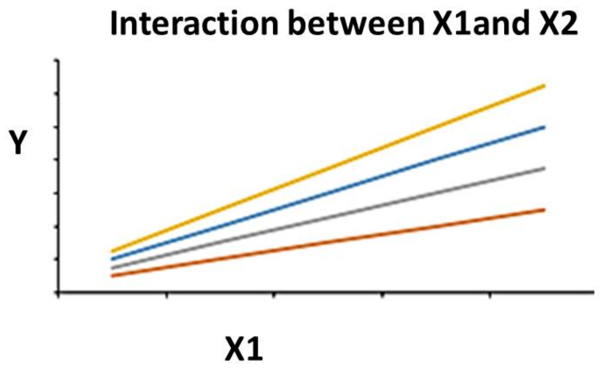
Data 603:Statistical Modelling with Data

Logistic Regression

Part III : Model Building in Multiple Logistic Regression Model and Assumptions

Model building in Multiple Regression (An Interaction Model with both Quantitative and Qualitative variables)

An interaction occurs if the relation between one predictor, X_1, and the outcome (response) variable, Y, depends on the value of another independent variable, X_2. The regression coefficient for the product term represents the degree to which there is an interaction between the two variables. The effect of X_1 on Y is not the same for all values of X_2, which, in linear regression, is graphically represented by non-parallel slopes.



Non-parallel slopes represent interation terms between X1 and X2

If slopes are parallel, the effect of X_1 on Y is the same at all levels of X_2, and there is no interaction. **Variable X1 and X2 may be binary or continuous** . Interactions are similarly specified in logistic regression if the response is binary. The right hand side of the logit equation includes coefficients for the predictors, X1,X2, and X1*X2.

If the interaction coefficient β_3 is significant, we conclude that the association between X_1 and the probability that $Y = 1$ depends on the values of X2, X1 and X2 may be binary or continuous.

The test of the interaction may be conducted with **the Wald chi square test** or **a likelihood ratio test** comparing models with and without the interaction term.

For example, using Default data to predict the probability of default, test the interaction term for the logistic regression model

```r
library(ISLR)
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

mylogit <- glm(default ~ balance+income, data = Default, family = "binomial")
summary(mylogit)#Wald z test

##
## Call:
## glm(formula = default ~ balance + income, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

```
interlogit <- glm(default ~ balance+income+balance*income, data = Default, fa
mily = "binomial")
summary(interlogit)
```

```
## 
## Call:
## glm(formula = default ~ balance + income + balance * income, 
##     family = "binomial", data = Default)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.5415  -0.1441  -0.0570  -0.0207   3.7546  
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -1.092e+01  9.489e-01 -11.504   <2e-16 ***
## balance         5.265e-03  5.648e-04   9.323   <2e-16 ***
## income          1.600e-06  2.683e-05   0.060    0.952    
## balance:income  1.193e-08  1.638e-08   0.728    0.466    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1578.4  on 9996  degrees of freedom
## AIC: 1586.4
## 
## Number of Fisher Scoring iterations: 8
```

```
anova(mylogit,interlogit,test="Chisq")
```

```
## Analysis of Deviance Table
## 
## Model 1: default ~ balance + income
## Model 2: default ~ balance + income + balance * income
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      9997     1579.0                     
## 2      9996     1578.4  1  0.53539   0.4644
```

```
#likelihood ratio test
lrtest(mylogit,interlogit)
```

```
## Likelihood ratio test
## 
## Model 1: default ~ balance + income
## Model 2: default ~ balance + income + balance * income
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -789.48                     
## 2   4 -789.22  1 0.5354     0.4644
```

From the output, by using **the Wald Z test** and **Likelihood ratio test** ,we see that the p-value =0.466>005 (from the Wald Z test)>0.05 and the p-value =0.4644>0.05 (from Likelihood Ratio test). Therefore, the interaction term is not significant. We should drop this term out of the model.

For Default data, consider the Multiple Logistic model with both Qualitative and Quantitative variables with interation terms. We add a Student predictor (qualitative variable) into the logistic model and also add all interaction terms.

```
library(ISLR)
library(lmtest)
mylogit<- glm(default ~ balance+income+factor(student), data = Default, famil
y = "binomial")

