Distribution	p.d.f.	Support	mean	variance	Dispersion ϕ	Cumulant $b(\theta)$	$\mu(\theta)$	Canonical Link $\theta(\mu)$	$V(\mu)$
Normal (μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$(-\infty,\infty)$	μ	σ^2	σ^2	$\theta^2/2$	θ	μ (identity)	1
Poisson (μ)	$\mu^x e^{-\mu}/x!$	$\mathbb{Z}^+$	μ	μ	1	e^θ	e^θ	$\log\mu$ (log)	μ
Binomial(n,p)	$\binom{n}{xn}p^{xn}(1-p)^{n-xn}$	$\{0,1,...,n\}/n$	p	$p(1-p)/n$	1/n	$\log(1+e^\theta)$	$e^\theta/(1+e^\theta)$	$\log(\frac{p}{1-p})$ (logit)	$\mu(1-\mu)$
Gamma(μ,ν)	$\frac{1}{\Gamma(\nu)}(\frac{\nu}{\mu})^\nu x^{\nu-1}\exp(\frac{-\nu x}{\mu})$	$(0,\infty)$	μ	μ^2/ν	$1/\nu$	$-\log(-\theta)$	$-\theta^{-1}$	$1/\mu$ (reciprocal)	μ^2

1. LINEAR ALGEBRA

Setup. Consider a matrix $\mathbf{A}_{n\times n}$ with eigenvalues λ_i.

$\text{Trace}(\mathbf{A}) = \sum\lambda_i$ and $\text{Det}(\mathbf{A}) = \prod\lambda_i$.
$\text{Trace}(\mathbf{AB}) = \text{Trace}(\mathbf{BA})$. More generally invariant under cyclic permutations: $\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{BCA}) = \text{Tr}(\mathbf{CAB})$.
$\mathbb{E}(x^T\mathbf{A}x) = \mathbb{E}(\text{Trace}(x^T\mathbf{A}x)) = \text{Trace}(\mathbf{A}\mathbb{E}(xx^T))$

Generalized Inverse $\mathbf{A}^- := \mathbf{G}$ where $\mathbf{AGA} = \mathbf{A}$. Always exists, in general not unique.

SVD $\mathbf{X}_{n\times p} = \mathbf{UDV}^T$ where $\mathbf{U}_{n\times p}$ and $\mathbf{V}_{p\times p}$ are orthogonal, and span $C(\mathbf{X})$, $C(\mathbf{X}^T)$, resp. $\mathbf{D} = \text{diag}(d_1,...,d_p)$ with $d_1 \geq ... \geq d_p \geq 0$. If $d_p = 0$, then $\mathbf{X}$ is singular.

Eigen Decomp $\mathbf{A} = \mathbf{VDV}^{-1}$, where $\mathbf{D}$ is diagonal, with $\mathbf{D}_i = \lambda_i$, and the columns of $\mathbf{V}$ are eigenvectors. If $\mathbf{A}$ is real symmetric, $\mathbf{V}$ is orthogonal.

Using SVD for $\mathbf{X}$, we have $\mathbf{X}^T\mathbf{X} = \mathbf{VD}^2\mathbf{V}^T$

Projection $\mathbf{P}_{n\times n}$ is a projection onto subspace W if (i.) $\mathbf{P}y = y$ all $y \in W$ and (ii.) $\mathbf{P}y = 0$ all $y \in W^\perp$.
· $\mathbf{P}$ is a projection matrix if and only if $\mathbf{P}$ is symmetric and idempotent ($\mathbf{P}^2 = \mathbf{P}$). (Note: $W = C(\mathbf{P})$.)
· The eigenvalues of $\mathbf{P}$ are all 0 or 1.
· $\text{Trace}(\mathbf{P}) = \text{Rank}(\mathbf{P})$
· $\mathbf{I} - \mathbf{P}$ projects onto $W^\perp$.

Every $y \in \mathbb{R}^n$ has unique decomposition $y = y_1 + y_2$ where $y_1 \in W$ and $y_2 \in W^\perp$.

$C(\mathbf{X})^\perp = N(\mathbf{X}^T)$. That is, $y \in C(\mathbf{X})^\perp \Leftrightarrow \mathbf{X}^T y = 0$.

For nested models $M_a \subset M_b$ with model matrices $\mathbf{X}_a, \mathbf{X}_b$ ($C(\mathbf{X}_a) \subset C(\mathbf{X}_b)$) and projections $\mathbf{P}_a, \mathbf{P}_b$, we have:
$\mathbf{P}_a\mathbf{P}_b = \mathbf{P}_b\mathbf{P}_a = \mathbf{P}_a$. And $\mathbf{P}_b - \mathbf{P}_a$ is a projection matrix.

Lemma. Suppose $\{\mathbf{P}_i\}$ are symmetric $n \times n$ matrices s.t. $\sum_i \mathbf{P}_i = \mathbf{I}$. The following are equivalent: (i.) $\mathbf{P}_i$ idempotent each i. (ii.) $\mathbf{P}_i\mathbf{P}_j = \mathbf{0}$ all $i \neq j$. (iii.) $\sum_i \text{Rank}(\mathbf{P}_i) = n$.

Lemma. $\nabla_a a^T\mathbf{B}a = (\mathbf{B} + \mathbf{B}^T)a$ ($= 2\mathbf{B}a$ when B sym)

$\mathbb{E}y_{n\times 1} = \mu \Rightarrow \text{Var}y = \mathbb{E}(y-\mu)(y-\mu)^T = \mathbb{E}(yy^T) - \mu\mu^T$.
$\hat{\Sigma}(X_{n\times p}) = n^{-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})^T$ ($= \frac{X^TX}{n}$ if centered)

$\text{proj}_a(b) = \frac{a^Tb}{a^Ta}a$. Orthogonalize b w.r.t. a: $b^* = b - \text{proj}_a(b)$

2. SUPERVISED LEARNING

$\text{EPE}(f) = \mathbb{E}(L(Y_{1\times 1}, f(X_{p\times 1})) = \mathbb{E}_X\mathbb{E}_{Y|X}(L(Y, f(X))|X)$.
Under sq err loss, minimizer is $f(x) = \mathbb{E}(Y|X = x)$.
f approx linear: $f(x) \approx x^T\beta \Rightarrow \hat{\beta} = \mathbb{E}(XX^T)^{-1}\mathbb{E}(XY)$

KNN assumes f well-approx by locally constant function attempt direct est of $\mathbb{E}(Y|X = x)$, $\hat{y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$ average over obs in $\mathcal{T}$ close (k-nearest) to x. Eff df is N/k.

For categorical response G with K classes $G_1, ..., G_K$,
$\text{EPE}(\hat{G}) = \mathbb{E}_X \sum_{k=1}^K L(G_k, \hat{G}(X))\mathbb{P}(G = G_k|X)$.
0-1 loss $\Rightarrow \hat{G}(x) = \text{argmax}_k\, \mathbb{P}(G = G_k|X = x)$ (Bayes)

KNN non-linear decision boundary, no strong assumptions, low bias, high variance/unstable, no longer local as p increases, ignores any special structure in data.

3. LINEAR METHODS FOR REG

Model: $y = X\beta + \varepsilon$, where $\mathbb{E}(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2 I$.

Def (Identifiability). β is identifiabile if $X\beta = X\beta^*$ implies $\beta = \beta^*$. More generally, $\ell^T\beta$ is identifiable if $X\beta = X\beta^* \Rightarrow \ell^T\beta = \ell^T\beta^*$. (Note: $\ell^T\beta$ can be identifiable when β is not.)

Def (Estimability). $\ell^T\beta$ is estimable if $\exists a$ s.t. $\mathbb{E}a^Ty = \ell^T\beta$. That is, $\ell \in C(X^T)$. Estimable quantities are linear functions $a^T\mu$ of $\mu = X\beta$.

