Topic 1

Review of Linear Regression/Normal Error Regression Model:

- Build and fit a model which describes the relationship between a dependent response variable and set of explanatory or predictor variables
- $Y_i = \alpha_0 + \alpha_1 * X_1 + \varepsilon_i$
- Y_i is the ith observation of dependent variable Y, X_i is the ith observation of independent variable X, α_0 and α_1 are unknown parameters and ε_i is a random error term which is $N(0, \sigma^2)$
- σ^2 is constant, not dependent on X_i and $Cov(\varepsilon_i, \varepsilon_j) = 0$, uncorrelated error
- Response Y_i is decomposed into systematic (non-random term): $\alpha_0 + \alpha_1 X_i$ and non-systematic (random term): ε_i
- $E[Y_i] = \alpha_0 + \alpha_1 X_i$ and $Var[Y_i] = \sigma^2$
- Normality Assumption: ε_i are normally distributed and so the response Y_i is normally distributed

Estimation of α_0 and α_1:

- Can be done under OLS or MLE
- $\widehat{\alpha_0} = \bar{Y} - \widehat{\alpha_1} * \bar{X}$
- $\widehat{\alpha_1} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$
- Under normality assumptions for ε_i, OLS and MLE methods produce the same estimates

Fitted Regression Line

- Obtain $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$
- The fitted line equation is $\widehat{Y_i} = \widehat{\alpha_0} + \widehat{\alpha_1} * X_i$
- $\widehat{Y_i}$ is the estimate of the mean response $E[Y_i]$
- Regression line is a straight line in the X-Y plane
- Residuals: $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$
- $\widehat{\varepsilon_i}$ is an estimate of ε_i
- MLE of σ^2 is the mean of the squared residuals
- $\widehat{\sigma^2} = \frac{1}{n} * \sum \widehat{\varepsilon_i}$, this is a biased estimate
- Unbiased estimate of σ^2 is $\frac{1}{n-2} * \sum \widehat{\varepsilon_i}$, subtract 2 degrees of freedom due to 2 estimations of alpha's

Testing Assumptions:

- Use residuals to test if the regression is linear, if random error terms have constant variance, if random error terms are independent and if error terms are Normal Dist
- If some assumptions are not valid, regression model doesn't fit the data well

Intro to GLMs:

- Normal error linear regression model may not be suitable as assumptions underlying the model are not appropriate
- A generalised linear model (GLM) is more flexible since it incorporates the non-normality of the response and nonlinear relationship between mean response $E[Y_i]$ and predictor $X_i{}'s$

- Detect normal linear model is not suitable through common sense, histogram of the data, a normal probability plot (Q-Q plot) or a normality test (Anderson darling)
- Q-Q plot = Need straightish line, density = symmetrical
- Detect non-constancy of variance through data plot or residual plot
- Implication of non-constancy – mean and variance are totally unrelated to each in the Normal distribution but for many non-normal distributions, the variance is linked to the mean. Hence the variance is not constant as it changes with mean, and the normality assumption for the response Y_i or ε_i is not valid and therefore normal error regression model is not likely to provide a good fit.
- Under a GLM, response Y_i can be any distribution from the exponential family (Normal, Exp, Poisson, Gamma, Binomial, Inverse Gaussian, Neg Bin)
- Exponential family has good mathematic properties to estimate the unknown parameters easily using the maximum likelihood estimation

Link Functions and Examples:

- In normal error regression model, mean of response is a linear function of the explanatory variables, if $\mu_i = E(Y_i)$, then
$$\mu_i = \beta_0 + \beta_1 * X_{1i} + \cdots + \beta_n * X_{ni}$$
- In a GLM, the relationship between the mean response and the explanatory variables may not be linear, can be exp fn, reciprocal etc
- Linear Predictor η_i
- $\eta_i = \beta_0 + \beta_1 * X_{1i} + \cdots + \beta_n * X_{ni}$
- There exists a relationship between the mean response μ_i and the predictors which can be expressed as $\mu_i = \eta_i$ or $\mu_i = \exp(\eta_i)$ etc
- These transformations are usually rearranged to be a function of μ_i and are called link functions
- $\eta_i = \mu_i$ is the identity link function
- $\eta_i = \ln(\mu_i)$ is the log link function
- $\eta_i = \frac{1}{\mu_i}$ is the reciprocal link function
- $\eta_i = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$ is the logit link function
- Generally, if $\eta_i = g(\mu_i)$, g(mu) is called the link function

Exponential Family:

- For GLM, assume the distribution of the random error term is in the exponential family (Exp, Bernoulli, Binomial, Normal, Poisson, Gama, NB)
- A distribution belongs to the exp family if it can be written in the form
$$f(y;\theta,\emptyset) = \exp\left(\frac{y\theta - b(\theta)}{a(\emptyset)} + c(y,\emptyset)\right)$$
- a,b,c are function
- θ is the natural parameter and $\emptyset$ is the scale parameter
- $Normal: \theta = \mu, \emptyset = \sigma^2, b(\theta) = \frac{\theta^2}{2}, a(\emptyset) = \emptyset, c(y,\emptyset) = -\frac{y^2}{2\emptyset} - \frac{1}{2}\ln(2\emptyset\pi)$
- $Poisson: \theta = \ln(\lambda), \emptyset = 1, b(\theta) = \exp(\theta), a(\emptyset) = 1, c(y,\emptyset) = -lny!$
- Method: Put everything into exp using $e^{\ln()}$ and rearrange to get form
- When find θ, sub back into f() to get all others