STAT6001

2° Linear Models

General Theory
- 2.0
- 2.1 Inference
  - 2.1.1 Estimation of params
  - 2.1.2 Est σ^2
  - 2.1.3 Weighted least squares
  - 2.1.4 CIs & tests

Multiple Regression
- 2.2 Multiple lin reg
  - 2.2.1 Basic results
  - 2.2.2 Interpretation
  - 2.2.3 Model adequacy

Robust Regression
- 2.3 Robust
  - 2.3.0
  - 2.3.1 M-estimation
  - 2.3.2 S-estimator
  - 2.3.3 MM-estimator

Model Selection
- 2.4 Var select
  - 2.4.1 t-tests, F-tests
  - 2.4.2 Best subset
  - 2.4.3 Criterion comp models
  - 2.4.4 Stepwise methods
  - 2.4.5 lasso

ANOVA
- 2.5 ANOVA

3° Generalised linear models

General Theory
- 3.1 Exp family
  - 3.1.1 Properties
- 3.2 GLM theory
  - 3.2.1 Estimation
  - 3.2.2 CIs for params
  - 3.2.3 Hypothesis test

Logistic Reg for Binary Data
- 3.3 Binomial data, logistic reg
  - 3.3.1 Param interpretation
  - 3.3.2 Basic results
  - 3.3.3 Model adequacy
  - 3.3.4 Model selection
  - 3.3.5 Analysis of deviance
- 3.4 Contingency tables

Log-Linear Models for Contingency Tables
  - 3.4.1 Background
  - 3.4.2 3-way
  - 3.4.3 Fitting models
  - 3.4.4 Goodness of fit, checks
- 3.5 GAMs
  - 3.5.1 Intro
  - 3.5.2 Structure
  - 3.5.3 Param est
  - 3.5.4 Inference

Assumptions: Linearity
Normality (e_i)
Homogeneity ($e_i \perp\!\!\!\perp x$)
Independence (e_i)

Linear Models

- $\underline{y} = \underline{\underline{X}}\underline{\beta} + \underline{e}$ $\quad \left(e_i \overset{iid}{\sim} N(0, \sigma^2),\ y_i \overset{ind}{\sim} N(\mu_i, \sigma^2),\ \mu_i = \underline{x}_i^T \underline{\beta} \right)$

- $S(\underline{\beta}) = \sum e_i^2 = \sum (y_i - \underline{x}_i^T \underline{\beta})^2 = (\underline{y} - \underline{\underline{X}}\underline{\beta})^T (\underline{y} - \underline{\underline{X}}\underline{\beta})$

Unbiased if y_i normal

- $\frac{\partial S}{\partial \underline{\beta}} = 2\underline{\underline{X}}^T\underline{\underline{X}}\underline{\beta} - 2\underline{\underline{X}}^T\underline{y} = 0 \Rightarrow \hat{\underline{\beta}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{y}$

- $\hat{\underline{\beta}} \sim N(\underline{\beta}, \sigma^2(\underline{\underline{X}}^T\underline{\underline{X}})^{-1})$

- $\hat{\underline{y}} = \underline{\underline{X}}\hat{\underline{\beta}} = \underline{\underline{H}}\underline{y} \Rightarrow \underline{\underline{H}} = \underline{\underline{X}}(\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T$ $\quad \left(\mathrm{tr}(\underline{\underline{H}}) = p,\ \underline{\underline{H}}^T\underline{\underline{H}} = \underline{\underline{H}},\ \underline{\underline{H}}^T = \underline{\underline{H}} \right)$

- $\hat{\underline{e}} = \underline{y} - \hat{\underline{y}} = (\underline{\underline{I}} - \underline{\underline{H}})\underline{y}$, $\quad \left(\underline{\underline{X}}^T\hat{\underline{e}} = 0 \ (\underline{\underline{X}}^T\underline{\underline{X}}\underline{\beta} - \underline{\underline{X}}^T\underline{y} = 0),\ \cancel{\sum e_i = 0} \right)$

% variance explained, low = don't trust model inferences

- $RSS = \hat{\underline{e}}^T\hat{\underline{e}}$, $E(\hat{\underline{e}}^T\underline{e}) = E(\underline{y}^T(\underline{\underline{I}} - \underline{\underline{H}})\underline{y}) = (N-p)\sigma^2$

$$\Rightarrow \hat{\sigma}^2 = \frac{RSS}{N-p}, \quad \left(\frac{RSS}{\sigma^2} \sim \chi^2_{N-p} \right)$$

- WLS: $e_i \sim N(0, V_{ii})$, $\underline{\underline{V}}$ diagonal

$$S(\underline{\beta}) = \sum w_i (y_i - \hat{y}_i), \quad w_i = \frac{1}{Var(y_i)}$$
$$= (\underline{y} - \underline{\underline{X}}\underline{\beta})^T \underline{\underline{V}}^{-1} (\underline{Y} - \underline{\underline{X}}\underline{\beta})$$

$$\Rightarrow \hat{\underline{\beta}} = (\underline{\underline{X}}^T\underline{\underline{V}}^{-1}\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{\underline{V}}^{-1}\underline{y}$$

$\left(E(\underline{\underline{A}}\underline{x}) = \underline{\underline{A}}E(\underline{x}),\ V(\underline{\underline{A}}\underline{x}) = \underline{\underline{A}}V(\underline{x})\underline{\underline{A}}^T \right)$

$$V(\hat{\underline{e}}) = V(\underline{y} - \hat{\underline{y}}) = V((\underline{\underline{I}} - \underline{\underline{H}})\underline{y}) = (\underline{\underline{I}} - \underline{\underline{H}})V(\underline{y})(\underline{\underline{I}} - \underline{\underline{H}})^T = \sigma^2(\underline{\underline{I}} - \underline{\underline{H}})$$

- $R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{RSS}{CTSS}$

~5% > 2

- $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$ $\quad \left(\hat{\underline{e}} \sim N(\underline{0}, \sigma^2(\underline{\underline{I}} - \underline{\underline{H}})) \right) \leftarrow (\Rightarrow \hat{e}_i \perp\!\!\!\perp \hat{e}_j, \text{ as } \sum e_i = 0)$

- Regularisation

$$\hat{\underline{\beta}}_{RR} = (\underline{\underline{X}}^T\underline{\underline{X}} + \lambda\underline{\underline{I}})^{-1}\underline{\underline{X}}^T\underline{y}, \quad E(\hat{\underline{\beta}}_{RR}) = \underline{\underline{X}}^T\underline{\underline{X}}(\lambda\underline{\underline{I}} + \underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\beta}$$

- Partial vs total regression coef

- t-test: $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$, $v_j = [(\underline{X}^T\underline{X})^{-1}]_{jj}$

$$\Rightarrow \frac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)} \sim t_{N-p} \qquad (se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_j})$$