Ordinary LS. $\hat{\beta} = \text{argmin}_\beta\, ||y - X\beta||^2 = (X^TX)^{-1}X^Ty$.
$\hat{\beta} \sim (\beta, \sigma^2(X^TX)^{-1})$ is BLUE for β (Gauss-Markov)
'hat' matrix $H = X(X^TX)^{-1}X^T$ is projection onto C(X)
SLS: $y = \beta_0 + \beta_1 x \Rightarrow \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$,

Data projection $\hat{y} = Hy$ gives unique least squares fit:
$||y - Hy|| \leq ||y - z||\ \ \forall z \in C(X)$, w/ equality iff $z = Hy$
Residuals $\hat{e} = y - \hat{y} = (I - H)y \in C(X)^\perp \Rightarrow \hat{e}^T\hat{y} = 0$
true and sample resid: $||y - \mu||^2 = ||\hat{e}||^2 + ||\hat{y} - \mu||^2$
data = fit + residuals: $y = Hy + (I - H)y$
$\Rightarrow y^Ty = y^THy + y^T(I - H)y \ \Rightarrow\ ||y||^2 = ||\hat{y}||^2 + ||\hat{e}||^2$

$\hat{\sigma}^2 :=: s^2 := \frac{1}{N-p}\sum_{i=1}^N (y_i - \hat{y}_i)^2 = \hat{e}^T\hat{e}/(N-p)$
$= y^T(I - H)y/(N - p)$, an unbiased estimate of σ^2.

$\hat{y}$ maximizes $\text{corr}(y, \hat{y}) = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}}$
$= \sqrt{\frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}} = \sqrt{\frac{SSR}{TSS}} = \sqrt{\frac{TSS - SSE}{TSS}} = +\sqrt{R^2}$

$R^2_{adj} = 1 - \frac{SSE/(n-p)}{TSS/(n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2)$.

$\text{Var}(\hat{y}) = \sigma^2 H$; $\text{Var}(\hat{e}) = \sigma^2(I - H) \Rightarrow \text{Var}(\hat{y}_i) = \sigma^2 h_i$; $\text{Var}(\hat{e}_i) = \sigma^2(1 - h_i)$, where $h_i = H_{ii}$ is leverage of i^{th} ob. Note $0 \leq h_i \leq 1$, and $\sum_i h_i = \text{Trace}(H) = \text{Rank}(H) = p$.

Standardized residuals: $r_i = \hat{e}/\hat{\sigma}\sqrt{1 - h_i}$. Want $|r_i| \leq 3$.
Influential pts have high leverage and large r_i.

Cook's Distance. $D_i = p^{-1}(\hat{\beta}_{-i} - \hat{\beta})^T[\text{Var}(\hat{\beta})]^{-1}(\hat{\beta}_{-i} - \hat{\beta})$
$= r_i^2 h_i/p(1 - h_i) = (y_i - \hat{y}_i)^2 h_i/p\hat{\sigma}^2(1 - h_i)$.

Generalized LS $\text{Var}(\varepsilon) = \sigma^2 V$, where V known.
Assume V full rank with spectral decomp $V = Q\Lambda Q^T$. Let $y^* = V^{-1/2}y$, $X^* = V^{-1/2}X$, where $V^{-1/2} = Q\Lambda^{-1/2}Q^T$. Then $\mathbb{E}y^* = X^*\beta$, $\text{Var}(y^*) = \sigma^2 I$, and we have $\hat{\beta} = (X^{*T}X^*)^{-1}X^{*T}y^* = (X^TV^{-1}X)^{-1}X^TV^{-1}y$.

Normal Linear Model Assume $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.
$\Rightarrow \hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$ and $\hat{y} \sim \mathcal{N}(X\beta, \sigma^2 H)$.

Cochran's Theorem. Suppose $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$ and $P_1, ..., P_K$ projection matrices satisfying $\sum_k P_k = I$. Then the quadratic forms $Q_k = (y - \mu)^T P_k(y - \mu)$, $k = 1, ..., K$, are independent with $Q_k \sim \sigma^2\chi^2_{r_k}$, where $r_k = \text{Rank}(P_k)$ and $\sum_k r_k = n$.

$(n - p)\hat{\sigma}^2 = \hat{e}^T\hat{e} = (y - X\beta)^T(I - H)(y - X\beta) \sim \sigma^2\chi^2_{n-p}$.

Note: $\chi^2_p \sim \Gamma(\nu = p/2, \mu = p)$. For $y \sim \mathcal{N}(\mu, V)$ with V nonsingular with rank p, $(y - \mu)^TV^{-1}(y - \mu) \sim \chi^2_p$.

$z/\sqrt{x/p} \sim t_p$ where $z \sim \mathcal{N}(0, 1)$ indep of $x \sim \chi^2_p$.

$\frac{x/p}{y/q} \sim F_{p,q}$ where $x \sim \chi^2_p$ indep of $y \sim \chi^2_q$.

$[SSR/(p-1)]/[SSE/(n-p)] \sim F_{p-1,n-p}$

Estimate $\tau = c^T\beta$ (non-random) with $\hat{\tau} = c^T\hat{\beta}$.
$\Rightarrow (\hat{\tau} - \tau)/\sqrt{\sigma^2 c^T(X^TX)^{-1}c} \sim \mathcal{N}(0, 1)$
$\Rightarrow (\hat{\tau} - \tau)/\sqrt{\hat{\sigma}^2 c^T(X^TX)^{-1}c} \sim t_{n-p}$, a valid t-dist since $\hat{\beta}$, $\hat{e}$ jointly normal as linear functions of y, and uncorrelated.

Predict future y_0 (random), at x_0 fixed with $\hat{y}_0 = x_0^T\hat{\beta}$.
$y_0 = \hat{y}_0 + e$ where $\mathbb{E}(e) = 0$, and by indep of obs,
$\text{Var}(e) = \text{Var}(y_0) + \text{Var}(\hat{y}_0) = \sigma^2(1 + x_0^T(X^TX)^{-1}x_0)$.
$\Rightarrow (y_0 - \hat{y}_0)/\hat{\sigma}^2\sqrt{1 + x_0^T(X^TX)^{-1}x_0} \sim t_{n-p}$.

One-Way ANOVA $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
treatment levels $i = 1, ..., I$; $j = 1, ..., n_i$; $N = \sum_i n_i$.
Assume $\varepsilon_{ij} \sim (0, \sigma^2 I)$. Let $\mu_i := \mu + \alpha_i = \mathbb{E}(Y_{ij})$.

Matrix form: $Y = X\beta + \varepsilon$, $\beta = (\mu, \alpha_1, ..., \alpha_I)^T$
$X_{N\times(I+1)}$ has rank I so β not identifiable.

Goal: minimize $\sum_{i=1}^I \sum_{j=1}^{n_i}(Y_{ij} - \mu - \alpha_i)^2$
Need constraint on α_i's for unique solution.
However, μ_i's identifiable with $\hat{\mu}_i = \bar{Y}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$
$SSE = \sum_{i=1}^I \sum_{j=1}^{n_i}(Y_{ij} - \hat{\mu}_i)^2$ is minimizer.
$\hat{\sigma}^2 = SSE/(I - 1)$ is unbiased for σ^2.
$\hat{\mu} = \bar{Y}_{\cdot\cdot}$ overall mean, $TSS = \sum_i \sum_j (Y_{ij} - \hat{\mu})^2$, and
$SSR = TSS - SSE = \sum_i n_i(\hat{\mu}_i - \hat{\mu})^2$ is SS reduction due to treatment.
If ε_{ij} i.i.d. $\mathcal{N}(0, \sigma^2 I)$, then $\frac{SSR/(I-1)}{SSE/(N-I)} \sim F_{I-1,N-I}$.

Contrasts: $\theta = \sum_i c_i\alpha_i$ is a contrast if $\sum_i c_i = 0$.
$\theta = \sum_i c_i\alpha_i$ is estimable if and only if θ is a contrast. Recall θ estimable if $\exists\hat{\theta} = \sum_i \sum_j c_{ij}Y_{ij}$ s.t. $\mathbb{E}\hat{\theta} = \theta$.

