($\min\sum\frac{1}{v_i}(y_i - f(x_i))^2$) `of Z versus X with weights` $w_i = 1/v_i = \hat{p}_i(1-\hat{p}_i)$ `to get new` $\hat{\theta}$

`4. Compute` $\hat{p}_i = logit^{-1}(x_i^T\hat{\theta})$

`Repeat 2-4 until convergence`

$\hat{\theta}_{IRLS} = \hat{\theta}_{MLE} \xrightarrow{d} \mathcal{N}(\theta, (X^TWX)^{-1})$, w. $W = dig(p_i(1-p_i))$, $se(\hat{\theta}_j) = \sqrt{(X^T\hat{W}X)^T}$

Cox reg: Let T_i be the failure time with pdf f and cdf F.

$\lambda(t) = \lim_{h\to 0}\frac{1}{h}P(t \leq T \leq t+h | T \geq t) = \frac{f(t)}{1-F(t)} = -\frac{d}{dt}\log(1-F(t))$=failure rate w.

$F(t) = 1 - \exp(-\int_0^t \lambda(u)du)$

$\lambda_i(t) = \exp(x_i^T\theta)\lambda_0(t)$=failure rate of i-th covar.

Partial likelihood: $\partial l(\theta) = \prod_i \frac{\exp(x_i^T\theta}{\sum_{j\in\Lambda_i}\exp(x_j^T\theta)}$, where $\Lambda_i = \{j : t_j \geq t_i\}$ and the i-th factor is the conditional probability of failure of the i-th observed unit in the interval $[t_i, t_i + dt)$.

Censored data: Let T_i be unknown but $1\{T_i > C_i\}$ known, C_i=censoring time.

Partial likelihood:

$\partial l(\theta) = \prod_i \frac{\exp(x_i^T\theta)}{\sum_{j\in\Lambda_i^1}\exp(x_j^T\theta)+\sum_{j\in\Lambda_i^2}\exp(x_j^T\theta)}$, where $\Lambda_i^1 = \{j : t_j \geq t_i$, j uncen.$\}$ and $\Lambda_i^2 = \{j : c_j \geq t_i$, j cen.$\}$.

$\hat{\lambda}_i(t) = \exp(x_i^T\theta)\hat{\lambda}_0(t)$, $\frac{\lambda_i(t)}{\lambda_j(t)} = \exp((x_i - x_j)^T\theta)$,

$\hat{\theta} = \arg\max_\theta \partial l(\theta)$

Pearson residual: $R_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$

Deviance residual:

$D_i = s_i\sqrt{-2(Y_i\log\hat{p}_i + (1-Y_i)\log(1-\hat{p}_i)} = \sqrt{\text{i-th summand in } -2l(\cdot)}$,

$s_i = 1\{Y_i = 1\} - 1\{Y_i = 0\}$

Interpretation: $\sum_{i=1}^n D_i^2 = -2l(\hat{\theta})$ =goodness of fit

Non-parametric reg: Let K be a symmetric probability density with $\text{supp}(K) = [-1,1]$ or K decays very rapidly, $h > 0$ bandwidth.

Nadaraya-Watson: $\hat{f}(x) = \frac{\sum_i y_iK((x-x_i)/h)}{\sum_i K((x-x_i)/h)}$

Gasser-Müller:(Ass. $0 \leq x_1 \leq \cdots \leq x_n \leq 1$)

$s_0 = -\infty$, $s_i = (x_i + x_{i+1})/2$ for $0 < i < n$ and $s_n = \infty$

$\hat{f}(x) = \sum_{i=1}^n y_i\int_{s_{i-1}}^{s_i}\frac{1}{h}K((x-u)/h)du$.

Local poly: $\hat{f}(x) = \hat{\theta}_0(x)$, with $\hat{\theta}(x) = \arg\min_\theta\sum_{i=1}^n K((x-x_i)/h)(y_i - \sum_{j=0}^p\theta_j(x_i-x)^j)^2$.

