① Univariate exponential family

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

θ_i parameter of interest, not θ_i

$E(Y_i) = b'(\theta_i)$
$Var(Y_i) = b''(\theta_i)\frac{\phi}{w_i}$

ϕ dispersion parameter, not of interest (nuisance)

Normal, Bernoulli, Binomial (regular), Poisson, Gamma ...

② MLR (prep)

$Y = X\beta + e, \quad e \sim N(0, \sigma^2 I)$

$E(Y) = X\beta, \quad Cov(Y) = \sigma^2 I$
$Y \sim N(X\beta, \sigma^2 I)$
or $Y_i \sim N(\mu_i, \sigma^2)$
$Y_1, ..., Y_n$ independent
univariate exp fam

③ MLR (prepared for GLM)

1) $Y_i \sim N_1, \quad \mu_i = E(Y_i), \quad \phi$
2) $\eta_i = x_i^T\beta$
3) $\eta_i = \mu_i$

④ GLM Model

1) $Y_i \sim f$ exp fam, $E(Y_i) = \mu_i$, $Var(Y_i) = $
2) $\eta_i = x_i^T\beta$
3) $\eta_i = g(\mu_i)$ and $\mu_i = h(\eta_i)$

canonical link:
$E(Y_i) = b'(\theta_i)$, $\eta_i = x_i^T\beta$

canonical link $\theta_i = \eta_i$
concave log-likelihood, $H(\mu): F(\beta)$

(2-4) Q: Why not

$Y_i = x_i^T\beta + \epsilon_i$ where $\epsilon_i \sim Poisson(0)$
$\mu_i \quad \epsilon_i \sim bin(1, \pi_i=!)$

If explosion then $x_i^T\beta + \epsilon_i$ not linear.
Need mean μ_i in specification of distribution.

⑪ LMM:

$Y_i = X_i\beta + U_i\gamma_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2 I)$
and $\gamma_i \sim N(0, Q)$
ϵ_i and γ_i independent for $i = 1, ..., m$

$N = \sum_{i=1}^m n_i$

Marginal model:
$Y_i = X_i\beta + \epsilon_i^*$ where $\epsilon_i^* = U_i\gamma_i + \epsilon_i \sim N(0, U_iQU_i^T + \sigma^2 I)$
so $Y_i \sim N(X_i\beta, V_i)$

Likelihood: $L(\beta, \phi) = \prod_{i=1}^m f(y_i | \beta, \phi)$

$\hat\beta = [\sum_{i=1}^m (X_i^T V_i^{-1} X_i)]^{-1} \sum_{i=1}^m (X_i^T V_i^{-1} Y_i)$
$\hat\beta \sim N(\beta, (\sum X_i^T V_i^{-1} X_i)^{-1})$ and when $\hat V_i$ replace V_i

REML: V_i: found from maximizing $A^T Y \sim N(0, A^T Cov(Y)A)$ where A could be a $(N-p)\times N$ version of $(I-H)$ and gives less biased estimator than using ML

Prediction of random effects γ (BLUP) and ϵ
$\hat\gamma_i = \hat Q U_i^T \hat V_i^{-1}(Y_i - X_i\hat\beta)$
$\hat\epsilon_i = Y_i - X_i\hat\beta - U_i\hat\gamma_i$
can be plotted to check model assumptions

⑤ GLM likelihood inference. Aim ML $\hat\beta$

$L(\beta) = \prod_{i=1}^n L_i(\beta)$ and $l(\beta) = \sum_{i=1}^n \ln(L_i(\beta)) = \sum_{i=1}^n l_i(\beta)$

Work with $L_i(\beta)$ and chain rule for differentiation

$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^n \frac{(y_i - \mu_i) x_i h'(\eta_i)}{Var(Y_i)} = X^T D \Sigma^{-1}(y - \mu)$