For $\theta = \sum_i c_i\alpha_i$ a contrast, $\hat{\theta} = \sum_i c_i\hat{\mu}_i$ is BLUE.
If ε_{ij} i.i.d. $\mathcal{N}(0, \sigma^2 I)$, $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2\sum_i (c_i^2/n_i))$; $\hat{\sigma}^2 \perp\!\!\!\perp \hat{\theta}$.
$\Rightarrow T_\theta := (\theta - \hat{\theta})/\sqrt{\hat{\sigma}^2\sum_i (c_i^2/n_i)} \sim t_{N-I}$ and
$100(1-\alpha)\%$ given by $\hat{\theta} \pm t_{N-I,1-\alpha/2} \cdot \hat{\sigma}\sqrt{\sum_i (c_i^2/n_i)}$

Multiple Comparisons: Given contrasts θ_j, $j = 1, ..., m$ and α-level, want CI's S_j s.t. $\mathbb{P}(\cap_j^m\{\theta_j \in S_j\}) \geq 1 - \alpha$.

Simple, conservative approach for small m:
Bonferroni's Inequality: $\mathbb{P}(\cap_j^n E_j) \geq \sum_j^n \mathbb{P}(E_j) - (n - 1)$. $\Rightarrow$ construct each S_j s.t. $\mathbb{P}(\theta_j \in S_j) = 1 - \alpha/m$.

Scheffe's Method: SCI for all contrasts θ, for larger m.
$\sup_\theta T^2_\theta = \frac{1}{\hat{\sigma}^2}\sum_i n_i[(\hat{\mu}_i - \hat{\mu}) - (\alpha_i - \bar{\alpha})] \sim (I - 1)F_{I-1,N-I}$,
$\hat{\alpha} = \frac{\sum_i n_i\alpha_i}{N} \Rightarrow \hat{\theta} \pm \sqrt{(I-1)F_{I-1,N-I,1-\alpha}} \cdot \hat{\sigma}\sqrt{\sum_i (c_i^2/n_i)}$

Two-Way ANOVA *Assuming balanced design*

Main Effects: $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$, $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2 I)$

Treatments $i = 1, ..., I$; Blocks $j = 1, ...J$; $k = 1, ..., K$.

$LSE(\mu, \alpha, \beta) = \text{argmin}_{\mu,\alpha,\beta} \sum_{ijk}(Y_{ijk} - \mu - \alpha_i - \beta_j)^2 \Rightarrow$
Normal Eqs: $\bar{Y}_{i\cdot\cdot} - \mu - \alpha_i - \bar{\beta} = 0$; $\bar{Y}_{\cdot j\cdot} - \mu - \bar{\alpha} - \beta_j = 0$;
$\bar{Y}_{\cdots} - \mu - \bar{\alpha} - \bar{\beta} = 0 \Rightarrow$ need 2 constraints $\Rightarrow \bar{\alpha} = \bar{\beta} = 0$
$\Rightarrow \hat{\mu} = \bar{Y}_{\cdots}$; $\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \hat{\mu}$; $\hat{\beta}_j = \bar{Y}_{\cdot j\cdot} - \hat{\mu}$

$SSE_M = \sum_{ijk}(Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 \sim \sigma^2 \chi^2_{N-I-J+1}$
$\hat{\sigma}^2 = SSE_M/(N - I - J + 1)$

Under $H_0^A : \alpha_1 = ... = \alpha_I$, model becomes one-way, with just factor B (βs) $\Rightarrow SSE_0 = \sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdot j\cdot})^2$
SS Reduc due to A: $SSA = SSE_0 - SSE_M = JK \sum_i \hat{\alpha}_i^2$.
Also $SSB = IK \sum_j \hat{\beta}_j^2$ and $TSS = SSB + SSA + SSE_M$.
Proj decomp: $I - P_0 = (P_B - P_0) + (P_M - P_B) + (I - P_M)$.
df decomp: $N - 1 = (J - 1) + (I - 1) + (N - I - J + 1)$.
Balanced design means factors A and B orthogonal.

Main Effects contrast $\theta = \sum_i c_i \alpha_i$ with $\sum_i c_i = 0$.
Est $\hat{\theta} = \sum_i c_i \hat{\alpha}_i = \sum_i \bar{Y}_{i\cdot\cdot}$ with $\text{Var}(\hat{\theta}) = \frac{\sigma^2}{JK} \sum_i c_i^2$.

Complete Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$
Constraints $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$.
$\hat{\gamma}_{ij} = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots} = \bar{Y}_{ij\cdot} - \hat{\alpha}_i - \hat{\beta}_j + \hat{\mu}$
$\hat{\mu}, \hat{\beta}_j, \hat{\alpha}_i$ same $\Rightarrow SSE_C = \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2 \sim \sigma^2 \chi^2_{N-IJ}$
$SS_{AB} = SSE_M - SSE_C = K \sum_{ij} \hat{\gamma}_{ij}^2 \sim \sigma^2 \chi^2_{IJ-I-J+1}$.

Let $\tau_{ij} = \alpha_i + \beta_j + \gamma_{ij} \Rightarrow$ one-way ANOVA w/ IJ levels.
$\theta = \sum_{ij} d_{ij} \tau_{ij}$ w/ $\sum_{ij} d_{ij} = 0 \Rightarrow \hat{\theta} = \sum_{ij} d_{ij} \bar{Y}_{ij\cdot}$ BLUE

ESL ch 3

Successive orthogonalization of the cols of $X_{n\times p} \Rightarrow$ **QR** decomp: $X = QR$, where $Q_{n\times p}$ orthogonal w/ $Q^TQ = I$ and $R_{p\times p}$ upper triangular. We can use this to rewrite LS solutions: $\hat{\beta} = R^{-1}Q^Ty$ and $\hat{y} = QQ^Ty$.

Looking **Beyond LS**: (1) *Prediction accuracy* - LS est often have low bias but high variance $\Rightarrow$ Shrink coeffs to $\downarrow$variance at the expense of some bias. (2) *Interpretability* - find smaller subset of preds w/ strongest effects.

Model Selection approaches oft use training set to fit seq of models indexed by complexity param c_p. Choose smallest c_p to minimize est of EPE, eg CV or AIC.

Best Subset minimizes RSS for each subset size $k < p$.
Forward/backward stepwise selection greedy algorithms

Binary inc/exc $\Leftrightarrow$ Var. **Shrinkage** methods more cts.

Ridge Regression. min $RSS(\lambda) = ||y - X\beta||_2^2 + \lambda||\beta||_2^2$
$\Rightarrow \hat{\beta}^{\text{ridge}} = (X^TX + \lambda I)^{-1}X^Ty$, where inv always exists
Can alleviate unstable coef est for correlated predictors

$\hat{\beta}^{\text{ridge}}$ *not* equivariant under input scaling $\Rightarrow$ Standardize and center x_j's, take $\hat{\beta}_0 = \bar{y}$ and solve $\hat{\beta}^{\text{ridge}}$ for centered y w/o intercept. Then $\hat{y} = \bar{y} + X\hat{\beta}^{\text{ridge}}$.

Given SVD $X = UDV^T$, we have $\beta_{LS} = UU^Ty$, and
$X\hat{\beta}^{\text{ridge}} = \sum_{j=1}^p u_j \frac{d_j^2}{d_j^2+\lambda} u_j^Ty$. $\text{df}(\lambda) = \sum_{j=1}^p \frac{d_j^2}{d_j^2+\lambda}$
v_1 is first principal component direction.
$z_1 = Xv_1 = d_1u_1$ is first principal component, z_1 has largest sample variance, d_1^2/n, for normalized $z \in C(X)$

Principal Component Regression. Perform LS reg on first $k < p$ principal components. We implicitly assume y varies most in directions of high var of inputs

$\hat{\beta}_j^{\text{pcr}} = \frac{\langle z_j, y\rangle}{\langle z_j, z_j\rangle} = \frac{\langle u_j, y\rangle}{d_j}$; $\hat{y} = \bar{y} + \sum_{j=1}^M u_j u_j^T y$.

Lasso. $\hat{\beta}^{\text{lasso}} = \text{argmin}_\beta ||y - X\beta||_2^2$, subject to $|\beta|_1 \le t$.
Equivalently: $\hat{\beta}^{\text{lasso}} = \text{argmin}_\beta\{\frac{1}{2}||y - X\beta||_2^2 + \lambda|\beta|_1\}$

When $X^TX = I$, $\hat{\beta}_j^{\text{lasso}} = \text{sign}(x_j^Ty)(|x_j^Ty| - \lambda)_+$
Here lasso shrinks the least square solution by λ (or to 0).