$\hat{\beta}_j \pm t_{N-p,\frac{1}{2}\alpha}$ confidence interval $(1-\alpha)\%$

- Robust Regression:
  - (Also L1 & Least Median) → M-estimator: $S(\underline{\beta}) = \sum \rho\left(\frac{e_i(\underline{\beta})}{\sigma}\right)$ (ρ: +ve, symmetric, $\rho(0)=0$)
    - ($\psi = \rho'$)
    - (bounded ψ = bounded effect of large e_i)
  - analagous to WLS, except $\underline{W}$ not constant → $\hat{\underline{\beta}}_M^{(k+1)} = (\underline{X}^T \underline{W}^{(k)} \underline{X})^{-1} \underline{X}^T \underline{W} \underline{y}$, (iterate to convergence)

$$w_i^{(k)} = \frac{\psi(u_i^{(k)})}{u_i^{(k)}}$$

$$u_i^{(k)} = (y_i - \hat{y}_i^{(k)})/\hat{\sigma}^{(k)}$$

$$\hat{y}_i^{(k)} = \underline{x}_i^T \underline{\beta}^{(k)}$$

$$\hat{\sigma}_{MAD}^{(k)} \propto MED\{|\hat{r}_i - MED\{\hat{e}_j\}|\}$$

- Variable Selection
  - Backwards & Forwards elimination ($p = d+1$)

$$AIC = -2\hat{l}(model) + 2p = N \log \frac{RSS}{N} + 2p \qquad \left(l(model) = \frac{-N}{2}\log\sigma^2 - \frac{RSS}{2\sigma^2}, \ \hat{\sigma}^2 = \frac{RSS}{N}\right)$$

  - LOO-CV
  - Lasso (standardise $x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j}$)

Generalized Linear Models

- $\eta_i = \underline{x}_i^T \underline{\beta}$

link function → $g(\mu_i) = \eta_i$, $\mu_i = E(y_i)$, $y_i \sim \text{Expfam}(\theta, \phi)$

default for R →
$y_i \sim N(\mu_i, \sigma^2) \Rightarrow g(\mu) = \mu$
$y_i \sim \text{Poisson}(\mu_i) \Rightarrow g(\mu) = \log(\mu)$
$y_i \sim \text{Binomial}(n,p) \Rightarrow g(\mu) = \text{logit}(\mu) = \log \frac{\mu}{1-\mu}$

Canonical links, $g(\mu) = \theta$

$l(\underline{\beta}) = \sum \log P(y_i)$

$$\frac{\partial l}{\partial \beta_j} = \sum \left(\frac{y_i - \mu_i}{V(\mu_i)}\right) \frac{d\mu_i}{d\eta_i} x_{ij} = 0$$

Estimate $\underline{\beta}$ → Fisher ~~Newton~~ iterations → $\Rightarrow (\underline{X}^T \underline{W} \underline{X})^{(s-1)} \hat{\underline{\beta}}^{(s)} = (\underline{X}^T \underline{W} \underline{z})^{(s-1)}$

$W_{ii} = \frac{1}{V(\mu_i^{(s)})} \left(\frac{d\mu_i}{d\eta_i}\right)^2$
$z_i^{(s)} = \eta_i + (y_i - \mu_i^{(s)}) \left(\frac{d\eta_i}{d\mu_i^{(s)}}\right)$

Fisher Info Matrix → $\underline{\mathcal{I}}_{ij} = E\left(\frac{-\partial l}{\partial \beta_i} \cdot \frac{\partial l}{\partial \beta_j}\right)$, $\underline{\mathcal{I}} = \underline{X}^T \underline{W} \underline{X} / \phi$ (= ~~inverse of~~ negative Hessian)

$$\hat{\underline{\beta}} \sim N(\underline{\beta}, \underline{\mathcal{I}}^{-1}) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{se(\beta_j)} \overset{\text{approx}}{\sim} N(0,1)$$

- Deviance:

$$D^* = 2[l(\hat{\underline{\beta}}_{sat}) - l(\hat{\underline{\beta}})]\phi$$

($l(\hat{\underline{\beta}}_{sat})$ is likelihood when $\hat{\mu}_i = y_i$)

$$D^* = D/\phi \sim \chi^2_{N-p}$$

"likelihood test" of H_0 → $D_0^* - D^* \sim \chi^2_{p-q}$, ($D_0^*$ = standardized deviance under H_0: $p-q$ params all 0)

$$D^*_{M_1} - D^*_{M_2} \sim \chi^2_{p_1 - p_2}$$

Estimate ϕ → $\cdot\ X^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$, $X^2/\phi \sim \chi^2_{N-p} \Rightarrow \hat{\phi} = \frac{\hat{X}^2}{N-p}$

X^2 = D

Exponential Family

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

	θ	ϕ	$b(\theta)$	$c(y,\phi)$
Poisson (μ)	$\log\mu$	1	e^θ	$-\log y!$
Bin (n,π)	$\text{logit}\,\pi$	1	$n\log(1+e^\theta)$	$\log\binom{n}{y}$
$N(\mu,\sigma^2)$	μ	σ^2	$\frac{1}{2}\theta^2$	$-\frac{1}{2}(y^2/\phi + \log(2\pi\phi))$

For GLM, canonical link is $g(\mu) = \theta$

$E(Y) = b'(\theta)$, $\text{Var}(Y) = b''(\theta)\,a(\phi) = V(\mu)a(\phi)$

Generalized Additive Models

$\eta = \underline{x}^T\underline{\theta} + \sum_j f_j(x_j)$ (instead of $\eta = \underline{x}^T\underline{\beta}$ for GLM)

f_j are smooth functions, eg splines

Spline: weighted sum of basis functions, $f_j(x_j) = \sum \beta_{jk} b_{jk}(x_j)$

- $\underline{\lambda}$ = smoothing parameters; punish excessively wiggly fits
- estimate parameters w/ fixed $\underline{\lambda}$ by maximum likelihood

STATG001- Statistical Models & Data Analysis

Linear Models

$$Y_i = \vec{x}_i^T \vec{\beta} + e_i, \quad i = 1, \dots N, \quad \vec{x}_i \in D^m$$

- often assume $e_i \overset{iid}{\sim} N(0, \sigma^2)$ $(y_i \overset{iid}{\sim} N(x_i^T \beta, \sigma^2)$ $D = \mathbb{R} \Rightarrow$ multiple linear regression
- components of $\vec{x}$ can be transforms of the explanatory variables, eg $\vec{x}_i = \begin{pmatrix} 1 \\ u_i^2 \\ \log(u_i) \end{pmatrix}$

D = discrete $\Rightarrow$ analysis of variance model
D = mix $\Rightarrow$ general linear regression