#Wald z test for testing individual predictors
summary(mylogit)
```

```
##
## Call:
## glm(formula = default ~ balance + income + factor(student), family = "bino
mial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## balance              5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income               3.033e-06  8.203e-06   0.370  0.71152
## factor(student)Yes -6.468e-01  2.363e-01  -2.738  0.00619 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

```
mylogit1<- glm(default ~ balance+factor(student), data = Default, family = "b
inomial")
summary(mylogit1)
```

```
##
## Call:
```

```
## glm(formula = default ~ balance + factor(student), family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4578  -0.1422  -0.0559  -0.0203   3.7435
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
## balance             5.738e-03  2.318e-04  24.750  < 2e-16 ***
## factor(student)Yes -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.7  on 9997  degrees of freedom
## AIC: 1577.7
##
## Number of Fisher Scoring iterations: 8
```

```
interlogit <- glm(default ~ balance+factor(student)+balance*factor(student),
data = Default, family = "binomial")
summary(interlogit)
```

```
##
## Call:
## glm(formula = default ~ balance + factor(student) + balance *
##     factor(student), family = "binomial", data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4839  -0.1415  -0.0553  -0.0202   3.7628
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.087e+01  4.640e-01 -23.438   <2e-16 ***
## balance                     5.819e-03  2.937e-04  19.812   <2e-16 ***
## factor(student)Yes         -3.512e-01  8.037e-01  -0.437    0.662
## balance:factor(student)Yes -2.196e-04  4.781e-04  -0.459    0.646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
```

```
##
## Number of Fisher Scoring iterations: 8

#Likelihood Ratio Test for testing the interation term
anova(mylogit1,interlogit,test='Chisq')

## Analysis of Deviance Table
##
## Model 1: default ~ balance + factor(student)
## Model 2: default ~ balance + factor(student) + balance * factor(student)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      9997     1571.7
## 2      9996     1571.5  1  0.20944   0.6472

lrtest(mylogit1,interlogit)

## Likelihood ratio test
##
## Model 1: default ~ balance + factor(student)
## Model 2: default ~ balance + factor(student) + balance * factor(student)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -785.84
## 2   4 -785.74  1 0.2094     0.6472
```

H_0: $\beta_3 = 0$
reduced model is true (with no interation term)
H_1: $\beta_3 \neq 0$
larger model is true(with interation term)

The likelihood ratio statistic is

$$\begin{aligned}\triangle G^2 &= -2logL \text{ from the reduced model} - (-2logL \text{ from larger model})\\ &= -2(-785.84) - (-2(-785.74)) = 0.2094\\ &\text{The p} - \text{value is } = 0.6472 > \alpha = 0.05\end{aligned}$$

Therefore, we reject the null hypothesis which means that the interaction term (student*balance) is insignificant to be in the model.

Inclass Practice Problem

Example: The German Credit Data contains data on 6 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.The independent variables are listed below

Creditability= (1 if good credit, 0 if bad credit)

Balance=Account Balance (Categorical variable with 4 levels)

$$Balance = \begin{cases} 1 \text{ if balance is more than } 5000 \\ 2 \text{ if balance is } 3001-5000 \\ 3 \text{ if balance is } 1001-3000 \\ 4 \text{ if balance is less than } 1000 \end{cases}$$

Duration= Duration of credit in months (months)

Employment=Length of current employment (years)

Amount=Credit amount (dollars)

Age=Age (year)

Build the logistic regression model for predicting the probability of hiring. Check whether interation terms should be added into the model or not.

$$\hat{y} = \frac{e^{0.61527+0.07785X_1-0.039772X_2+0.426465X_{3i}+0.559255X_{4i}+0.423117X_{5i}+0.024819X_1*X_{3i}+0.163106X_1*X_{4i}+0.48}}{1+e^{0.61527+0.07785X_1-0.039772X_2+0.426465X_{3i}+0.559255X_{4i}+0.423117X_{5i}+0.024819X_1*X_{3i}+0.163106X_1*X_{4i}+0}}$$

$$where,$$

$$Balance = \begin{cases} 1 \text{ if balance is more than } 5000 \\ 2 \text{ if balance is } 3001-5000 \\ 3 \text{ if balance is } 1001-3000 \\ 4 \text{ if balance is less than } 1000 \end{cases}$$

Inclass Practice Problem

Experience in hiring. Suppose you are investigating the hiring practices of a particular firm. Build the logistic regression model for predicitng the probability of hiring. Check whether interation terms should be added into the model or not. The data are provided in **DISCRIM.csv file**

Logistic Regression Assumptions

Logistic regression is widely used because it is a less restrictive than other techniques such as simple and multiple linear regression. Because of it, many researchers do think that LR has no an assumption at all.

First, logistic regression does not require _a linear relationship between the dependent and independent variables__.

Second, the error terms (residuals) do not need to be normally distributed_.

Third, homoscedasticity (constanct varaince) is not required.

Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale.

However, there are some assumptions still apply.

First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.

Independence Assumption (Independent observations and errors)

Identical to linear regression, the assumption of independent errors states that errors should not be correlated for two observations. In logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data. This typically occurs when the data for both dependent and independent variables are observed sequentially over a period of time-called **time-series data**. Therefore, if cases are selected at random, the independent observations condition is met. If no time series data have been used, the independent errors condition is met.

Linearity Assumption (Linear relationship between between response and predictors)

For linear regression the assumption is that the outcome variable has a linear relationship with the explanatory variables, but for logistic regression this is not possible because the outcome is binary.