The df is $\approx$ num of 'active' covariates, with non-zero coefs.

Under certain conditions the lasso identifies correct predictors with high prob. But nonzero lasso coef estimates are biased towards 0 and not consistent. Modified versions SCAD and adaptive lasso modify penalty so larger coefs shrunk less.

Least Angle Regression. (1) Select covariate most correlated w/ residual and add it to the active set $\mathcal{A}$.
(2) Move $\hat{\beta}_{\mathcal{A}}$ continuously towards OLS solution, until a new covariate is equally correlated with the residual. Add new covariate and repeat. $\min(N, p-1)$ steps

Residual at k^{th} "step": $r_k = y - X_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$
Direction of the step: $\delta_k = (X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1} X_{\mathcal{A}_k} r_k$
Coefficients within step: $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha\delta_k$
Effective df at end of kth step is k.

Partial Least Squares. Use y (unlike PCR) to construct LCs of inputs for reg. Standardize x, set $\hat{y}^{(0)} = \bar{y}$, $x_j^{(0)} = x_j$.
mth step: $z_m = \sum_j \langle x_j^{(m-1)}, y\rangle x_j^{(m-1)}$. $\hat{\theta}_m = \frac{\langle z_m, y\rangle}{\langle z_m, z_m\rangle}$.
$\Rightarrow \hat{y}^{(m)} = y^{(m-1)} + \hat{\theta}_m z_m$; $x_j^{(m)} = x_j^{(m-1)} - \text{proj}_{z_m}(x_j^{(m-1)})$
Solution path is nonlinear function of y. PLS seeks directions that have high variance *and* have high correlation with y.

For minimizing prediction error, ridge regression generally preferable to subset selection, PCR and PLS (w/ only slight improvement over latter two).

Forward Stagewise - FS, FS_ϵ, FS_0**:** Standardize predictors and start with residual $r = y$, $\hat{\beta} = 0$.

FS: At k^{th} step find x_j most correlated with current resid r
Update $\hat{\beta}_j \leftarrow \hat{\beta}_j + \frac{\langle x_j, r\rangle}{\langle x_j, x_j\rangle}$, leaving other β's unchanged.
Cont 'slow fitting' until correlations between r and x_j's all 0

FS_ϵ: As above find x_j most correlated with r. Update $\hat{\beta}_j \leftarrow \hat{\beta}_j + \delta_j$ where $\delta_j = \epsilon \cdot \text{sign}(\langle x_j, y\rangle)$. Set $r \leftarrow r - \delta_j x_j$.

FS_0: let $\epsilon \to 0$. Linear solution paths similar to LAR except coefs cannot move in opposite direction to their correlation.

4. LINEAR CLASSIFICATION

We have a linear decision boundary whenever some monotone transformation of $\delta_k(x)$ or $\mathbb{P}(G = k|X = x)$ is linear in x.

LDA/QDA: Use *discriminant* functions $\delta_k(x)$ for classification: $\hat{G}(x) = \text{argmax}_k\, \delta_k(x)$. Model classes as MV Gaussian:
$f_k(x) = 1/[(2\pi)^{p/2}|\Sigma_k|^{1/2}]\text{exp}\{-\frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k)\}$
$\mathbb{P}(G = k|X = x) = f_k(x)\pi_k / \sum_l f_l(x)\pi_l$ (Bayes w priors π_k)

LDA: $f_k(x) \sim \mathcal{N}(\mu_k, \Sigma)$, common cov Σ across classes.
$\Rightarrow \delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$.

In practice, need to estimate $(K-1) \times (p+1)$ parameters:
$\hat{\pi}_k = N_k/N$ $\qquad$ $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$
$\hat{\Sigma} = \sum_{k=1}^K \sum_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N - K)$

QDA: $f_k(x) \sim \mathcal{N}(\mu_k, \Sigma_k)$; $(K-1) \times \{p(p+3)/2+1\}$ params.
$\Rightarrow \delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k) + \log\pi_k$.

Regularized Discriminant Analysis: $\hat{\Sigma}(\alpha) = \alpha\hat{\Sigma}_k + (1 - \alpha)\hat{\Sigma}$.
can also shrink $\hat{\Sigma}$ toward scalar cov: $\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1 - \gamma)\hat{\sigma}^2 I$

LDA projects data onto centroid subspace H_{K-1}. Want subspace $H_L \subset H_{K-1}$ of dim $L < (K-1)$ where projected centroids are maximally spread out in terms of variance.

B - cov between centroids $(\mu_1, ..., \mu_K)^T$. W - common, w/in class covariance. $B + W = T$, total cov of X, ignoring class.

Fisher: *'Find linear combination $Z = a^TX$ s.t. btw class var, a^TBa, is maximized relative to the w/in-class var, a^TWa.'*

$\Rightarrow \max_a(a^TBa/a^TWa) \Leftrightarrow \max_a a^TBa$ subject to $a^TWa = 1$.
Solution a^* is eigenvector for largest eigenvalue of $W^{-1}B$.

Optimal H_L spanned by *discriminant coordinates* $a_1, ..., a_L$.

LDA vs Logistic Regression: LDA maximizes the full log-likelihood based on the joint density $Pr(X, G = k) = Pr(X)Pr(G = k|X) = \phi(X; \mu_k, \Sigma)\pi_k$, where ϕ is the Gaussian density; whereas logistic regression maximizes the conditional likelihood.

Separating Hyperplanes. $\{x : \hat{f}(x) = x^T\hat{\beta} + \hat{\beta}_0 = 0\}$.
Find the plane that minimizes the distance of misclassification points to the boundary. The classification is given by $G = \text{sign}(\hat{f}(x_i))$.
Rosenblatt's Perceptron's Algorithm. Minimize $D(\beta) = -\sum_{i\in\mathcal{M}} y_i(x_i^T\beta)$, where $\mathcal{M}$ is the set of misclassified points. Search solution with *stochastic gradient descent*, and update one observation at a time: $\beta \leftarrow \beta + y_iz_i$, where z_i is normalized x_i. Can show this converges in at most $||\beta_{\text{start}} - \beta_{\text{sep}}||^2$ steps, but there are multiple solutions, and which one gets picked depends on the initial guess.
Optimal separating hyperplane. Separates the two classes and maximizes the distance to the closest point from either class. The corresponding optimization problem is
$\min\frac{1}{2}||\beta||^2$ subject to $y_i(x_i^T\beta + \beta_0) \ge 1, i = 1, ..., N$. Can show the points at the boundary determine β.
Note: If the data is separable, logistic regression always finds a separating hyperplane.

5. BASIS EXPANSIONS

Idea. Construct *derived input features* $h_m(X) : \mathbb{R}^p \to \mathbb{R}$ and use the linear model $f(X) = \sum_{m=1}^M \beta_m h_m(X)$.

Order-M spline with K knots ξ_j, is a piecewise polynomial of order M (degree $M - 1$) with continuous derivatives up to order $M - 2$. $M + K$ df.

Truncated-power basis set: $h_j(X) = X^{j-1}, j = 1, \cdots, M$.
$h_{M+l}(X) = (X - \xi_l)_+^{M-1}, l = 1, \cdots, K$.

Boundary behavior of polynomial fit erratic, exacerbated by splines (less global smoothing).Trade boundary bias for $\downarrow$var

Natural cubic splines: Additional constraint that function be linear beyond the boundary knots. Frees up 4 df: $N_1(X) = 1$, $N_2(X) = X$, $N_{k+2}(X) = d_k(X) - d_{K-1}(X)$, $k = 1, ..., K-2$, where $d_k(X) = [(X - \xi_k)_+^3 - (X - \xi_K)_+^3]/(\xi_K - \xi_k)$.

