2.6 Validation and Sensitivity Analysis – R Functions

Part 1 – Theoretical concepts

R2 Function

# model y = 2x + z + 6  
df <- data.frame(x <- seq(0, 2, by = 0.05) + rnorm(n = 41, mean = 0, sd = 1)  
, z <- seq(-1, 3, by = 0.1) + rnorm(n = 41, mean = 0, sd = 1)  
, y <- 6 + 2 \* x + z + rnorm(n = 41, mean = 0, sd = 1))  
  
# modelling   
res <- lm(y ~ x + z, data = df)  
summary(res)

##   
## Call:  
## lm(formula = y ~ x + z, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.44205 -0.48973 -0.08631 0.45026 1.56860   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.07712 0.18236 33.33 < 2e-16 \*\*\*  
## x 1.92834 0.12665 15.22 < 2e-16 \*\*\*  
## z 0.92658 0.08389 11.04 2.01e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7699 on 38 degrees of freedom  
## Multiple R-squared: 0.9395, Adjusted R-squared: 0.9364   
## F-statistic: 295.3 on 2 and 38 DF, p-value: < 2.2e-16

ROC Function

# modelling of expiration in icu against age, gender and vent\_day  
logit.mod <- glm(icu\_exp\_flg ~ vent\_day + age + gender  
 , data = aline.cohort, family = "binomial")  
summary(logit.mod)

##   
## Call:  
## glm(formula = icu\_exp\_flg ~ vent\_day + age + gender, family = "binomial",   
## data = aline.cohort)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7980 -0.4811 -0.2650 -0.1432 3.0724   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -16.00197 535.41121 -0.030 0.976   
## vent\_day 0.13913 0.02451 5.677 1.37e-08 \*\*\*  
## age 0.06090 0.00563 10.817 < 2e-16 \*\*\*  
## genderF 9.56460 535.41123 0.018 0.986   
## genderM 9.47683 535.41122 0.018 0.986   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1120.93 on 1775 degrees of freedom  
## Residual deviance: 912.18 on 1771 degrees of freedom  
## AIC: 922.18  
##   
## Number of Fisher Scoring iterations: 12

Only vent\_day and age appear to be significant in the model.

library(pROC) # install.packages("pROC")  
plot(roc.res <- roc(aline.cohort$icu\_exp\_flg, logit.mod$fitted.values))

##   
## Call:  
## roc.default(response = aline.cohort$icu\_exp\_flg, predictor = logit.mod$fitted.values)  
##   
## Data: logit.mod$fitted.values in 1606 controls (aline.cohort$icu\_exp\_flg 0) < 170 cases (aline.cohort$icu\_exp\_flg 1).  
## Area under the curve: 0.8159

Part 2 – Case study

Caliper Function

## Full Model  
dependent <- "aline\_flg"  
independent <- c(#personnal variables  
 "age", "gender\_num", "weight\_first", "sofa\_first"  
 , "service\_num", "day\_icu\_intime\_num", "hour\_icu\_intime"  
 # flags  
 , "chf\_flg", "afib\_flg", "renal\_flg", "liver\_flg", "copd\_flg"  
 , "cad\_flg", "stroke\_flg", "mal\_flg", "resp\_flg"  
 # labs var  
 , "map\_1st", "hr\_1st", "temp\_1st", "spo2\_1st", "wbc\_first"  
 , "hgb\_first", "platelet\_first", "sodium\_first", "potassium\_first"  
 , "tco2\_first", "chloride\_first", "bun\_first", "creatinine\_first"  
 , "po2\_first", "pco2\_first")  
# Missing values: replace by -1, as flag or positive values only  
indep.cohort <- aline.cohort[, independent]  
indep.cohort[is.na(indep.cohort)] <- -1  
full.mod <- glm.fit(x = indep.cohort  
 , y = aline.cohort[, dependent], family = binomial(link = "logit"))  
# summary(full.mod)  
  
independent2 <- c(#personnal variables  
 "service\_num", "day\_icu\_intime\_num", "hour\_icu\_intime", "sofa\_first"  
 # flags  
 , "liver\_flg", "copd\_flg"  
 # labs var  
 , "map\_1st", "hr\_1st", "temp\_1st", "spo2\_1st", "wbc\_first"  
 , "platelet\_first", "sodium\_first", "potassium\_first"  
 , "chloride\_first", "bun\_first")  
  
 ## random Model  
rand.mod <- glm.fit(x = indep.cohort[, independent2]  
 , y = aline.cohort[, dependent], family = binomial(link = "logit"))  
  
 library(Matching)

model.calip.calc <- function(model, seq.cal) { # model <- "full" ; seq.cal <- 0.01  
 # one-to-one matching without replacement (the "M=1" option)  
 # Estimating the treatment effect on the treated (the "estimand" option defaults to ATT)  
 if(model == "full") X <- full.mod$fitted else X <- rand.mod$fitted  
 ps <- Match(Y=NULL, Tr=aline.cohort$aline\_flg, X=X, M=1, estimand='ATC', caliper=seq.cal, replace=F)  
   
 # data set of the index used for the match  
 data2 <- aline.cohort[c(ps$index.treated,ps$index.control)  
 , c("day\_28\_flg", "aline\_flg")]  
   
 # fit the newly created dataset  
 glm2 <- glm(day\_28\_flg~aline\_flg, family=binomial, data=data2)  
   
 # output: caliper sequence, name of the model  
 # aline\_flg coefficients and p-value for model and seq.caliper  
 data.frame(s = seq.cal, mod = model  
 , coef = exp(glm2$coefficients[2])  
 , p\_value = coef(summary(glm2))[2, 4], stringsAsFactors = F)  
 }  
   
 # list of the results  
 res.caliper <- mapply(model.calip.calc, c(rep("full", 10), rep("rand", 10)), 1:10/100, SIMPLIFY = F)  
  
 # final data frame  
res.caliper.df <- do.call("rbind", res.caliper)  
  
qplot(x = s, y = coef, color = p\_value, data = res.caliper.df, size = 2) +  
 facet\_grid(.~mod) + scale\_size(guide = F)

Cohort Function

library(ggplot2)  
aline.cohort$d28\_Survival <- ifelse(aline.cohort$hosp\_exp\_flg == "1", "N", "Y")  
qplot(x = vent\_day, data = aline.cohort, bins = 28) + facet\_grid(d28\_Survival~., scales = "free\_y") + ggtitle("Survival 28 days")

Ventilation Function

vent.mod <- glm(hosp\_exp\_flg ~ vent\_day + aline\_flg, data = aline.cohort, family = "binomial")  
summary(vent.mod)

##   
## Call:  
## glm(formula = hosp\_exp\_flg ~ vent\_day + aline\_flg, family = "binomial",   
## data = aline.cohort)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5435 -0.5278 -0.4960 -0.4701 2.1426   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.19495 0.11503 -19.081 < 2e-16 \*\*\*  
## vent\_day 0.13544 0.02233 6.064 1.33e-09 \*\*\*  
## aline\_flg 0.10813 0.15185 0.712 0.476   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1421.5 on 1775 degrees of freedom  
## Residual deviance: 1377.9 on 1773 degrees of freedom  
## AIC: 1383.9  
##   
## Number of Fisher Scoring iterations: 4

library(ResourceSelection)  
  
hoslem.test(vent.mod$y, fitted(vent.mod), g = 10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: vent.mod$y, fitted(vent.mod)  
## X-squared = 16.624, df = 8, p-value = 0.03428