Multicolinearity

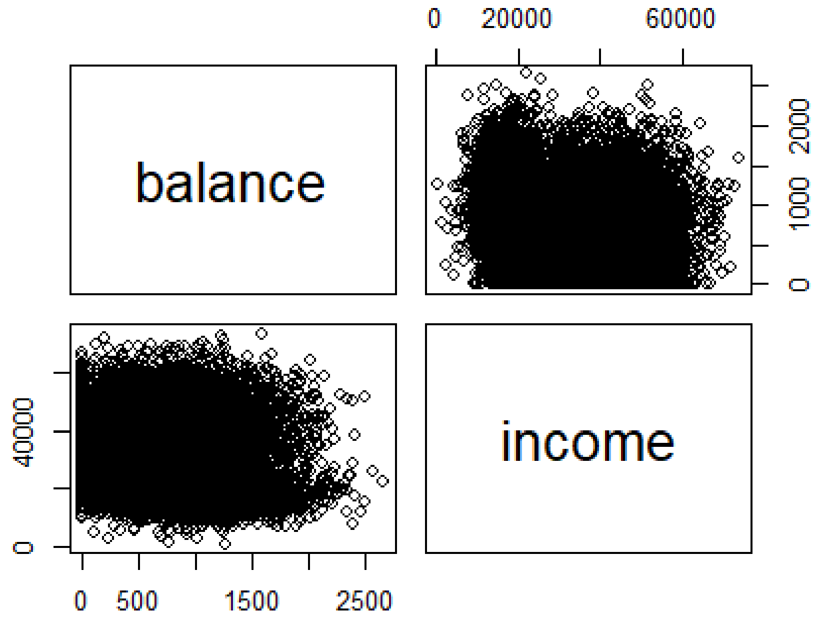
Logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other. We can apply ggpairs and compute VIF from multiple linear regression to check for multicollinearity.

For example, using Default data to predict the probability of default, check Multicolinearity Assumption for the fitted model

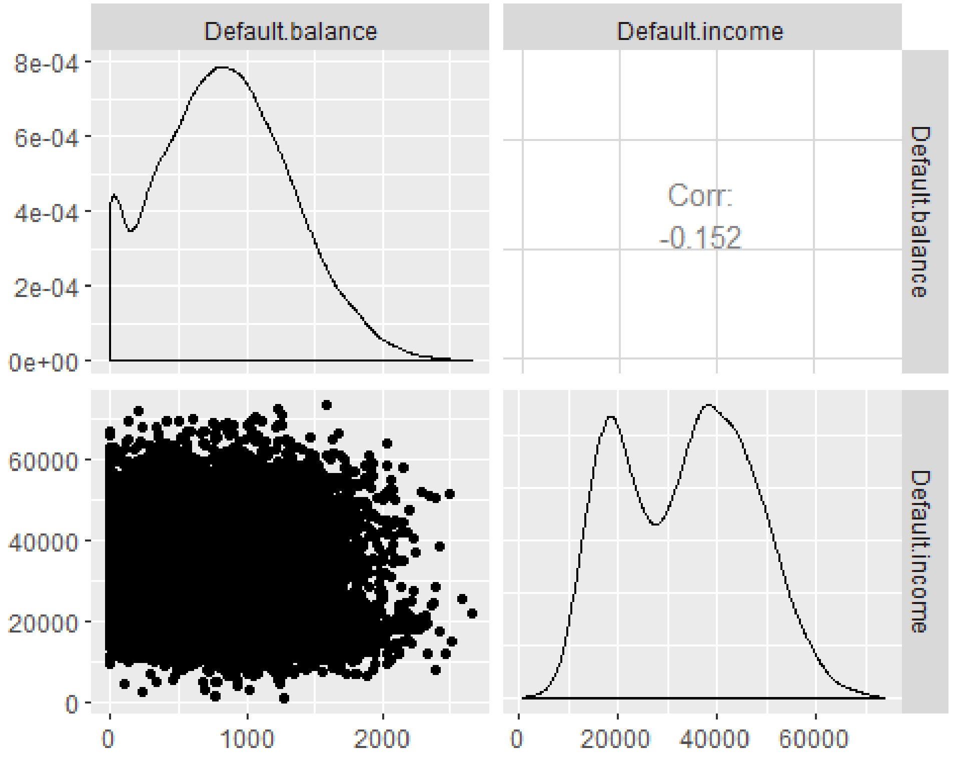
```
library(GGally)

## Loading required package: ggplot2

library(ISLR)
#Multicolinearity Assumption
pairs(~balance+income, data=Default)
```



```
defaultdata <-data.frame(Default$balance,Default$income)
ggpairs(defaultdata)
```



```r
library(mctest)
imcdiag(defaultdata,as.numeric(Default$default), method="VIF")
```

```
##
## Call:
## imcdiag(x = defaultdata, y = as.numeric(Default$default), method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                    VIF detection
## Default.balance 1.0237         0
## Default.income  1.0237         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```