Smoothing splines. Given smoothing param λ, find $f \in \mathcal{C}^2$ to minimize $RSS(f, \lambda) = \sum_{i=1}^N\{y_i - f(x_i)\}^2 + \lambda\int\{f''(t)\}^2dt$
Solution is a natural cubic spline with knots at the unique values of $x_i \Rightarrow f(x) = \sum_{j=1}^N N_j(x)\theta_j$, using the basis defined above. *Gives spline basis method that avoids knot selection.*

Let $\{\mathbf{N}\}_{ij} = N_j(x_i)$, $\{\mathbf{\Omega}_N\}_{jk} = \int N_j''(t)N_k''(t)dt$. $\Rightarrow$ need to

min $RSS(\theta,\lambda) = (\mathbf{y}-\mathbf{N}\theta)^T(\mathbf{y}-\mathbf{N}\theta) + \lambda\theta^T\mathbf{\Omega}_N\theta \Rightarrow$ Generalized ridge regression solution: $\hat{\theta} = (\mathbf{N}^T\mathbf{N}+\lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y}$.

$\hat{\mathbf{f}} = \mathbf{N}(\mathbf{N}^T\mathbf{N}+\lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y} = \mathbf{S}_\lambda\mathbf{y}$, linear in $\mathbf{y}$. $\mathbf{S}_\lambda$ is called the *smoother matrix*. Compare to hat matrix from LS.
$\cdot\mathbf{S}_\lambda$ symmetric, positive-semidefinite w/ $\text{Rank}(\mathbf{S}_\lambda) = N$.
$\cdot\mathbf{S}_\lambda$ nearly banded $\Rightarrow$ smoothing spline a local fitting method.
$\cdot\mathbf{S}_\lambda\mathbf{S}_\lambda \preceq \mathbf{S}_\lambda$, meaning that $\mathbf{S}_\lambda$ exceeds $\mathbf{S}_\lambda\mathbf{S}_\lambda$ by a PSD matrix. This is due to the shrinking nature of $\mathbf{S}_\lambda$.
$\cdot$*Reinsch form*: $\mathbf{S}_\lambda = (\mathbf{I}+\lambda\mathbf{K})^{-1}$, where penalty matrix $\mathbf{K}$ is free of λ. $\Rightarrow \hat{\mathbf{f}} = \text{argmin}_\mathbf{f}\{(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f}) + \lambda\mathbf{f}^T\mathbf{K}\mathbf{f}\}$.
$\cdot$Eigen decomp: $\mathbf{S}_\lambda = \sum_k \rho_k(\lambda)\mathbf{u}_k\mathbf{u}_k^T$, where $\rho_k(\lambda) = 1/(1+\lambda d_k)$, and d_k is corresponding eigenvalue of $\mathbf{K}$.
$\cdot \Rightarrow \hat{\mathbf{f}} = \sum_k \rho_k(\lambda)\mathbf{u}_k\langle\mathbf{u}_k,\mathbf{y}\rangle$, so $\mathbf{y}$ is decomposed w.r.t. $\{\mathbf{u}_k\}$ and differentially shrunk by $\rho_k(\lambda)$.
$\cdot\rho_1(\lambda) = \rho_2(\lambda) = 1, d_1 = d_2 = 0$ corresponds to 2-D eigenspace of functions N_1, N_2 linear in X which are never shrunk.
$\cdot\rho_k(\lambda) \to 0$ as $\lambda \to \infty$ for $k = 3, ..., N$
$\cdot\text{df}_\lambda = \text{Trace}(\mathbf{S}_\lambda) = \sum_k \rho_k(\lambda)$.

Multidimensional Splines. Given bases $h_{1k}(X_1)$ and $h_{2j}(X_2)$, define the tensor product basis as $g_{jk}(X) = h_{1k}(X_1)h_{2k}(X_2)$. Gives the estimator $g(X) = \sum_{j=1}^{M_1}\sum_{k=1}^{M_2}\theta_{jk}g_{jk}(X)$.

Multi-D Smoothing Splines. $\min_f \sum_{i=1}^N\{y_i - f(x_i)\}^2 + \lambda J[f]$. Solution is *thin-plate* spline: $f(x) = \beta_0 + \beta^T x + \sum_{j=1}^N \alpha_j h_j(x)$, where $h_j(x) = ||x-x_j||^2\log||x-x_j||$.

Wavelet bases can capture both smooth & locally bumpy functions - *time and freq localization.* Haar basis $\phi(x) = I(x \in [0,1])$. Father wavelet $\phi_{0,k}(x) = \phi(x-k), j \in \mathbb{Z}$ span V_0. Mother wavelet $\phi_{j,k} = 2^{j/2}\phi(2^j x - k)$ span V_j, $V_j \supset V_{j-1}$. Writing $V_{j+1} = V_j \oplus W_j$, then W_j represents the detail, e.g. W_0 is spanned by $\psi(x-k)$, where $\psi(x) = \phi(2x) - \phi(2x-1)$, the mother wavelet.

Symmlet-p wavelet. Has p vanishing moments, that is $\int \psi(x)x^j dx = 0, j = 0, \cdots, p-1$.

Adaptive Wavelet Filtering. "SURE" shrinkage popular: $\min_\theta ||\mathbf{y} - \mathbf{W}\theta||_2^2 + 2\lambda||\theta||_1$, same as lasso criterion. With $N = 2^J$ lattice-points, design $\mathbf{W}$ is orthogonal and $\hat{\theta}_j = \text{sign}(\mathbf{y}_j^*)(|\mathbf{y}_j^*| - \lambda)_+$ where $\mathbf{y}^* = \mathbf{W}^T\mathbf{y} \Rightarrow \hat{\mathbf{f}} = \mathbf{W}\hat{\theta}$.

6. KERNEL SMOOTHING METHODS

Idea. Compute $\hat{f}(x_0)$ by assigning weights to points (x_i, y_i) that die off smoothly with distance from x_0.

Nadaraya-Watson (NW) kernel-weighted average.
$\hat{f}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0,x_i)y_i}{\sum_{i=1}^N K(x_0,x_i)}$, like cts version of KNN, where $\hat{f}_{knn}(x_0) = \sum_i w_i y_i$, and $w_i = 1/k$ if $x_i \in N_k(x_0)$, else 0.

Epanechnikov quadratic kernel $K_\lambda(x_0,x) = D(|x-x_0|/\lambda)$ where $D(t) = 3/4(1-t^2)$ if $|t| \le 1$, else 0.

To accommodate ties in x_i, ave y_i's at ties & use weights w_i

Bias-variance tradeoff in choice of λ/width of kernel. Epanechnikov/tri-cube $\Rightarrow$ λ = radius of support. Gaussian $\Rightarrow \lambda = \sigma$. KNN $\Rightarrow \lambda = k$.

Local linear and polynomial regression. Kernel smoothing is biased at the boundaries. Local linear and polynomial (of degree d) regression remove the first-order and up to d^{th} order bias, smoothness allowing, and solve:
$\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^N K_\lambda(x_0,x_i)[y_i - \sum_{j=0}^d \beta_j(x_0)x_i^j]^2$

Let $b(x)^T = (1, x, x^2, ..., x^d)$. Let B be $N \times (d+1)$ regression matrix with i^{th} row $b(x_i)^T$, and $W(x_0)$ the $N \times N$ diagonal matrix with i^{th} diagonal element $K_\lambda(x_0,x_i)$. Then $\hat{f}(x_0) = b(x_0)^T(B^TW(x_0)B)^{-1}B^TW(x_0)y = \sum_{i=1}^N l_i(x_0)y_i$

Taylor expanding: $E\hat{f}(x_0) = \sum_{i=1}^N l_i(x_0)f(x_i)$
$= f(x_0)\sum_{i=1}^N l_i(x_0) + f'(x_0)\sum_{i=1}^N(x_i-x_0)l_i(x_0)$
$+\frac{1}{2}f''(x_0)\sum_{i=1}^N(x_i-x_0)^2 l_i(x_0) + R$

$\sum_{i=1}^N l_i(x_0) = 1$ and $\sum_{i=1}^N(x_i-x_0)^j l_i(x_0) = 0$ for $1 \le j \le p$
$\Rightarrow \text{Bias}\,\hat{f}(x_0) = \frac{f^{(d+1)}(x_0)}{(d+1)!}\sum_i(x_i-x_0)^{d+1}l_i(x_0) + ...$

Pay price in variance for increasing d, particularly at boundary. Assuming $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$, then
$\text{Var}\,\hat{f}(x_0) = \sigma^2||l(x_0)||$, where $||l(x_0)||$ increases with d.