$D = diag(h'(\eta_i))$
$\Sigma = diag(Var(Y_i))$
$s(\beta) = $ score function
$E(s(\beta)) = 0$ (also $E(s_i(\beta)) = 0$)

$F(\beta) = $ expected Fisher info matrix
$F(\beta) = Cov(s(\beta)) = E(-\frac{\partial^2 l}{\partial \beta \partial \beta^T}) = X^T W X$
$W = diag(\frac{h'(\eta_i)^2}{Var(Y_i)})$

⑥ Fisher-scoring to find $\hat\beta$ numerically

$\hat\beta^{(t+1)} = \hat\beta^{(t)} + F^{-1}(\hat\beta^{(t)}) \cdot s(\hat\beta^{(t)})$

For GLM this can be written as an
IRWLS Iteratively reweighted least squares

$\hat\beta^{(t+1)} = (X^T W(\hat\beta^{(t)}) X)^{-1} X^T W(\hat\beta^{(t)}) \tilde z^{(t)}$

working response vector
$\tilde z_i = x_i^T \hat\beta^{(t)} + \frac{y_i - h(x_i^T\hat\beta^{(t)})}{h'(x_i^T\hat\beta^{(t)})}$

Solve $s(\beta) = 0$ to get $\hat\beta$ ML

Asymptotic distribution of $\hat\beta$
$\hat\beta \approx N(\beta, F^{-1}(\hat\beta))$
TWO important uses of $F^{-1}(\beta)$

Requirements: $n \to \infty$, $\lambda_{min} F_n \to \infty$
for all $\delta > 0$ $\max_i \|V_n(\beta) - I\| \to 0$ normed info matrix too advanced for us!

⑦ Requirements GLM model

+ Y_i's independent
+ X full rank ($F(\beta)^{-1}$ possible)
+ one-to-one and twice differentiable link function

⑧ Further GLM inference [CI, hyp, deviance]

CI for β_j: $[\hat\beta_j \pm z_{\alpha/2} SD(\hat\beta_j)]$ where $SD(\hat\beta_j) = \sqrt{F^{-1}(\hat\beta)_{[j,j]}}$

Hypothesis tests (a,b,c): $H_0: C\beta = d$ vs $H_1: C\beta \neq d$ r hypotheses
$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

a) **Wald test:** $w = (C\hat\beta - d)^T [CF^{-1}(\hat\beta)C^T]^{-1}(C\hat\beta - d) \approx \chi_r^2$
P-value: $P(\chi_r^2 > z_?^2)$ observed value
NB: $(\frac{\hat\beta_j - 0}{SD(\hat\beta_j)})^2 \sim \chi_1^2$ is also Wald test (summary print-out gives in R)

b) **Likelihood-ratio test (LRT):** B nested in A
A: larger model [H₁ true] → $\hat\beta_A$ with $L(\hat\beta_A)$
B: smaller model [H₀ true] → $\hat\beta_B$ with $L(\hat\beta_B)$
$LRT = -2[\ln L(\hat\beta_B) - \ln L(\hat\beta_A)] \approx \chi^2_{\#A - \#B}$
P-value: $P(\chi_r^2 > lrt)$
Requirement $LRT \approx \chi_r^2$: $n \to \infty$

c) **The score test:** $U = (s(\tilde\beta) - 0)^T F^{-1}(\tilde\beta)(s(\tilde\beta) - 0) \approx \chi_r^2$
$\tilde\beta$: param estimate when H₀ true
$s(\tilde\beta)$: subvector of score function where only the elements in H₀ but not in H₀ are present but evaluated based on $\tilde\beta$.
$F^{-1}(\tilde\beta)$: Submatrix of F^{-1} evaluated at $\tilde\beta$.
Used: test small (H₀) model against many H₁'s

⑨ Deviance (test) for GLM (perfect fit to the data)
A: saturated model - same number of param. as number of covariate patterns $n \in G$ patterns
B: candidate model: our model
Deviance $D = 2[\ln L(\hat\beta_A) - \ln L(\hat\beta_B)] \approx \chi^2_{G-p}$
$\hat\beta_A$ a function of the y's in each cov pattern
observed = d

⑨+ If we find that $P(\chi^2_{G-p} > d) \leq \alpha$ then our model is very far from the saturated model - and that's a bad thing. Then we "reject" H₀ that our model is good.
NB: for Poisson and binomial (grouped) GLM we need the number of obs for each covariate pattern to be large to trust the χ^2-distribution assumption.

