Generalized Linear Models

Want a model for $(Y|X_1, \ldots, X_p)$.

1. Specify a **family of distributions**.
2. Specify a **link function** $g()$ such that:

$$g\{E(Y|X_1, \ldots, X_p)\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

Note: $\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ referred to as **linear predictor**.

```
> glm(y~x, family=..., link=...)
```

Three primary examples:

$Y \sim N(\mu, \sigma^2)$, $\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

$Y \sim Bernoulli(p)$, $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Note: Binomial data can be treated as Bernoulli data. (Some issues lurking here - what is sample size? also deviance.)

$Y \sim Poisson(\lambda)$, $\log \lambda = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Note: Have given the 'default' link function in each case. This is more than convention, tied up with "exponential family" theory - the **canonical link function** is that which equates the **natural parameter of the distribution** with the **linear predictor**.

Other links sometimes used in Bernoulli case: probit $\Phi^{-1}(p)$, complementary log-log: $log\{-log(1-p)\}$.

Variance function and scale parameter

Write $\mu = E(Y)$ (so link function links μ to predictors).

Choice of distribution will lead to $Var(Y) = \phi v(\mu)$ for some **scale parameter** ϕ and **variance function** $v()$ describing the *mean-variance relationship*.

Sometimes ϕ is known, sometimes a nuisance parameter to be estimated.

Normal: $v(\mu) = 1$, ϕ unknown

Bernoulli: $v(\mu) = \mu(1 - \mu)$, $\phi = 1$.

Poisson: $v(\mu) = \mu$, $\phi = 1$.

Computation (say ϕ known):

Distribution + link = fully-specified prob. model. Gives log-likelihood function $l(\beta)$.

NR algorithm to maximize $l()$, requires evaluation of $l'()$ and $l''()$.

Fisher-scoring modification: Replace $l''()$ with $E_{Y|X}\{l''()\}$.

Turns out we have an **iteratively reweighted least squares** (IRLS) algorithm.

Important/curious property. **Algorithm turns out to be 100% determined by data, link function, and variance function.** That is, choice of distribution only matters in terms of variance function it yields, also ϕ doesn't matter.

Gives strategy for add-on estimation of ϕ when needed. Pretend $\phi = 1$ and get $\hat{\beta}$, then **deviance** (recall defn, D_M in text). Then estimate $\hat{\phi} = (n-p)^{-1} D_M$.

Rationale: real/scaled/residual deviance is $\phi^{-1} D_M$.

Also, the role of the variance function leads to **quasi-likelihood** ideas: choose a variance function as one sees fit, don't worry if it corresponds to a real probability model or not.

So, have a general regression strategy for multiple types of Y variable.

Point estimation.

SE / Interval estimation.

Hypothesis testing: nested, q predictors versus p predictors.

Hypothesis testing: general goodness-of-fit (p predictors versus saturated).

Residual-based diagnostics

Poisson regression example

Deaths from coronary heart disease in a (famous) cohort study...

