Team 1:
Chin Pak Cheng
Kaiyi Wang
Kwok Leung Ong

Generalised Linear Models Analysis Process with R Programming :: **CHEAT SHEET**

Step 1 : Establish Research Question

It is always a good habit to set up a clear **research question** at the very beginning. Review and revise it and don't get lost when doing modelling.

Step 2 : Predictors Selection

- Anderson Healthcare Utilization Model
  - Predisposing factors
  - Enabling factors
  - Need
- Sociodemographic information
- Health status
- Risk factors
- Etc.

Step 3 : Exploratory Data Analysis and Univariate analysis

- Data structures
  - `summary(data)`
  - `str(data)`
  - `dplyr::glimpse(data)`
- Distribution of the variables
  - `plot(y ~ x)`
  - `boxplot(y ~ x)`
  - `hist(x)`
- Categorical data relationship
  - `mosaic(~x + y, data)`

Step 4 : Multivariable GLM Fitting

Model Fitting

- Model fitting with all the variables
  - Logistics model: `glm(y ~ x, family = binomial, data)`
  - Poisson model: `glm(y ~ x, family = poisson, data)`
  - Binomial model: `glm(cbind(cases, controls) ~ x, family = binomial, data)`
  - Log-binomial model : `glm(y ~ x, family='binomial'(link='logit'), data)`
- Show the GLM result
  - `summary(model)`

  Note: Identified the significant variables with 5 – 20% levels, examine the overall significance

Model Diagnostics Methods

- Model Predictions
  - Linear predictor scale: `predict(model)`
  - Predicted probability scale : `predict(model, type="response")`
- Raw residuals
  - `residuals(model, type="response")`
  - `binnedplot(predict(model), residuals(model)`
- Detecting unusual observations
  - By Q-Q plot `qqnorm(residuals(model))`
  - By Leverage `halfnorm(hatvalues(model))`

Step 5 : Model Re-fitted

- Re-fit the model by excluding non-significant variables
  - `drop1(model, test="Chi")`
  - `step(model)`
- Examine the model with the above diagnostics methods

Step 6 : Model Comparisons

Statistical Inference

- ANOVA
  - `anova(model_1, model_2, test="Chi")`
- AIC comparison $(AIC=-2\log L+2q)$

Goodness of Fit

- Brier score test
  - `predprob <- predict(model, type="response")`
  - `Brier_score <- mean((predprob - df$outcome)^2)`
- PseudoR2 (DescTools package)
  - `PseudoR2(model, which = "all")`
- Check ROC curve and calculate AUC (pROC package)
  - `roc(df$outcome, df$predprob, percent=TRUE, plot=TRUE)`

Step 7 : Check Interaction Between Variables

- Test the interaction between the variables when necessary.
  - `glm( y ~ x1 * x2 , family = binomial, data)`

Step 8 : Interpretation

- For logistic model: `Odds = exp(β)`
- Confidence Intervals: `confit(model)`
- Confusion matrix: examine the sensitivity and specificity of the model
- Interpretate the **OR, RR, HR** in a correct way