⑩ AIC for GLM: can be used to choose between models - both nested and not
$AIC = -2l(\hat\beta, \hat\phi) + 2p^*$
$p^* = p + \#par$ in ϕ
can also compare different link functions

⑫ LMM: random intercept model $q=0$

$Y_{ij} = x_{ij}^T\beta + \gamma_{0i} + \epsilon_{ij}$
γ_{0i} iid $N(0, \tau_0^2)$

For the random intercept model $V_i = \sigma^2 I + \tau_0^2 11^T$ and
$Corr(Y_{ij}, Y_{il}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$ called intraclass correlation (ICC)

Plots of x_{ij} vs $\hat\beta_0 + \gamma_{0i} + \beta_1 x_{ij}$ (many parallel lines)

LMM: random intercept and slope (one x_{ij}), $q=1$
$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \gamma_{1i} x_{ij} + \epsilon_{ij}$
$\begin{bmatrix}\gamma_{0i}\\\gamma_{1i}\end{bmatrix} \sim N(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_0^2 & \tau_{01}\\\tau_{01} & \tau_1^2\end{bmatrix})$
$Cov(Y_{ij}, Y_{il})$ not so easy to interpret since dependent on x_{ij} and x_{il}.

⑬ Model selection for LMM

Fixed effects can be tested with Wald or LRT from ML estimates
Wald: $H_0: C\beta = d$ vs $H_1: C\beta \neq d$
$(C\hat\beta - d)^T [C(\sum X_i^T V_i^{-1} X_i)^{-1} C^T]^{-1}(C\hat\beta - d) \approx \chi_r^2$
LRT: $-2[\ln L(\hat\beta_B) - \ln L(\hat\beta_A)] \approx \chi^2_{\#par A - \#par B}$
A: model under H₁
B: model under H₀

For testing of random effects: trouble because on boundary of param space and normality of $\hat Q$ problematic
→ instead: use AIC to compare models
need to either choose ML or REML
$AIC_{ML} = -2l(\hat\beta, \hat\theta) + 2p^*$ $AIC_{REML} = -2l(\hat\beta, \hat\theta) + 2p^*$
both β's and # par in θ
$p + 1 + 1$ for random intercept and slope

⑭ GLMM: correlated non-normal responses Y_{ij}, $i=1,...,m$, $j=1,...,n_i$

1) Distributional assumption:
$Y_{ij}|\gamma_i$ has conditional distribution $f(y_{ij}|\gamma_i)$ [binomial, Poisson,...]

2) Structural assumption:
$\mu_{ij} = E(Y_{ij}|\gamma_i)$ linked to $\eta_{ij} = x_{ij}^T\beta + u_{ij}^T\gamma_i$
$\eta_{ij} = g(\mu_{ij})$
$\mu_{ij} = h(\eta_{ij})$

3) Distributional assumption for random effects
γ_i iid $N(0, Q)$

Marginal model: $f(y_{ij}) = \int f(y_{ij}|\gamma_i)f(\gamma_i)d\gamma_i$
not in closed form!
Which makes maximum likelihood estimation numerically challenging. Several solutions exists
We have not looked at CI and testing since asymptotic distr. of parameters challenging for small samples. Use asymptotic results on $\hat\beta \sim N(\beta, F(\beta^*)^{-1})$ for fixed effects.

