21 NOV - LEC 10

GENERALIZED LINEAR MODELS (GLMs)

Let's start by reviewing the hypotheses of the normal linear model, but highlighting some components. In particular, we can identify three elements:

1. stochastic component: $Y_i \sim N(\mu_i, \sigma^2)$ indep. $i=1,\dots,n$ (Gaussian assumption)
2. systematic component: $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} = \tilde{x}_i^T \underline{\beta}$ (linearity)
3. a function that relates μ_i and η_i: for the lm, identity function: $\mu_i = \eta_i$

What happens if these hypotheses are not satisfied?

- the response variable is not Gaussian:
  - → estimate the model anyway relying on the OLS estimate. You still have good properties, but you can not do inference.
  - → transform the Y and fit a model on the transformed data (careful: if linearity was ok, after transforming the data you may lose it)
- the relationship between μ_i and η_i is not linear:
  - → transform the data (if you don't lose normality and homoscedasticity...)

Sometimes these remedies are not sufficient: you need more flexible models.

The normal linear model is not always adequate to describe the data.
GLMs extend the LM in two main directions:

- NONLINEAR relationship between μ_i and η_i
- NON-GAUSSIAN distribution of Y_i

Moreover, they no longer assume homoscedasticity of the response ($var(Y_i) \neq \sigma^2 \ \forall i$)

ASSUMPTIONS OF A GLM

1. DISTRIBUTION (hypothesis on the stochastic component)
   $Y_i \sim f(y_i ; \theta)$ with f DENSITY THAT BELONGS TO THE EXPONENTIAL FAMILY
2. LINEAR PREDICTOR
   $\eta_i = \tilde{x}_i^T \underline{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$ linearity w.r.t. $\underline{\beta}$
3. MONOTONE LINK FUNCTION that relates μ_i and η_i:
   $g(\mu_i) = \eta_i$ with $g(\cdot)$ invertible ($\Rightarrow \mu_i = g^{-1}(\eta_i)$)

Remark: the distributive assumption

The exponential family is a set of probability distributions. All densities in this set have a common "special" structure that allows the derivation of several inferential properties within a unified framework.
This means that it is possible to study the properties of a general GLM and they will apply to all particular cases.
A lot of commonly used distributions belong to this class. Some examples are: Gaussian, Bernoulli, binomial, Poisson, negative binomial.
We will only study two cases: Bernoulli and Poisson.

Remark 2

Notice that, different from the Gaussian LM, here we CAN NOT separate the random and the systematic component.
For the Gaussian we could write $Y = \underbrace{\mu}_{\text{systematic}} + \underbrace{\varepsilon}_{\text{random}}$
This additive form only holds for the Gaussian.
This is clear from the fact that, for example, if $Y \sim Poisson(\lambda)$, then it does <u>not hold</u> that $Y + \mu \sim Poisson(\lambda + \mu)$.