- useful if model assumptions (linearity, normality) violated for original variables
- or multiplicative model $Z_i = u_{i1} u_{i2} \cdots u_{im} V_i$ (error), take logs $\Rightarrow$

$$\log Z_i = \log u_{i1} + \log u_{i2} + \cdots + \text{lots of } e_i$$

- or variable interactions, $\vec{x}_i = (1, u_i, v_i, u_i v_i)^T$
- or comparison of group means

$$Y_{ij} = \mu_i + e_{ij} \qquad i = 1, \dots I \leftarrow \#\text{groups}$$
$$= x_{ij1} \cdot \mu_1 + x_{ij2} \mu_2 + \cdots + e_{ij} \qquad j = 1, \dots n_i$$
$$Y_{ij} = \vec{x}_{ij}^T \vec{\mu} + e_{ij}$$

where ~~x_{ijk}~~ $x_{ijk} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$, $\vec{x}_{ij} \in \{0,1\}^I$

- Inference:
  - Parameter estimation:
    - Least squares:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, S(\vec{\beta}), \quad S(\vec{\beta}) = \sum_i e_i^2 = \sum_i (Y_i - \vec{x}_i^T \vec{\beta})^2$$
$$= (\vec{Y} - X\vec{\beta})^T (\vec{Y} - X\vec{\beta})$$

$$\Rightarrow X^T X \hat{\beta} = X^T \vec{Y}$$

if $p < N$, $X^T X$ invertible $\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T \vec{Y}$

- $E(\hat{\beta}) = \vec{\beta}$, $V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ (if e_i independent w/ var σ^2)

$\hat{\beta} \sim N_d(\vec{\beta}, \sigma^2 (X^T X)^{-1})$ if $e_i \overset{IID}{\sim} N(0, \sigma^2)$

- Maximum likelihood estimation

$$L(\vec{\beta},\sigma^2)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[\frac{-S(\vec{\beta})}{2\sigma^2}\right]$$

max of $L(\vec{\beta},\sigma^2)$ for given σ equivalent to min $S(\vec{\beta})$

- Estimation of σ^2
  - Residuals $\hat{e}=\vec{Y}-X\hat{\beta}$, $\hat{\beta}=(X^TX)^{-1}X^T\vec{Y}$
    $=(I_N-H)\vec{Y}$, $H=X(X^TX)^{-1}X^T$

Residual Sum of Squares → $RSS=\hat{e}^T\hat{e}$

$$E(RSS)=E(\hat{e}^T\hat{e})=E(\vec{Y}^T(I_N-H)\vec{Y})=(N-p)\sigma^2$$

$$\Rightarrow \underline{\hat{\sigma}^2=\frac{RSS}{N-p}} \quad \left(\text{if } e_i\sim N(0,\sigma^2), \text{IID} \Rightarrow \frac{RSS}{\sigma^2}\sim\chi^2_{N-p}\right)$$

- Weighted Least Squares Estimation
  - use when error variance not constant ← $e_i\sim N(0,V)$ (diagonal, ie independent errors)
  - minimise $S(\vec{\beta})=\sum w_i(Y_i-\vec{x}_i^T\vec{\beta})^2$ ($w_i\uparrow$ = more reliable measurements, e_i lower variance)

if $w_i=\frac{1}{Var(Y_i)}$, $S(\vec{\beta})=(\vec{Y}-X\vec{\beta})^TV^{-1}(\vec{Y}-X\vec{\beta})$ ← $w_i=\frac{1}{Var(Y_i)}$ ideally

  - correlated errors:
    - V no longer diagonal
    - normal equations: $X^TV^{-1}X\hat{\beta}=X^TV^{-1}Y$
    - problem: need to estimate V, see later

- Tests & Confidence intervals:
  - Regression params:

$e_i\sim N(0,\sigma^2)$

$\hat{\beta}_j\sim N(\beta_j,\sigma^2V_j)$ ← (j,j) element of $(X^TX)^{-1}$

$\frac{RSS}{\sigma^2}\sim\chi^2_{N-p}$

$\hat{\sigma}^2=\frac{RSS}{N-p}$

$$\Rightarrow t=\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2V_j}}\sim t_{N-p} \qquad se(\hat{\beta}_j)=\sqrt{\hat{\sigma}^2V_j}$$

$100(1-\alpha)\%$ CI $=\hat{\beta}_i\pm t_{N-p,\frac{1}{2}\alpha}\cdot se(\hat{\beta}_j)$

- Test hypothesis of the form $\beta_j = \beta^*$ for some β^*, eg $H_0: \beta_j = 0$

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim A_{N-p} \quad \text{under } H_0$$

- Linear combination of regression params:

$$\psi = \vec{c}^T \vec{\beta}, \quad \hat{\psi} = \vec{c}^T \hat{\beta}$$

$$\hat{\psi} \sim N(\psi, \sigma^2 v) \quad \text{where } v = \vec{c}^T (X^T X)^{-1} \vec{c}$$

$$\downarrow$$

$$\frac{\hat{\psi} - \psi}{\sqrt{\hat{\sigma}^2 v}} \sim t_{N-p}$$

- Tests about multiple params
  - Test subset $p-q$ params are all 0, leaving max q non-zero
  - H_1: at least 1 of $p-q$ params in H_0 isn't 0

RSS_0, RSS_1 are RSS under H_0, H_1

$$\frac{RSS_0 - RSS_1}{\sigma^2} \sim \chi^2_{p-q} \quad \left(\text{as } \frac{RSS_0}{\sigma^2} \sim \chi^2_{N-q}, \ \frac{RSS_1}{\sigma^2} \sim \chi^2_{N-p} \right.$$

$$\downarrow$$

$$\frac{(RSS_0 - RSS_1)/(p-q)}{RSS_1/(N-p)} \sim F_{p-q, N-p} \quad \text{under } H_0$$

Linear Models – Multiple Linear Regression

$$Y_i = \vec{x}_i^T \vec{\beta} + e_i$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad \hat{\beta} \sim N(\vec{\beta}, \sigma^2 (X^T X)^{-1})$$

$$RSS = (\vec{Y} - X^T \vec{\beta})^T (\vec{Y} - X^T \vec{\beta}) = \vec{Y}^T \vec{Y} - \hat{\beta} X^T \vec{Y}$$

$$\sigma^2 = \frac{RSS}{N-m-1}$$

$$H_0: \beta_j = 0 \Rightarrow t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{N-m-1}$$

H_0: ν of parameters are all 0

$$\downarrow$$

$$F = \frac{(RSS_0 - RSS)/\nu}{RSS/(N-m-1)} \sim F_{\nu, N-m-1}$$

Interpretation of Regression params:

- Partial regression coefficient:
  - β_j in multiple linear regression model measures rate of change in $\hat{}$ (mean response when) $\vec{x}_j$ with all others fixed
- Total regression coefficient:
  - when only 1 response variable, β_j is total regression coef
  - measure rate of change of mean response with $\vec{x}_j$, ignoring other explanatory variables
- Partial & total regression coefs only equal if X-matrix columns are orthogonal

Assumptions:
- linearity
- Normality of e_i (outliers/skew bad)
- Homogeneity of variances
  - variance independent of predictor variables
- independence of e_i

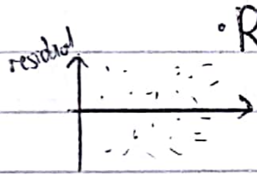
- Checking model adequacy:
  - Checking assumptions:

- Matrix plot:
  - predictors vs predictors:
    - check colinearity & leverage points ↙ not violations of assumptions, but problematic
  - predictors vs response:
    - check linearity, outliers & homoscedasticity
- Residuals & Standardized results

$\hat{e}_i = (Y_i - \vec{X}_i^T \hat{\beta})$ estimators for e_i

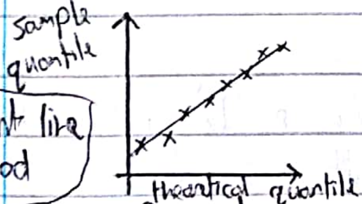
standardized residuals → $r_i = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ $\quad (H = X(X^TX)^{-1}X^T)$

r_i should have mean 0, stdev 1; assess their size (~5% larger than 2)



random pattern = good

- Residual plots:
  - Predictor vs residuals: non-linearity, heteroscedasticity (autocorrelation)
  - Fitted val vs residuals: same as above
  - Observation order vs residuals: (if order is informative, eg time)
    - autocorrelation, heteroscedasticity



Straight line = good

- Normal probability plot:
  - should look like a straight line
  - indicate ~~to~~ deviations from normality & outliers

- Remedies for violated assumptions:
  - Non-linearity:
    - transform predictors (eg log, exp...) or non-linear models
  - Non-normality:
    - transformation (help skew), robust regression (esp help outliers)
  - Heteroscedasticity:
    - weighted least squares, robust regression if only a few outliers
  - Dependence (of errors):
    - possibly time series models
    - only really effects stdevs & CIs

- R^2, coefficient of determination

$$R^2 = \frac{\sum(\hat{u}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{CTSS} \quad \leftarrow CTSS = \sum(y_i - \bar{y})^2$$

  - proportion of total variation explained by model
  - small = assumption violated, or predictors missing or large error variance

- Outliers:
  - Regression outliers: large residuals on plots
  - Leverage points: potential of a single point to influence model
    - don't violate model assumptions
    - good: is 'in-line' w/ other points, ie omitting this point won't change model much
    - bad: is out of line, removing point changes model a lot

standardised measure of leverage → $MD_i = \sqrt{(\vec{x}_i - \bar{x})^T \hat{\Sigma}_x^{-1} (\vec{x}_i - \bar{x})}$; $MD_i^2 = (N-1)\left[h_{ii} - \frac{1}{N}\right]$

- Cook's Statistic: $(MD_i^2 \sim \chi^2_{p-1})$
  - $\hat{\beta}_{(i)}$ is estimator of $\vec{\beta}$, omitting point i

$$D_i = \frac{1}{p\hat{\sigma}^2} (\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})$$

$$\left(= \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2\right) \leftarrow \text{efficient}$$

(r_i: residual)

- Problems:
  - $X^T X$ non-invertible; Cook's statistic unstable or $\hat{\beta}$ unstable
  - Masking effect:
    - leverage points 'hide' each other; need to use robust estimators, can then look at residual plots to help find outliers

Linear Models - Robust Regression

- Outlies are common in real world data, often some explanatory variables are inevitably missed
  - may want to ignore them, or not (eg Finance)
- In LS regression, squared residuals make sensitive to outliers (ie try to minimize $\hat{\sigma}^2$)
  - Instead, try to minimize absolute sum or median residual (ie try to minimize $\hat{\sigma}$)

- Effect of Leverage points:
  - in simple linear regression, $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum v_i y_i$ (slope estimator) $\left(v_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}\right)$
    - largest weight to large $(x_i - \bar{x})$ points (ie large leverage)
    - R^2 analogy $\Rightarrow$ low good leverage $\uparrow R^2$
      bad leverage $\downarrow R^2$

- Efficiency:
  - ~~designed to be~~ LS is optimal, robust methods designed to be less effected by outliers
    $\Rightarrow$ asymptotic variance higher than that of LS estimator

- M-estimation:
  - minimize sum of objective function, ρ, or residuals $\sum \rho(r_i)$, where $\rho(u) \geq 0$, $\rho(0) = 0$, $\rho(u) = \rho(-u)$, $\rho(u) \geq \rho(u')$ if $|u| < |u'|$

- $\hat{\beta}_M$ minimizes $\sum \rho\left(\frac{e_i(\vec{\beta})}{\sigma}\right) = \sum \rho\left(\frac{Y_i - \vec{x}_i^T \vec{\beta}}{\sigma}\right)$

$$\Rightarrow \sum \vec{x}_i \rho'\left(\frac{Y_i - \vec{x}^T \hat{\beta}_M}{\sigma}\right) = 0 = \sum \vec{x}_i w_i u_i,$$

'robustness weights' $w_i = \frac{\rho'(u_i)}{u_i}$, $u_i = (Y_i - \vec{x}^T \hat{\beta}_M)/\sigma$

σ is unknown: estimate w/ $\sigma_{MAD} = MED(|\hat{e}_i - MED(\hat{e})|)/.6745$ (need to iterate this too)

if w_i fixed, $\hat{\beta}_M = (X^T W X)^{-1} X^T W Y$ $\quad (W = diag(w_1, \ldots, w_N))$

as not, iterate

$$\hat{\beta}_M^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} y \quad \text{until convergence}$$

- residuals only enter through $\rho' \Rightarrow$ bounded ρ' = bounded effect of large residuals

- Distribution of M-estimators & efficiency:
  - Consistent if $E(\rho'(Z)) = 0$, where $Z \sim F$ (dist of errors)
    - holds for all symmetric F if ρ' bounded
  - Tests & CIs:

$$\sqrt{n}(\hat{\beta}_M - \vec{\beta}) \overset{n\to\infty}{\sim} N(0, V(\rho', F)L^{-1}), \quad V(\rho', F) \text{ complicated matrix}$$

$$L = \lim \frac{1}{N} X^T X$$

- Efficiency:

$$C = \sigma^2 (X^T X)^{-1} \quad \text{for LS}$$

$$V(\psi, F)L^{-1} = bC \quad \text{for M-estimation}$$

$1/b$ = efficiency, need $b\times$ as many observations to match LS estimator precision

- Objective function:

$\rho(u) = u^2 \Rightarrow$ LS

$\rho(u) = |u| \Rightarrow L_1$-Regression

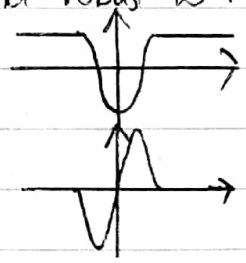
ρ': not small for large residuals; ~~unbounded~~ not robust to ↑ leverage points

Bisquare $\longrightarrow$

$$\rho_B(u) = \begin{cases} \frac{1}{6}\left(1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right) & \text{if } |u| \le c \\ \frac{1}{6} & \text{if } |u| > c \end{cases}$$

0 for large residuals! $\longrightarrow$

$$\rho_B'(u) = \begin{cases} \frac{u}{c^2}\left[1 - \left(\frac{u}{c}\right)^2\right]^2 & \text{if } |u| \le c \\ 0 & \text{if } |u| > c \end{cases}$$



- c tunes efficiency vs robustness

- S-estimator:
  - Bisquare M-estimator relies on (unknown) σ

M-estimator of scale $\longrightarrow$

$$M(\vec{z}) = \frac{1}{N}\sum_i \rho\left(\frac{z_i}{S_M(\vec{z})}\right) = E_{N(0,1)}\rho(Z) \quad \text{for observations } (z_1, \dots, z_N)$$

  - $\rho(u) = u^2 \Rightarrow E_{N(0,1)}\rho(Z) = 1$ and $S_M^2(\vec{z}) = \frac{1}{N}\sum_i z_i^2$

- Linear regression:
  - minimise $S_M(e_1(\vec{\beta}), \dots, e_N(\vec{\beta})) = S_M$ (residuals)

$$= \min S_M\{e_1(\vec{\beta}), \dots, e_N(\vec{\beta})\}$$

- Computation:
  - find $\beta^{(k+1)}$ for given $S_M^{(k)}(\vec{z}^{(k)})$ & $\beta^{(k)}$
  - find $S_M^{(k+1)}$ for given $\vec{z}^{(k+1)} = \{e_1(\beta^{(k+1)}), \dots, e_N(\beta^{(k+1)})\}$
  - repeat until convergence
    - maybe local optima; repeat w/ diff $\beta^{(0)}$

- MM-estimator:
  - combine S & M-estimation
  - S_M estimator of σ^2, then use this for M-estimator
  - iterate

Linear Models - Variable Selection

$$Y_i = \vec{x}_i^T \vec{\beta} + e_i$$

- aim to reduce #no. variables, ie find $\beta_i = 0$ (or close to 0)
- motivation:
  - + p large, N small $\Rightarrow X^TX$ nearly colinear, unstable estimation
    - N > 10p desired
  - + simpler, clear explanation
  - − worse for prediction
  - − unstable variable selection if some are correlated
  - − bias β remaining β to be high, so p-values biased low
- t-tests & F-tests:
  - F-test anova table: no variable affects response
  - F-test: y of $\beta_i = 0$ (given other m-y in model)
  - t-test: $\beta_j = 0$, given other m-1 in model
  - note: dropping 1 variable changes p-values for others!
- Best subset selection:
  - m variables $\Rightarrow 2^m$ possible models; which is best?
  - for fixed k no. variables, find model which:
    - minimizes RSS, $\hat{\sigma}^2$, maximizes R^2
    - cannot use this to compare models w/ different k
- Criteria to compare models:
  - Akaike Information Criteria (AIC)

$$AIC = -2\hat{l}(\text{model}) + 2p$$

($\hat{l}$(model) is max of (log-likelihood function); p: no. regression params (=d+1))

$$\hat{l}(\text{model}) = -\frac{N}{2}\log\sigma - \frac{RSS}{2\sigma^2} + \text{const.}$$

$$\Rightarrow AIC = N \log\frac{RSS}{N} + 2p, \text{ using } \hat{\sigma}^2_{ML} = \frac{RSS}{N}$$

  - Other functions of RSS:
    - Mallows C_p or adjusted R^2; in general too 'soft'; doesn't reduce no. vars enough

Leave-one-out cross validation

↓

- LOO-CV:
  - train on N-1 points, test on holdout point
  - repeat N times & average losses
  - − computationally demanding;
  + not based on any model assumptions
- Stepwise methods:
  - best subset is exponential in no. variables; too slow!
  - Backward elimination:
    - fit full model
    - for k=m to 0:
      - fit ~~not~~ all models w/ k-1 vars (k dif models)
      - find best
    - get sequence of 'best' models for all k, use AIC or LOO-CV to pick best from sequence
  - Forward selection
    - start from 0 variables, keep adding variable which produces largest decrease in RSS
    - again select from sequence produced w/ AIC or LOO-CV
  - can use t/F-tests here to decide when to stop
  - backwards/forwards give dif models; often backwards better
  - − not guaranteed to find best subset model, & very unstable
- Lasso:
  - standardise variables (to mean 0, stdev 1)

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$ (makes β_i comparable size), minimise $S(\vec{\beta}) = \sum e_i^2$

subject to $\sum_{i=1}^{m} |\beta_i| \leq t$

  - set t by LOO-CV
  - encourages sparse solutions, & equivalent to L_1-regularization

+ more stable & 'smooth' for large m & close-to collinearity situations
− biased, worse than stepwise in 'clear-cut' situations
− computation is complicated

Linear Models: ANOVA ← Analysis of Variance

- Y_i depends on categorical variables $\in \{0, 1\}$
- ANOVA refers to breakdown of ~~RSS~~ CTSS into sum of RSS & SS

- Comparison of Groups - One Way Layout

$$Y_{ij} = \mu_i + e_{ij}, \quad i = 1, \dots, I \leftarrow \text{groups}, \quad j = 1, \dots, n_i$$

$$\vec{Y} = X\vec{\beta} + \vec{e} \qquad \vec{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{In_I} \end{pmatrix}, \; X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{pmatrix} \; (\rightarrow I), \; \vec{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}, \; \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$$

- cols of $X = I$ indicator variables, rows $= n$ observations
  - eg $n_1 = 2, n_2 = 3, n_3 = 1 \Rightarrow X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
  - this way, $\vec{\beta} = \vec{\mu}$
- usual hypothesis: $\mu_1 = \mu_2 = \dots = \mu_I$: H_0, H_1: at least 2 differ
  - test with $RSS_0 = CTSS$

- Two Way Layout:
  - observations classified according to 2 factors

$$Y_{ijk} = \underbrace{\mu}_{\text{overall expected value}} + \underbrace{\alpha_i}_{\substack{\text{effect of} \\ \text{level } i \text{ of} \\ \text{factor 1}}} + \underbrace{\beta_j}_{\substack{\text{effect of} \\ \text{level } j \text{ factor 2}}} + \underbrace{\gamma_{ij}}_{\substack{\text{interaction} \\ \text{effect}}} + e_{ijk} \qquad \begin{matrix} i = 1, \dots, I, \; j = 1, \dots, J \\ k = 1, \dots, n_{ij} \end{matrix}$$