⑮ Binary regression

$[f(y_i) = (\frac{\pi_i}{1-\pi_i})^{y_i}\pi_i^{?}(1-\pi_i)^{?}]$

1) $Y_i \sim bin(1, \pi_i)$, $\mu_i = E(Y_i) = \pi_i$ ← univariate exponential family (show!)
2) $\eta_i = x_i^T\beta$ [x_i is $p\times1$ column vector]
3) $\eta_i = g(\mu_i) = \log(\frac{\pi_i}{1-\pi_i})$ logit link (canonical)
$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$

other popular link functions: probit, complementary log-log

Interpretation of β with logit link
If x_{ij} increase from x_{ij} to $x_{ij}+1$ then
$\frac{\pi_i}{1-\pi_i}$ change with a multiplicative factor e^{β_j} (odds)
$\frac{P(Y_i=1|x_{ij}+1)}{P(Y_i=0|x_{ij}+1)} = \frac{P(Y_i=1|x_{ij})}{P(Y_i=0|x_{ij})} \cdot e^{\beta_j}$

Extension to $Y_i \sim bin(n_i, \pi_i)$ for grouped data
Still use logit link $\eta_i = logit(\pi_i)$
Same covariate pattern for all n_i, $i=1,...,G$ groups
$n_i = 1$: only continuous covariates

Likelihood: $L(\beta) = \prod_{i=1}^G \pi_i^{y_i}(1-\pi_i)^{n_i-y_i}(\begin{smallmatrix}n_i\\y_i\end{smallmatrix})$
Score function: $s(\beta) = \sum_{i=1}^G x_i(y_i - n_i\pi_i)$
Expected Fisher info: $F(\beta) = \sum_{i=1}^G x_i x_i^T n_i \pi_i(1-\pi_i)$
Deviance: $D = 2\sum_{i=1}^G[y_i \ln(\frac{y_i}{n_i\hat\pi_i}) + (n_i - y_i)\ln(\frac{n_i-y_i}{n_i-n_i\hat\pi_i})]$
saturated model has $\tilde\pi_i = \frac{y_i}{n_i}$ (insterted in D alone) instead of $\hat\pi_i$

⑯ Poisson regression

1) $Y_i \sim Poisson(\mu_i)$, $f(y_i) = \frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}$, $y_i = 0,1,2...$
$E(Y_i) = Var(Y_i) = \mu_i$, $\mu_i > 0$
2) $\eta_i = x_i^T\beta$
3) $\eta_i = \log(\mu_i)$ ← canonical link, log link
$\mu_i = \exp(\eta_i)$

Interpretation: $\mu_i = \exp(\eta_i) = \exp(x_i^T\beta)$
$= e^{\beta_0} \cdot e^{\beta_1 x_{i1}} ... e^{\beta_k x_{ik}} = e^{\beta_0}(e^{\beta_1})^{x_{i1}}...(e^{\beta_k})^{x_{ik}}$
If $\beta_j > 0$: then $\exp(\beta_j) > 1$ so μ_i increases if x_{ij} increases
$\beta_j < 0$: < 1 μ_i decreases
$\beta_j = 0$: $= 1$ no change in μ_i if x_{ij} change

Likelihood: $L(\beta) = \prod_{i=1}^n \frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}$
Log-likelihood: $l(\beta) = \sum_{i=1}^n [y_i\mu_i - \mu_i - \ln(y_i!)] = \sum_{i=1}^n y_i x_i^T\beta - \sum \exp(x_i^T\beta) + C$
Score function: $s(\beta) = \sum_{i=1}^n (y_i - \mu_i)x_i$
Expected Fisher info: $F(\beta) = \sum_{i=1}^n x_i x_i^T \mu_i$, $\hat\mu_i = \exp(x_i^T\hat\beta)$
Deviance: $D = 2\sum_{i=1}^n [y_i \cdot \ln(\frac{y_i}{\hat\mu_i}) - y_i + \hat\mu_i]$