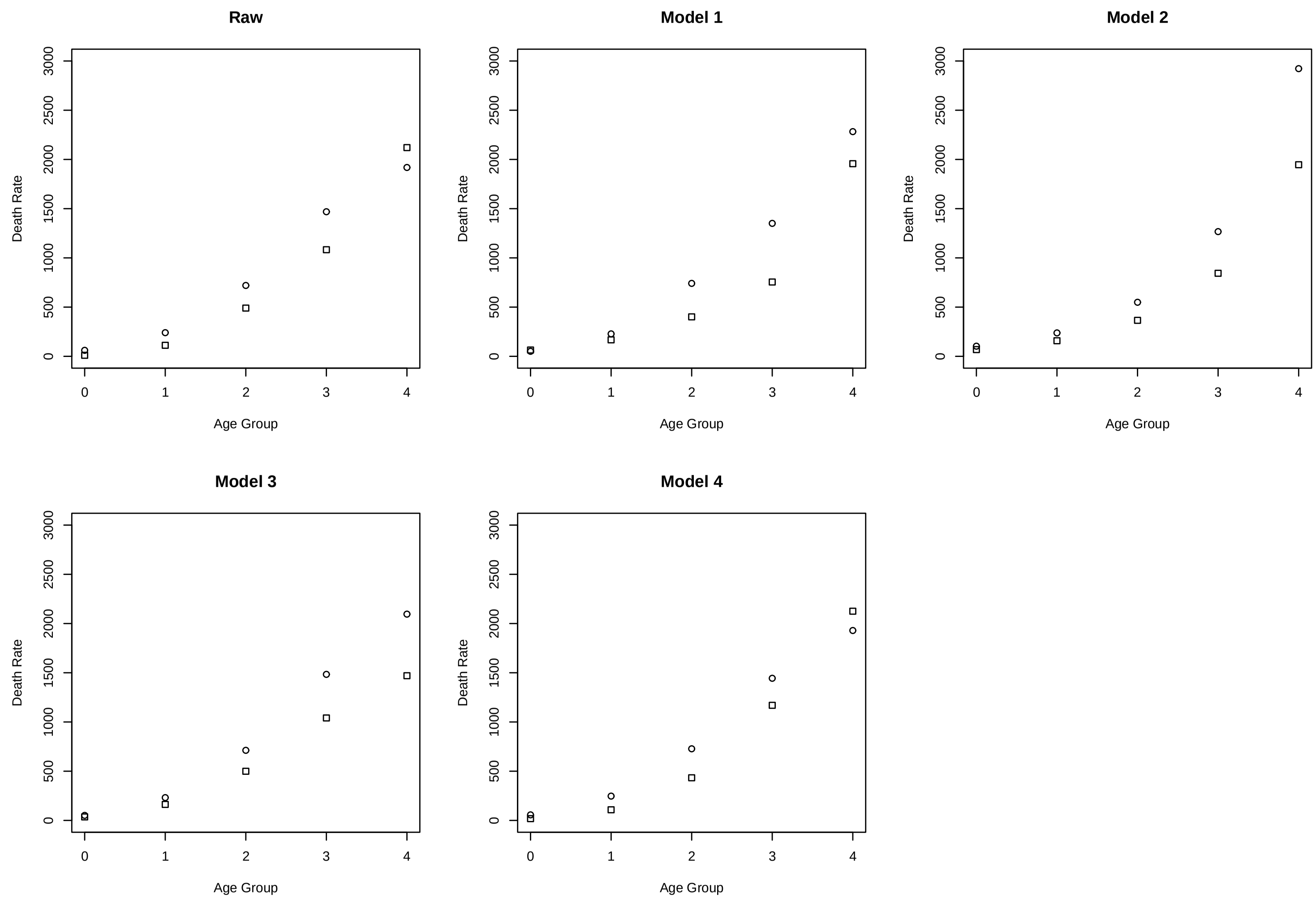
	Smokers		Non-smokers	
age	deaths	person-years	deaths	person-years
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

```
glm(formula = deaths ~ I(log(pyears)) + age.grp + smoke,
    family = poisson, data = dat)


               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -21.4739     2.2556  -9.520  < 2e-16 ***
I(log(pyears))   2.4352     0.2270  10.727  < 2e-16 ***
age.grp          1.7702     0.1542  11.478  < 2e-16 ***
smoke           -1.6991     0.3548  -4.789 1.68e-06 ***

    Null deviance: 644.269  on 9  degrees of freedom
Residual deviance:  25.576  on 6  degrees of freedom
```

```
glm(formula = deaths ~ offset(log(pyears)) + age.grp +
              smoke,
    family = poisson, data = dat)

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.28250    0.12076 -60.304  < 2e-16 ***
age.grp      0.83583    0.02904  28.777  < 2e-16 ***
smoke        0.40637    0.10720   3.791  0.00015 ***

    Null deviance: 935.067  on 9  degrees of freedom
Residual deviance:  69.182  on 7  degrees of freedom
```

```
glm(formula = deaths ~ offset(log(pyears)) +
              age.grp + I(age.grp^2) + smoke,
    family = poisson, data = dat)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.93363    0.16395 -48.391  < 2e-16 ***
age.grp       1.70594    0.12824  13.303  < 2e-16 ***
I(age.grp^2) -0.19438    0.02715  -7.159 8.14e-13 ***
smoke         0.35452    0.10737   3.302  0.00096 ***

    Null deviance: 935.067  on 9  degrees of freedom
Residual deviance:  12.176  on 6  degrees of freedom
```

```
glm(formula = deaths ~ offset(log(pyears)) + age.grp +
              I(age.grp^2) + smoke + smoke:age.grp,
    family = poisson, data = dat)

              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.61296    0.29172 -29.524  < 2e-16 ***
age.grp        1.98113    0.16025  12.363  < 2e-16 ***
I(age.grp^2)  -0.19768    0.02737  -7.223 5.08e-13 ***
smoke          1.13342    0.28077   4.037 5.42e-05 ***
age.grp:smoke -0.30755    0.09704  -3.169  0.00153 **

    Null deviance: 935.0673  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
```


Raw
Model 1
Model 2
Model 3
Model 4
Death Rate
0
500
1000
1500
2000
2500
3000
Age Group
0
1
2
3
4