- for $I = J = n_{ij} = 2$

$$Y = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ \vdots \\ Y_{222} \end{pmatrix}, \; X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \; \vec{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{pmatrix}$$

- but $X^T X$ not invertible as too many params
  - need to introduce constraints, eg

$$\sum \alpha_i = 0, \; \sum \beta_j = 0 \text{ and } \sum_i \gamma_{ij} = 0, \; \sum_j \gamma_{ij} = 0$$

(allow one of each to be determined by all others, effectively allowing us to ~~remove~~ shrink X)

• these lead to $X = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$, $\vec{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \gamma_{22} \end{pmatrix}$ (other params from constraints)

• Usual F-tests: (or AIC/LOO-CV)

no interactions → all $\gamma_{ij} = 0$

can remove a factor → all $\alpha_i = 0$

can remove a factor → all $\beta_j = 0$

] done in order is 'restricted back selection'

Generalised Linear Models – General Theory

- **Components:**
  - Random $Y_1, \dots, Y_N$
  - Systematic η_i
  - Link function $g(\mu_i) = \eta_i, \quad \mu_i = E(Y_i)$ (monotonic & differentiable)
- **Linearity:**
  $\eta_i = \vec{x}_i^T \vec{\beta}$
- **Special cases:**
  - Linear Models: $g(\mu) = \mu$ (identity link)
  - Logistic model:
    $g(\mu) = \log\left(\frac{\mu}{n-\mu}\right)$ ← $n=1 \Rightarrow$ logit of μ
    - used for Bernoulli trials ($n=1$) data, or $Y_i \sim \text{Binomial}(n_i, \pi_i)$
  - Log-linear model:
    $Y_i \sim \text{Poisson}(\mu_i) \Rightarrow \log(\mu_i) = \vec{x}_i^T \vec{\beta} \Rightarrow g(\mu) = \log(\mu)$
  - Contingency tables:

- **Exponential family:**
$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$
(ϕ: dispersion / scale param; $a(\phi) = \phi/w$ for known w)

Distribution	θ	ϕ	$b(\theta)$	$c(y, \phi)$
Poisson(μ)	$\log \mu$	1	e^θ	$-\log y!$
Bin(n, π)	$\log \frac{\pi}{1-\pi}$	1	$n\log(1+e^\theta)$	$\log\binom{n}{y}$
$N(\mu, \sigma^2)$	μ	σ^2	$\frac{1}{2}\theta^2$	$-\frac{1}{2}\left(y^2/\phi + \log(2\pi\phi)\right)$

(for all, $w=1$)

- known $\phi \Rightarrow$ exp family
- unknown $\phi \Rightarrow$ exp dispersion family
- **Properties:**
  (differentiate $\int f(y)dy = 1$, use $a(\phi) \neq 0$)
  - $E(Y) = b'(\theta)$
  - $\text{Var}(Y) = b''(\theta)\, a(\phi)$
  - $\text{Var}(Y) = V(\mu)\, a(\phi) = V(\mu)\frac{\phi}{w}$

	$V(\mu)$
Poisson(μ)	μ
Bin(n, π)	$\mu(n-\mu)/n$
$N(\mu, \sigma^2)$	1

  - Canonical links:
    - g such that $g(\mu) = \theta$
    - log, logit & identity for Poisson, Bin & N

$$f(y_i;\theta_i,\phi)=\exp\left[\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right]$$

- Some GLM theory:
  - Y_i ~~~ from distribution w/ pdf in exp family (& independent)
  - let $E(Y_i)=\mu_i$, $\eta_i=\sum_j \beta_j x_{ij}$
  - Estimation:
    - $l=\sum l_i$, $l_i=\log f(y_i;\theta_i,\phi)\left[=\log P(y)\right]$

$$\Rightarrow \frac{\partial l}{\partial \beta_j}=\sum_i \left(\frac{y_i-\mu_i}{V(\mu_i)}\right)\frac{d\mu_i}{d\eta_i}x_{ij}=0$$

  - Iterative procedure:
    - Estimation of $\vec{\beta}$:
      - cannot solve likelihood eqns algebraically
      - Fisher scoring method:
        - initialize $\hat{\beta}^{(0)}$,
        - iterate

$$w_{ii}=\frac{1}{V(\mu_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2 \qquad z_i=\eta_i+(y_i-\mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)$$

$$\rightarrow \underbrace{(X^TWX)^{(s-1)}}_{\text{evaluated w/ } \hat{\beta}^{(s-1)}}\hat{\beta}^{(s)}=(X^TW\vec{z})^{s-1}$$

    - Estimation of ϕ:

Pearson statistic $\longrightarrow X^2=\sum_i \frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}$] 0 mean, unit variance RVs

$$\frac{X^2}{\phi}\sim\chi^2_{N-p} \Rightarrow \hat{\phi}=\frac{\hat{X}^2}{N-p}$$

- Sampling distribution of $\hat{\beta}$:

Taylor expand around true β_0, evaluate at $\hat{\beta}$ $\longrightarrow$

$$\left.\frac{\partial l}{\partial \beta}\right|_{\hat{\beta}} \approx \left.\frac{\partial l}{\partial \beta}\right|_{\beta_0}+(\hat{\beta}-\beta_0)\left.\frac{\partial^2 l}{\partial \beta^2}\right|_{\beta_0}$$

$$\Rightarrow (\hat{\beta}-\beta_0)\approx -\frac{\left.\frac{\partial l}{\partial \beta}\right|_{\beta_0}}{\left.\frac{\partial l}{\partial \beta^2}\right|_{\beta_0}} \sim N(0,I^{-1}) \text{ as } n\to\infty \text{ (by CLT)}$$

$$\Rightarrow \hat{\beta}\sim N_p(\vec{\beta},I^{-1})$$

($I=X^TWX/\phi$, for W at convergence of Fisher Scoring method)

GLM - General Theory

- Covariance Matrix of $\hat{\beta}$:

Fisher information matrix $(j,k)^{th}$ element → $E\left(\frac{-\partial \lambda}{\partial \beta_j}\frac{\partial \lambda}{\partial \beta_k}\right) = \sum_i \sum_{i'} \left(\frac{E[(Y_i-\mu_i)(Y_{i'}-\mu_{i'})]}{Var(Y_i)Var(Y_{i'})}\right)\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \mu_{i'}}{\partial \eta_{i'}} x_{ij} x_{i'k}$

(i.e. the -Hessian!)