Note: Local regression smoothers are linear: $\hat{f} = S_\lambda y$.

Local Regression generalizes naturally to $p > 1$ dimensions. Typically use radial kernel $K_\lambda(x_0,x) = D(||x-x_0||/\lambda)$ where $||\cdot||$ Euclidean distance norm.
But we run into curse of dimensionality. As $p \uparrow$, need exponential inc in N to simultaneously maintain *localness* (low bias) and a sizable sample in neighborhood (low var). Boundary effects amplify as fraction of pts near boundary $\to 1$.

Structured local regression. Pick a PSD matrix A to weight diff coordinates: $K_{\lambda,A}(x_0,x) = D\left(\frac{1}{\lambda}(x-x_0)^TA(x-x_0)\right)$.

Varying coefficients models. Pick $q < p$ predictors and let $Z = \{X_{q+1}, ..., X_p\}$. Assume conditionally linear model $f(X) = \alpha(Z) + \beta_1(Z)X_1 + ... + \beta_1(Z)X_q$, which is linear for given Z but coefs can change w/ Z. Fit by locally weighted LS: $\min_{\alpha,\beta}\sum_{i=1}^N K_\lambda(z_0,z_i)(y_i - \alpha(z_0) - \sum_{j=1}^q x_{ji}\beta_j(z_0))^2$

Kernel density estimation and classification.
KDE is an unsupervised learning procedure. Given random sample $x_1, ..., x_N \sim f_X(x), X \in \mathbb{R}$, want to estimate $f_X(x_0)$.
Natural local estimate $\hat{f}_\lambda(x_0) = \#\{i : d(x_i,x_0) \le \lambda\}/(N\lambda)$.
Prefer smooth *Parzen estimate* $\hat{f}_\lambda(x_0) = \frac{1}{N\lambda}\sum_i K_\lambda(x_0,x_i)$.
With Gaussian kernel $K_\lambda(x_0,x_i) = \phi(|x_i-x_0|/\lambda)$, $\hat{f}_\lambda(x_0) = \frac{1}{N}\sum_i \phi_\lambda(x_0-x_i) = (\hat{F} * \phi_\lambda)(x_0)$, where $\phi_\lambda \sim \mathcal{N}(0,\lambda)$.
We can use these estimates for classification. For J class problem, fit $f_j(x_0)$, $j = 1, ..., J$ separately, and est class priors $\hat{\pi}_j$ (N_j/N). $\Rightarrow \hat{\mathbb{P}}(G = j|X = x_0) = \hat{\pi}_j\hat{f}_j(x_0)/\sum_{k=1}^J \hat{\pi}_k\hat{f}_k(x_0)$

Naive Bayes classifier. Given a class $G = j$, assume the features X_k are independent: $f_j(X) = \prod_{k=1}^p f_{jk}(X_k)$.
Then f_{jk} can be estimated separately with a 1-d KDE and $\log\frac{\Pr(G=l|X)}{\Pr(G=J|X)} = \log\frac{\pi_l}{\pi_j} + \sum_{k=1}^p \log\frac{f_{lk}(X_k)}{f_{Jk}(X_k)}$.
NB can work well for high p, when joint density est unattractive. Indep assumption often untrue, yet still performs well. Despite biased class densities, posteriors may hold up near decision regions. Lower variance.

Radial basis functions and kernels. Treat kernel functions $K_\lambda(\xi,x)$ as basis. $f(x) = \sum_{j=1}^M K_{\lambda_j}(\xi_j,x)\beta_j$.
Often choose $\{\lambda_j,\xi\}$ in unsupervised way, using only X.

Mixture models. $f(x) = \sum_{m=1}^M \alpha_m\phi(x;\mu_m,\Sigma_m)$, a weighted ave of (here Gaussian) densities. Mixing props α_m.

7. MODEL ASSESSMENT AND SELECTION

Setup. $Y = f(X) + \varepsilon$, where $\varepsilon \sim (0,\sigma^2)$, indep of X.

Test error is prediction error over an independent test sample, $\text{Err}_\mathcal{T} = \mathbb{E}\{L(Y,\hat{f}(X)|\mathcal{T}\}$, where X, Y are drawn randomly from their joint dist, and the training set $\mathcal{T}$ is fixed.
EPE or expected test error $\text{Err} = \mathbb{E}L(Y,\hat{f}(X)) = \mathbb{E}(\text{Err}_\mathcal{T})$.

Bias-Variance Decomp. For sq error loss,
$\text{Err}(x_0) = \mathbb{E}\{(Y-\hat{f}(x_0))^2|X = x_0\} = \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2$
$+E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 = \text{IrreducibleError} + \text{Bias}^2 + \text{Variance}$

Further decomp of expected Bias^2 possible for linear model with $\beta_* = \text{argmin}_\beta\,\mathbb{E}(f(X) - X^T\beta)^2$, and $\hat{f}_\alpha(x) = x^T\hat{\beta}_\alpha$, where $\hat{\beta}_\alpha$ possibly restricted like ridge regression:
$\mathbb{E}_{x_0}\text{Bias}\,\hat{f}_\alpha(x_0) = E_{x_0}[f(x_0) - x_0^T\beta_*]^2 + E_{x_0}[x_0^T\beta_* - Ex_0^T\hat{\beta}_\alpha]^2$
$= \text{Ave[Model Bias]}^2 + \text{Ave[Estimation Bias]}^2$
where model bias is between best linear fit and true f; estimation bias between est and best linear fit (0 for OLS).

Training error: $\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^N L(y_i,\hat{f}(x_i))$.

In-sample error: $\text{Err}_{\text{in}} = \frac{1}{N}\sum_i E_{Y^0}\{L(Y_i^0,\hat{f}(x_i))|\mathcal{T}\}$, where Y_i^0 a new response value at each of the training pts x_i

Optimism: $\text{op} := \text{Err}_{\text{in}} - \overline{\text{err}}$, and average optimism $\omega := E_y(\text{op})$, expectation over the training set outcome values, with predictors fixed. In general, $\omega = \frac{2}{N}\sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i)$.

Close connection to *effective degrees of freedom,* $\text{df}(\hat{y}) = (1/\sigma_\varepsilon^2)\sum_i \text{Cov}(\hat{y}_i, y_i)$. For linear fit $\hat{y} = Sy$, $\text{df}(\hat{y}) = \text{tr}(S)$.

For d-param linear fit under squared error loss, in-sample error est: $\hat{\text{Err}}_{\text{in}} = \overline{\text{err}} + \hat{\omega}$ given by $C_p = \overline{\text{err}} + 2(d/N)\hat{\sigma}_\epsilon^2$.

AIC $:= -2\text{loglik} + 2d$ gives est for $\hat{\text{Err}}_{\text{in}}$ for *log-likelihood loss* $L(y,\theta) = -2 \times \log p_\theta(y)$. Normal linear model $\Rightarrow$ AIC $\equiv C_p$

BIC $:= -2\text{loglik} + \log(N)d$ $(= (N/\sigma_\epsilon^2)\overline{\text{err}} + \log N\frac{d}{N}\sigma_\epsilon^2$ for normal linear model with known variance.)

Let data Z and models $\mathcal{M}_m$, $m = 1, ..., M$, with params θ_m. Use posterior prob of $\mathcal{M}_m$, $\mathbb{P}(\mathcal{M}_m|Z) \propto \mathbb{P}(\mathcal{M}_m)\mathbb{P}(Z|\mathcal{M}_m)$ to compare 2 models through their posterior odds $\frac{\mathbb{P}(\mathcal{M}_m|Z)}{\mathbb{P}(\mathcal{M}_\ell|Z)}$.