$p = 1$ and $p = 3$ often chosen. R-func=loess: chooses variable h nearest neighbor to ensure fixed number of obsv. in $[x-h, x+h]$

Smoothing spline:

$\hat{f} = \arg\min_f\sum_{i=1}^n(y_i - f(x_i))^2 + \lambda\int_0^1 f''(x)^2dx$

Linear on $[0, x_1]$ and $[x_n, 1]$. $\lambda \to 0$: interpolates data, $\lambda \to \infty$: least squares.

Bias-Variance tradeoff:

$\mathbb{E}[\hat{f}(x)] - f(x) \sim C(K,p)h^{p+1}f^{(p+1)}(x)$ and $\text{Var}[\hat{f}(x)] \sim C(K,p)\frac{\sigma_\epsilon^2}{nh}(\frac{1}{nh}\sum K((x-x_i)/h))^{-1}$, assuming $h_n \to 0$ and $h_nn \to \infty$.

h small $\to$ small (absolute) bias

h large $\to$ small variance

$\mathbb{E}[(\hat{f}(x) - f(x))^2] = \text{Var}[\hat{f}(x)] + (\mathbb{E}[\hat{f}(x)] - f(x))^2 = O((nh)^{-1}) + O(h^{2(p+1)})$.

Above is min. when both summands have same order, i.e. $h = O(n^{-1/(2p+3)})$, thus $O(n^{-(2p+2)/(2p+3)})$

High-dimensional stuff:

Smoothing spline: Let $x_1 < \cdots < x_n$. Chose SS s.t. $f(x) = \sum_{j=1}^n N_j(x)\theta_j$ w. $N_1(x) = 1$, $N_2(x) = x$, $N_{k+1} = d_k(x) - x_{k-1}(x)$ w.

$d_k(x) = \frac{(x-x_k)_+^3 - (x-x_n)_+^3}{x_n - x_k}$

$\hat{\theta} = \arg\min_\theta ||Y - N\theta||^2 + \lambda\theta^T\Omega\theta$,

$N = [N_j(x_i)]_{i,j=1}^n$, $\Omega_{jk} = \int_j N_j''(x)N_k''(x)dx \Rightarrow$ $\hat{\theta}_\lambda = (N^TN + \lambda\Omega)^{-1}N^TY$

Ridge Reg: $\hat{\theta} = \arg\min_\theta ||Y - X\theta||_2^2 + \lambda||\theta||_2^2 = (X^TX + \lambda I_p)^{-1}X^TY$ w. SVD:

If $p < n$: $X = UDV^T$ $((n\times p)\times(p\times p)\times(p\times p))$, $U^TU = I_n$, $V^TV = VV^T = I_p$

If $p \geq n$: $X = UDV^T$ $((n\times n)\times(n\times n)\times(n\times p))$, $U^TU = UU^T = I_n$, $V^TV = I_p$

$col(U) = col(X)$, $col(V) = row(X)$

$\hat{\theta} = (VD^TDU^TUDV^T + \lambda I)^{-1}VDU^TY = V\Lambda U^TY$ w. $\Lambda = diag(\frac{D_{ii}}{D_{ii}^2+\lambda} : i = 1, .., \min(n,p))$

$(D_{11} \geq D_{\min(n,p)} > 0)$

$\mathbb{E}[\hat{\theta} = VD\Lambda V^T\theta \to VV^T\theta$ = projection of θ onto row space of X

If $\min(n,p) = p$: $VV^T\theta = \theta$

If $p > n$: $\hat{\theta}$ has bias $(VV^T - I)\theta$

Identities: $\text{Cov}[AY + b] = A\,\text{Cov}[Y]A^T \Rightarrow$ $a^T\,\text{Cov}[Y]a \leq 0$.

$\phi(x) := \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$, $\Phi(x) = \int_{-\infty}^x\phi(y)dy$

$X \sim \mathcal{N}(0,1) \Rightarrow \mathbb{E}[X^3] = 0, \mathbb{E}[X^4] = 3$

$X \sim \mathcal{N}(\mu, \Sigma) \Rightarrow f_X(x) =$ $(2\pi)^{-n/2}(|\det\Sigma|)^{-1/2}\exp[(x-\mu)^T\Sigma^{-1}(x-\mu)]$, $\Sigma = AA^T$