$$\mathcal{I} = \sum_i \left(\frac{x_{ij}x_{ik}}{Var(Y_i)}\right)\cdot\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

(as $E((Y_i-\mu_i)(Y_{i'}-\mu_{i'})) = Var(Y_i)$ if $i=i'$, $Cov(Y_i,Y_{i'})=0$ if $i\neq i'$)

$$\mathcal{I} = X^T W X/\phi$$

$\therefore$ Covariance matrix of $\hat{\beta}$ is $\mathcal{I}^{-1} = (X^T W X)^{-1}\phi$

- Confidence intervals for model parameters:
  - assume known dispersion parameter, ϕ (unknown ϕ same, except must estimate ϕ & use appropriate t-distribution)
  - Single parameter:

$$\hat{\beta} \sim_{approx} N_p(\vec{\beta}, \mathcal{I}^{-1})$$

approx $(1-\alpha)\%$ CI → $\Rightarrow \hat{\beta}_j \pm z_{\alpha/2}\, se(\hat{\beta}_j)$, where $se(\hat{\beta}_j) = \sqrt{(\mathcal{I}^{-1})_{jj}}$

  - linear combination of parameters:

$$\psi = \vec{c}^T\hat{\beta}$$

$$\psi \sim_{approx} N(\psi, \vec{c}^T\mathcal{I}^{-1}\vec{c})$$

- Hypothesis testing:
  - $H_0: \beta_j = 0 \Rightarrow \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{approx}{\sim} N(0,1)$ under H_0
  - likelihood test → H_0: subset of $p-q$ params all 0

$$\Rightarrow 2[\lambda(\hat{\beta}) - \lambda(\hat{\beta}_0)] \sim \chi^2_{p-q} \text{ under } H_0$$ (applies when ϕ known)

Similar to RSS →

- Deviance:

$$D = 2[\lambda(\hat{\beta}_{sat}) - \lambda(\hat{\beta})]\phi$$ ($\lambda(\hat{\beta}_{sat})$ is saturated likelihood, achieved when $\hat{\mu}_i = y_i$)

scaled deviance → $D^* = D/\phi \sim \chi^2_{N-p}$ (under H_0 above)

- can also express likelihood test

$$D_0^* - D^* \sim \chi^2_{p-q}$$

- unknown ϕ:
  - under same H_0, $D_0^* - D^* \sim \chi^2_{p-q}$ & $D^* \sim \chi^2_{N-p}$
  - if $(D_0^* - D^*) \perp\!\!\!\perp D^*$, under H_0 & in large sample limit

$$F = \frac{(D_0^* - D^*)/p-q}{D^*/(N-p)} = \underbrace{\frac{(D_0 - D)/p-q}{D/(N-p)}}_{\text{no need to know } \phi} \sim F_{p-q,N-p}$$

GLM - Binomial Data & Logistic Regression

$Y_1, \dots, Y_N$ independent, $Y_i \sim \text{Bin}(n_i, \pi_i)$

$g(\mu_i) = \vec{x}_i^T \vec{\beta}$, $\mu_i = E(Y_i) = n_i \pi_i$

$\eta = \vec{x}_i^T \vec{\beta} = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ ← logit link gives linear logistic model (CDF of logistic dist) $\pi = \frac{1}{1+e^{-\eta}}$

(alternatives: probit (inverse standard normal CDF) $\pi = \Phi(\eta)$

complementary log-log ($\vec{x}_i^T \vec{\beta} = \log[-\log(1-\pi_i)]$ (CDF of extreme value dist) $\pi = 1 - \exp(-e^{\eta})$))

- these choices ensure $0 \le \pi \le 1$ & $-\infty < g(\mu) < \infty$ ($\Rightarrow$ no constraints on $\vec{\beta}$)

- Interpretation of logistic regression parameters:
  - $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) =$ log-odds
  - β_j measures rate of change of log odds w/ x_j (& other x_i held constant)
  - eg $x_1 \in \{0, 1\}$, $\Rightarrow$ odds on success w/ $x_1 = 1$ is odds of success w/ $x_1 = 0 \times \exp[\beta_1]$
- Some basic results:
  - likelihood: $\frac{\partial l}{\partial \beta_j} = \sum_i \frac{y_i - \mu_i}{V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \sum_i (y_i - \hat{\mu}_i) x_{ij} = 0$
  - MLE:

Fisher Iterations

$$(X^T W X)^{(s-1)} \hat{\beta}^{(s)} = (X^T W \vec{z})^{(s-1)}$$

for logistic model, $W_{ii} = \frac{1}{V_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$ $\left(V_i = \text{Var}(Y_i) = V(\mu_i) = n_i \pi_i (1-\pi_i)\right.$, $\left.\frac{d\mu_i}{d\eta_i} = n_i \pi_i (1-\pi_i)\right)$

$W_{ii} = n_i \pi_i (1-\pi_i)$

$z_i = \eta_i + (y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)$, $\left(\frac{d\eta_i}{d\mu_i} = \frac{1}{n_i \pi_i (1-\pi_i)}\right)$

- Sampling distribution of $\hat{\beta}$:

$$\hat{\beta} \overset{\text{approx}}{\sim} N_p(\vec{\beta}, (X^T W X)^{-1}) \quad \text{for large } N$$

- Deviance: $D = 2\sum_i \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]$ ($l(\hat{\beta}_{sat})$, $l(\hat{\beta})$)

  $D \sim \chi^2_{N-p}$ if model is true

- $H_0: \beta_j = 0$

  $\Rightarrow \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1)$

- H_0: ν of $\beta_1, \ldots, \beta_m$ are 0

  $\Rightarrow D_0 - D \sim \chi^2_\nu$

alternative goodness of fit test →

- Pearson chi-squared statistic:

  $X^2 = \sum_i \frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \sim \chi^2_{N-p}$ under model (observed, expected)

  (can be shown $D \approx X^2$) $\left(X^2 = \sum \frac{(O-e)^2}{e}\right)$

- Binary data:
  - if $n_i = 1$ $\forall i$, results still hold but D no longer good for assessing quality of fit
  - Can test by grouping observations into $g \approx 10$ groups according to $\hat{\pi}_j$, then calculate $X^2 \sim \chi^2_{g-2}$

- Checking model adequacy:
  - Plot residuals:
    - Raw: $\hat{e}_i = Y_i - n_i\hat{\pi}_i$
    - Pearson: $X_i = \hat{e}_i / \sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$
    - Standardized Pearson: $r_{P_i} = X_i / \sqrt{1-h_{ii}}$
    - Deviance: $\sum d_i^2 = D$
    - Standardized Deviance: $r_{D_i} = d_i / \sqrt{1-h_{ii}}$

    $\begin{cases} h_{ii} = H_{ii} \\ H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2} \end{cases}$

  - Cooks statistic: $D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_{P_i}^2$