Now $\log\mathbb{P}(Z|\mathcal{M}_m) = \log\mathbb{P}(Z|\hat{\theta}_m\mathcal{M}_m) - (d_m/2)\log N + O(1)$, where $\hat{\theta}$ is MLE and d_m is the # of free params in $\mathcal{M}_m \Rightarrow$ $\text{BIC}_m = -2\log\mathbb{P}(Z|\hat{\theta}_m\mathcal{M}_m) + d_m\log N \approx -\log\mathbb{P}(Z|\mathcal{M}_m)$.
$\Rightarrow$ assuming uniform prior $\mathbb{P}(\mathcal{M}_m) = 1/M$, estimate posterior prob $\mathbb{P}(Z|\mathcal{M}_m)$ with $\exp\{-\frac{1}{2}\text{BIC}_m\}/\sum_\ell\exp\{-\frac{1}{2}\text{BIC}_\ell\}$.

Cross-Validation. $\text{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^N L(y_i,\hat{f}^{-\kappa(i)}(x_i))$, where $\kappa : \{1, ..., N\} \to \{1, .., K\}$ partitions sample into K folds. Let L_k ave loss in kth fold, and $\bar{L} = K^{-1}\sum_k \bar{L}_k$. then $\text{sd}(L)/\sqrt{K} = \sqrt{\sum_k(L_k-\bar{L})^2/K}$ gives SE for CV est of Err.

Generalized Cross-Validation. A convenient approx for leave-one-out CV, for linear fitting ($\hat{f} = Sy$) under sq-error loss. For many linear methods (OLS, smoothing splines), the exact form is $\frac{1}{N}\sum_{i=1}^N[y_i - \hat{f}^{-i}(x_i)]^2 = \frac{1}{N}\sum_{i=1}^N\left[\frac{y_i-\hat{f}(x_i)}{1-S_{ii}}\right]$ The GCV approx uses $\text{trace}(S)/N$ instead of the S_{ii}.

The Bootstrap is a general tool for assessing statistical accuracy. For a training set Z, draw B data sets from Z with replacement. Can estimate prop of a statistic $S(Z)$, like variance $\hat{\text{Var}}[S(Z)] = \frac{1}{B-1}\sum_b(S(Z^{*b}) - \bar{S}^*)^2$. *Estimate of prediction error:* $\hat{\text{Err}}_{\text{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^B\sum_{i=1}^N L(y_i,\hat{f}^{*b}(x_i))$. Optimistic, as bootstrap samples overlap with training set, which we here use as a test set.

$\mathbb{P}(\text{ob } i \in b\text{th sample}) = 1 - \left(1-\frac{1}{N}\right)^N \approx 1 - e^{-1} = 0.632$

Taking C^{-i} the bootstrap sample indices which do not contain ob i, $\hat{\text{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in c^{-i}} L(y_i, \hat{f}^{*b}(x_i))$
Pessimistic in low data regime, where model learning curve is steep. $\hat{\text{Err}}^{(.632)} = .368\overline{\text{err}} + .632\hat{\text{Err}}^{(1)}$ is middle ground approach, but this breaks down as $\overline{\text{err}} \downarrow 0$ (extreme overfitting)

8. GENERALIZED LINEAR MODELS

Random Component y from expo family:
$y \sim p(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}$
Monotone, differentiable *Link Function* g connects random component to linear predictor $\eta = X\beta$. $g(\mu) = \eta$, where $\mu = \text{E}(Y)$. Inverse response function $g^{-1}(\eta) = \mu$.

Let $l_i = \log p(y_i;\theta,\phi)$, contribution of y_i to $l = \text{loglik} = \sum_i l_i$.
Bartlett eqs: $\mathbb{E}(\partial l/\partial\theta) = 0$ and $-\mathbb{E}(\partial^2 l/\partial\theta^2) = \mathbb{E}[(\partial l/\partial\theta)^2]$.

Then $\mathbb{E}y = b'(\theta)$ and $\text{Var}(y) = a(\phi)b''(\theta)$.
Often $a(\phi) = \phi$ and we define $V(\mu) = \text{Var(y)}/\phi$.

With *canonical link* $g(\mu) = \theta = X\beta$, $X^T y$ is sufficient for β.

Likelihood eqs $\partial l/\partial\beta = \sum_i \partial l_i/\partial\beta = 0$. For $j = 1, ..., p$,
$\partial l/\partial\beta_j = \sum_i \frac{\partial l_i}{\partial\beta_j} = \frac{\partial l_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} = \sum_i \frac{(y-\mu_i)x_{ij}}{\text{Var}(y_i)}\frac{\partial\mu_i}{\partial\eta_i} \Rightarrow$
$XDV^{-1}(y-\mu) = 0$ for $D = \text{diag}(\partial\mu_i/\partial\eta_i)$, $V = \text{diag}\,\text{Var}(y_i)$, nonlinear functions of β (through $\mu = g^{-1}(X\beta)$).

$\mathcal{I}(\beta) = \mathbb{E}\frac{-\partial^2 l}{\partial\beta_j\beta_k} = \sum_i \frac{x_{ij}x_{ik}}{\text{Var}(y_i)}(\frac{\partial\mu_i}{\partial\eta_i})^2 = X^T W X$, where $W = \text{diag}([\partial\mu_i/\partial\eta_i]^2/\text{Var}(y_i))$. MLE $\hat{\beta} \overset{asy}{\sim} \mathcal{N}(\beta, (X^T W X)^{-1})$.

If the hessian H_f is negative-definite, there is a unique maximizer for l. $(H_f)_{ij} = \partial^2 f/\partial x_i \partial x_j$
If l is concave and has a unique MLE $\hat{\mu}$, then $X^T y = X^T\hat{\mu}$.

Taylor Expansion. $f(x) = f(x_0) + \nabla_f^T(x_0)(x-x_0) + \frac{1}{2}(x-x_0)^T H_f(x_0)(x-x_0)$.

Newton-Raphson. $\theta^{(k+1)} = \theta^{(k)} - [H_f(\theta^{(k)})]^{-1}\nabla_f(\theta^{(k)})$

Fisher-Scoring. Compute $\text{E}(H_l(\theta)) = -\mathcal{I}(\theta)$ where $\mathcal{I}$ is the Fisher information. Consider then the gradient of the log-likelihood $s(\theta) = \sum_i \nabla_{f_i}(Y_i;\theta)/f_i(Y_i;\theta)$, as a score function. Reasoning as above: $\theta^{(k+1)} = \theta^{(k)} + \mathcal{I}(\theta^{(k)})^{-1}s(\theta^{(k)})$

Iterative Re-weighted Least Squares.
(1) Start with initial guess β_0, and compute η_0 and μ_0.
(2) Compute adjusted response: $z_0 = \eta_0 + (y-\mu_0)g'(\mu_0)$.
(3) Compute the quadratic weight as $W_0^{-1} = (g'(\mu_0))^2 V_0$.
(4) Can then show the Fisher scoring step becomes: $X^T W_0 X\beta_1 = X^T W_0 z_0$, which the gradient of the RSS at a stationary point. Hence, β_1 solves a weighted least square problem.

Nonlinear Least-square. Minimize $||y - f(\beta)||^2$ when f is nonlinear. Do Taylor approximation around $\beta^(k)$ and iteratively solve linear least square until reaching convergence.

Fitting GLMs. Minimize $S = ||\sqrt{V_{(k)}^{-1}}(y-\mu(\beta))||^2$.

(1) Let $\widetilde{S}$ be the first-order Taylor approximation of S.
(2) Let $W_{ii}^{(k)} = [V(\mu_i^{(k)})g'(\nu_i^{(k)})^2]^{-1}$ and $z_i^{(k)} = g'(\mu^{(k)})(y_i - \mu_i^{(k)} + \eta_i^{(k)})$.
Then $\widetilde{S} = ||\sqrt{W^{(k)}}(z^{(k)} - X\beta)||^2$ and can obtain $\beta^{(k+1)}$.
Often $\hat{\beta}^{(k)} \xrightarrow{P} \hat{\beta}^{\text{MLE}}$.

MLE minimizes *deviance* $D(y,\hat{\mu}) = \phi\times 2[l(y;\phi,y) - l(\hat{\mu};\phi,y)]$.
χ^2 test. For two nested model $M_0 \subset M_1$, with p_1 and p_2 params, under M_0, $X^2_{1,0} = [D(y,\hat{\mu}_0) - D(y,\hat{\mu}_1)]/\phi \overset{d}{\approx} \chi^2_{p_1-p_0}$.