Rate models and offset:
Want model for $\frac{Y_i}{t_i}$ instead of Y_i so $E(\frac{Y_i}{t_i}) = \frac{\mu_i}{t_i}$
Let $\log(\mu_i/t_i) = x_i^T\beta$ so $\log(\mu_i) = \log(t_i) + x_i^T\beta$
Proceed as usual, only add offset $\log(t_i)$ in R (with log link)

⑰ Gamma regression:

1) $Y_i \sim Gamma(\mu_i, v)$, $f(y_i) = \frac{1}{\Gamma(v)}(\frac{v}{\mu_i})^v y_i^{v-1}\exp(-\frac{v}{\mu_i}y_i)$, $y_i > 0$
2) $\eta_i = x_i^T\beta$
3) $\eta_i = g(\mu_i) = -\frac{1}{\mu_i}$ is canonical link, but $\eta_i = g(\mu_i) = \log(\mu_i)$ is more popular (same interpretation as for Poisson with log-link)
The gamma has v as nuisance parameter.

⑱ Categorical regression:

Y: one of c+1 possible categories $\sim multinomial(m, \pi)$
(c+1) reference category, dummy variable coding
Independent pairs (x_i, Y_i), $Y_{ir} = 1$ if obs i from category r
Aim: model and estimate the probabilities $\pi_{ir} = P(Y_{ir} = 1|x_i)$

a) Nominal responses (unordered) [alligator food]
$\ln(\frac{\pi_{ir}}{\pi_{i,c+1}}) = x_i^T\beta_r$ each response has its own β!
$\pi_{ir} = \frac{\exp(\eta_{ir})}{1 + \sum_{s\neq r}\exp(\eta_{is})}$
C·p parameters to estimate
$D = 2\sum_{i=1}^G\sum_{s=1}^{c+1} y_{is}\ln(\frac{\theta_{is}}{\hat\pi_{is}})$ Deviance
$\sim \chi^2_{(Gc-cp)}$ G: covariate patterns with n_i obs in group i.

Multivariate GLM
1) Y_i multinomial $\mu_i = \pi_i$ (grouped data $Y_i \sim M(n_i, \pi_i)$)
2) $\eta_i = \begin{bmatrix}\eta_{i1}\\ \vdots \\ \eta_{ic}\end{bmatrix} = \begin{bmatrix}x_{i1}^T\beta_1\\ \vdots \\ x_{ic}^T\beta_c\end{bmatrix}$
3) Link functions $g(\mu_i) = \eta_i$

b) Ordinal responses (ordered) [mental health]
$Y_i = r \Leftrightarrow \theta_{r-1} < U_i \leq \theta_r$
$P(Y_i \leq r) = F(\theta_r + x_i^T\beta)$ common β's for all categories but introduced θ's (instead of β_0's)
$\pi_{ir} = F(\eta_{ir})$
$\pi_{ir} = F(\eta_{ir}) - F(\eta_{i,r-1})$ ⟹ k+c parameters to estimate
We only consider F as cdf logistic distribution
$F(\eta_{ir}) = P(Y_i \leq r) = \frac{\exp(\eta_{ir})}{1 + \exp(\eta_{ir})} = \frac{\exp(\theta_r + x_i^T\beta)}{1 + \exp(\theta_r + x_i^T\beta)}$
Proportional odds model: $\frac{P(Y_i \leq r|x_i)}{P(Y_i > r|x_i)} = \frac{P(Y_i \leq r|x^*)}{P(Y_i > r|x^*)}$
$D = $ as above but $\sim \chi^2_{(Gc - c - k)}$

c) Parameter estimation: ML with "smart" design matrix [not on reading list]
d) Inference: $\hat\beta \sim N(\beta, F^{-1}(\hat\beta))$ as before

Univariate exponential family GLM:
Normal, Binomial, Poisson, Gamma
Multivariate exponential family GLM:
Multinomial (nominal, ordinal)