GLM - Binomial Data & Logistic Regression

- Model selection:
  - Stepwise / best subset selection:
    - can then use AIC, which for binomial data

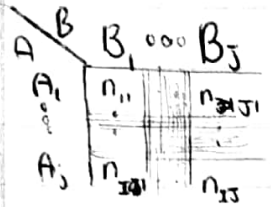
$$AIC = D + 2p$$

  - Compare model deviances:

$$D_{M_1} - D_{M_2} \sim \chi^2_{p_1 - p_2}$$ (equivalent to likelihood test)

- ~~Analysis of deviance:~~

GLM - Contingency tables: log-linear modelling



π_{ij} = prob observation in cell (i,j)

$\mu_{ij} = N\pi_{ij}$ = expected frequency

expected row frequency ↓ $\frac{\mu_{i+} \cdot \mu_{+j}}{N}$ ← expected column frequency

if $A \perp\!\!\!\perp B$, $\pi_{ij} = \pi_{i+}\,\pi_{+j}$, $\mu_{ij} = \frac{\mu_{i+}\cdot\mu_{+j}}{N}$

- $H_0: A \perp\!\!\!\perp B \Rightarrow \log \mu_{ij} = \log \mu_{i+} + \log \mu_{+j} - \log N$
  $= \alpha_i + \beta_j + \chi$
- $H_1: A \not\perp\!\!\!\perp B \Rightarrow \log \mu_{ij} = \alpha_i + \beta_j + \lambda + \phi_{ij}$ (log-linear models (log of expected response))
  $(= A + B + A:B)$
  - this is over-parameterised; need constraints on parameters to solve
    - eg $\alpha_1 = \beta_1 = \phi_{1j} = \phi_{i1} = 0$
    - or $\sum \alpha_i = \sum_i \phi_{ij} = \sum_j \phi_{ij} = 0$
  - H_1 is saturated; as many effective params as table cells,
    $\Rightarrow$ can fit perfectly, $N\pi_{ij} = n_{ij}$

- no clear interpretation of params; mainly want to test if zero

- Three-way contingency tables:
  - now variables A, B, C w/ I, J, K categories
  - Saturated model:

$$\log \mu_{ijk} = A + B + C + A:B + A:C + B:C + A:B:C = A*B*C$$

  - again, over-parameterised, need similar constraints to reduce to IJK effective parameters
  - Submodels:

	Independences	π_{ijk}
1. $A + B + C$	$A \perp\!\!\!\perp B \perp\!\!\!\perp C$	$\pi_{i++}\cdot\pi_{+j+}\cdot\pi_{++k}$
2. $A + B + C + A:B$	$A, B \perp\!\!\!\perp C$	$\pi_{ij+}\cdot\pi_{++k}$
3. $A + B + C + A:B + A:C$	$B \perp\!\!\!\perp C \mid A$	$\pi_{ij+}\cdot\pi_{i+k}/\pi_{i++}$
4. $A + B + C + A:B + A:C + B:C$		

$P(B,C|A)\cdot P(A)$ = $P(B|A)\cdot P(C|A)\cdot P(A)$

- Fitting the models:
  - maximal likelihood (joint dist of observation is multinomial)
    - equivalent to joint distribution of independent Poission RVs, conditioned on their sum
      $\Rightarrow$ fit as if observed frequencies are independent Poisson RVs
- Goodness of fit & model checking:
  - done as for Poisson data, using G^2
    $G^2 = 2\sum o \log\left(\frac{o}{e}\right)$ (o=observed, e=expected)
  - can compare nested models using G^2
  - Residuals:
    - Raw $\hat{e}_{ij} = n_{ij} - \hat{\mu}_{ij}$
    - Pearson $X_{ij} = \hat{e}_{ij} / \sqrt{\hat{\mu}_{ij}}$
    - Standardised Pearson $r_{P_{ij}} = X_{ij} / \sqrt{1-h_{ij}}$
    - Deviance $\sum_{ij} d_{ij}^2 = D$
    - Standardised Deviance $r_{D_{ij}} = d_{ij} / \sqrt{1-h_{ij}}$

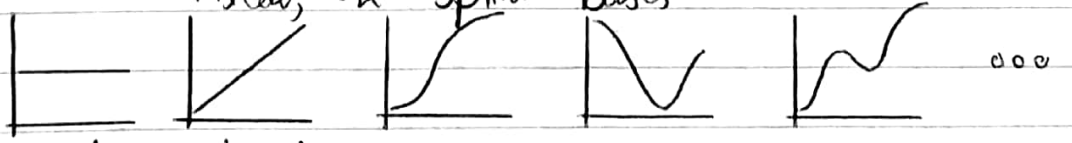
Generalized Additive Models: An Overview

$$g[E(Y_i)] = \eta_i = \vec{X}_i^* \vec{\Theta} + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}, x_{i4}) + \dots$$

$X_i^* = i^{th}$ row of X^*
$X^* =$ model matrix for any parametric model components
$\vec{\Theta} =$ parameter vector
$f_j =$ smooth functions for covariates x_{ij}
note $\sum_i f_j(x_{ij}) = 0$

- Difference w/ GLM:
  - GLM models effect of x_{ij} as $\Theta_0 + \overbrace{\sum_j \Theta_j x_{ij}}^{\text{linear!}}$
  - GAM models as $\sum_j f_j(x_{ij})$

- Smooth functions:
  - represented using regression splines
    $f_j(x_j) = \sum_k \beta_{jk} b_{jk}(x_j)$, where b_{jk} are basis functions & β_{jk} regression parameters to learn
  - basis functions could be polynomial $(x, x^2, x^3 \dots)$, but these become highly collinear
    - instead, use spline bases



- Parameter estimation:
  - estimated using MLE, adding smoothing parameters $\vec{\lambda}$ to reduce overfitting (choose $\vec{\lambda}$ by, eg, LOO-CV)

maximise $l(\vec{\beta}) - \frac{1}{2} \sum_j \lambda_j \int [f_j^{d_j}(x_j)]^2 dx_j$ ← derivative of f (d_j usually = 0); encourage 'less rough' functions

$= l(\vec{\beta}) - \frac{1}{2} \sum_j \lambda_j \vec{\beta}^T S_j \vec{\beta}$ ← known coefficients matrix → $S = \sum_j \lambda_j S_j$

don't estimate $\vec{\lambda}$; would just choose very low regularisation!

$$\Rightarrow \hat{\beta} = (X^T W X + S)^{-1} X^T W \vec{z}$$ (note biased estimator, due to penalty)

- Inference:
  - Inference not standard due to fixed λ_j
  $$\vec{\beta} \mid \vec{y} \sim N(\hat{\beta}, (I+S)^{-1})$$ allows approximate inference (eg CIs) to be done
  - Variable selection & model comparison also not standard