If we need to approximate ϕ under M_0, we can use $\frac{(D_{M_0} - D_{M_1})/(p-q)}{D_{M_1}/(n-p)} \overset{d}{\approx} F_{p-q,n-p}$, an approximation that can fail in non-Gaussian cases.

def (Pearson Residual). $r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$, which approximates a standard normal. An estimate for the dispersion parameter is $S^2 = \frac{\sum(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)(n-p)}$ where p is the rank of X.

def (Anscombe Residual). Let $A(z) = \int_a^z \frac{\text{d}\mu}{V^{1/3}(\mu)}$, where a is the lower limit of the range of μ. Then $r_{A_i} = \frac{A(y_i) - A(\hat{\mu}_i)}{V^{1/6}(\hat{\mu}_i)}$.

Deviance Residual Let $d_i = 2w_i\{y_i(\bar{\theta}_i - \hat{\theta}_i) - b(\bar{\theta}_i) + b(\hat{\theta}_i)\}$ and $r_{D_i} = \text{sign}(y-\hat{\mu}_i)\sqrt{(d_i)}$.

setup. Categorize population according to class variables. For example treatment exposure and disease outcome.

Def (Prospective Sampling). Number of patients who receive treatment X is fixed. Number of patients with response D is random. Estimate $P(D|X)$.

Def (Retrospective Sampling). Number of responders is fixed. Let Z denote the event "a patient has been selected". Estimate $P(D|Z,X)$ using Bayes rule.
Remark. Prospective and retrospective model have the same parameter β, when using the logit link.
Remark 2. Let $\pi_0 = P(Z=1|D)$ and $\pi_1 = P(Z=1|\bar{D})$. We can show $\hat{\beta} = \text{logit}(\hat{\pi}_0) - \text{logit}(\hat{\pi}_1)$ and $\text{Var}(\text{logit(Y)}) \approx \frac{1}{\text{EY}} + \frac{1}{n-\text{EY}}$, with the δ method.

def (Over-dispersion). Occurs when the variance of the response Y exceeds the model variance. For example, if Y is modeled as poisson, but $\text{Var}(Y) > \mu$. We could then estimate the over-dispersion parameter $\hat{\sigma}^2 = \frac{1}{n-p}\sum(y_i - \hat{\mu}_i)^2/\hat{\mu}_i$. Over-disperion can be caused by clustering or more generally a hierarchical data generating process.

Ordinal logistic regression. We have a soft measures, such as a scale from 1 to 10. Hence $z = j$ if $C_{j-1} < c \le C_j$.

Cumulative logit model. $\frac{P(z\le C_j)}{P(z>C_j)} = \frac{\pi_1+...+\pi_j}{\pi_{j+1}+...+\pi_J} = x^T\beta_j$

Propotional odds model. Assume only the intercept depends on j. $\log\frac{\pi_1+...+\pi_j}{\pi_{j+1}+...+\pi_J} = \beta_{0,j} + \beta_1 x_1 + ... + \beta_p x_p$.

Contingency table. Table with two (or more) categories. The probability of belonging to a cell is θ_{ij}. An estimate is $\hat{\theta}_{ij} = y_{ij}/n$.

Independence of categories. Often we want to test if two categories are independent, that is $H_0 : \theta_{ij} = \theta_{i.}\theta_{.j}$. Let $e_{ij} = n\hat{\theta}_{i.}\hat{\theta}_{.j}$ (theoretical prediction under the null). Then $X^2 = \sum_{i,j}\frac{(y_{ij}-e_{ij})^2)}{e_{ij}} \overset{\text{approx}}{\sim} \chi^2_{(I-1)(J-1)}$.

10. Advanced Topics.

General Additive Model. $\mathbb{E}(Y|X) = \alpha + f_1(X_1) + ... + f_p(X_p)$, a concept that applies to linear models and GLMs. If the model is not identifiable, use $\hat{\alpha} = \text{ave}(y_i)$, and $\sum f_i(X_i) = 0$

Regression Trees. Splits the data in regions and classifies according to $f(x) = \sum c_m I(x \in R_m)$, with $c_m = \text{Ave}(y_i|x_i \in R_m)$. At each stage, select splitting variable and node to minimize RSS.

Complexity criterion. We can construct sub-trees by collapsing internal (non-terminal nodes) nodes, i.e prune the tree. Then we pick a tree that, for a penalty α, minimizes $C_\alpha(T) = \sum_m N_m Q_m(T) + \alpha|T|$, where $N_m = \#\{x_i \in R_m\}$ and $Q_m(T) = \frac{1}{N_m}\sum_{x_i\in R_m}(y_i - \hat{c}_m)^2$.

Classification tree. In each region, compute $\hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i\in R_m} I(y_i = k)$, and classify according to the class k that maximizes $\hat{p}_k$.

Misclass. error. $\frac{1}{N_m}\sum_{i\in R_m} = I(y_i \ne k(m)) = 1 - \hat{p}_{mk(m)}$
Gini index. $\sum_k \hat{p}_{mk}(1-\hat{p}_{mk})$
Cross entropy. $-\sum_k \hat{p}_{mk}\log\hat{p}_{mk}$.

Remark. The Gini and cross entropy indices favor splits that create pure nodes, i.e. nodes with only one class.

MARS. Additive model with piecewise linear basis functions of the form $(x-t)_+, (t-x)_+$. The coefficients are estimated by standard linear regression.

Model building. a. Select a pair basis function, and take its product with all the functions currently in the model, $\mathcal{M}$.
b. Retain the product that most decreases the training error.
c. After the full model is built, remove, step by step, the term that causes the smallest increase in the RSS.
d. Select the model size using GCV. Approximate $S_{ii} \simeq r + cK$, where r is the number of linearly independent functions in the model, K is the number of nodes, and $c = 3$ (2 if we require the model to be additive).

Neural Networks. $z_m = \sigma(\alpha_{0m} + \alpha_m^T X)$, $T_k = \beta_0 + \beta_k^T z$ and $f_k(X) = g_k(T)$ where σ is the activation function (often a sigmoid).
Model fitting is done recursively using back propagation. With weights (α and β) near 0, the model is roughly linear; hence stopping the model fitting before reaching convergence shrinks the network towards the linear model.
Alternatively, we can use a penalized objective function.

Effective at capturing nonlinearity in the data and making predictions. Very difficult to interpret.

SVM. *Support Vector Classifier.* Extends separating hyperplanes to non-separable data. Let M be the margin width and ξ_i the distance of x_i inside the margin. Need: $\min_{\beta,\beta_0}\frac{1}{2}||\beta||^2 + C\sum_i \xi_i$ s.t. $\xi_i \ge 0$, $y_i(x_i^T\beta + \beta_0) \ge (1-\chi_i), \forall i$, where C controls how many points we allow to be misclassified. Solution: $\hat{\beta} = \sum_i \hat{\alpha}_i y_i x_i$, where $\hat{\alpha}_i$ is 0 for points not on the margin, C for points at the edge of the margin; else $0 < \alpha_i < C$.
SVC, unlike LDA, is robust to outliers.

Support Vector Machine. Same as SVC, but the data is transformed using M basis functions, h_m, and $h(x_i) = (h_1(x_i), h_2(x_i), ..., h_M(x_i))$.

11. HANDY RESULTS

Thm (Bayes rule). $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(B)/\mathbb{P}(B)$

Thm (Delta method). If $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$, then if $g'(\theta) \ne 0$, $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$.
If $g'(\theta) = 0$, $n(g(X_n) - g(\theta)) \xrightarrow{d} \frac{\sigma^2}{2}\chi_1^2 g''(\theta)$.

Total variance. $\text{Var}(Y) = \text{E}(\text{Var}(Y|X)) + \text{Var}(\text{E}(Y|X))$.

Multinomial$(\text{n}, \pi) \to p(y,\pi) = \frac{n!}{y_1!...y_J!}\pi_1^{y_1}...\pi_J^{y_J}$
The joint distribution of of k Poissons, with $Y_i \sim \text{Poisson}(\lambda_i)$, conditional on $\sum_i Y_i = n$, is a multinomial with parameters $(n, \pi_1, ..., \pi_k)$ where $\pi_i = \lambda_i/\sum_j^k \lambda_j$.