

**Regresi Logistic**

**Modul 3 Praktikum**

**Statistika Sains Data**

**Program Studi Sains Data**

**Fakultas Sains**

**Institut Teknologi Sumatera**

**2024**

**Eksplorasi Data**

**Modul Praktikum 1**

**Statistika Sains Data**

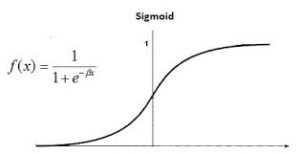
**Eksplorasi Data**

**Modul Praktikum 1**

**Statistika Sains Data**

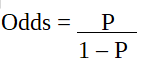
1. **Tujuan Praktikum**
2. Mahasiswa mampu menaksir model regresi logistic menggunakan software RStudio.
3. Mahasiswa mampu menentukan kualitas dari model regresi logistic yang terbentuk.
4. **Teori Dasar**

Regresi logistik dalam [Pemrograman R](https://www.geeksforgeeks.org/introduction-to-r-programming-language/) adalah algoritma klasifikasi yang digunakan untuk mencari kemungkinan keberhasilan suatu peristiwa dan kegagalan suatu peristiwa. Regresi logistik digunakan jika variabel terikatnya bersifat biner (0/1, Benar/Salah, Ya/Tidak). Fungsi logit digunakan sebagai fungsi link dalam distribusi binomial.  Probabilitas variabel hasil biner dapat diprediksi menggunakan teknik pemodelan statistik yang dikenal sebagai regresi logistik. Ini banyak digunakan di berbagai industri, termasuk pemasaran, keuangan, ilmu sosial, dan penelitian medis. Fungsi logistik, biasa disebut fungsi sigmoid, merupakan ide dasar yang mendasari regresi logistik. Fungsi sigmoid ini digunakan dalam regresi logistik untuk menggambarkan korelasi antara variabel prediktor dan kemungkinan hasil biner.



Regresi logistik juga dikenal sebagai regresi logistik binomial . Hal ini didasarkan pada fungsi sigmoid di mana keluaran adalah probabilitas dan masukan dapat berkisar dari -infinity hingga +infinity.

Regresi logistik juga dikenal sebagai model linier umum. Karena digunakan sebagai teknik klasifikasi untuk memprediksi respons kualitatif, Nilai y berkisar antara 0 hingga 1 dan dapat direpresentasikan dengan persamaan berikut:



p adalah probabilitas karakteristik yang diinginkan. Rasio odds didefinisikan sebagai probabilitas keberhasilan dibandingkan dengan probabilitas kegagalan. Ini adalah representasi utama dari koefisien regresi logistik dan dapat mengambil nilai antara 0 dan tak terhingga. Rasio odds 1 adalah ketika probabilitas keberhasilan sama dengan probabilitas kegagalan. Rasio odds 2 adalah ketika peluang sukses dua kali lipat peluang gagal. Rasio odds 0,5 adalah ketika peluang kegagalan dua kali lipat peluang sukses.



Karena kita bekerja dengan distribusi binomial (variabel terikat), kita perlu memilih fungsi tautan yang paling cocok untuk distribusi ini.



Ini adalah fungsi logit . Dalam persamaan di atas, tanda kurung dipilih untuk memaksimalkan kemungkinan mengamati nilai sampel daripada meminimalkan jumlah kesalahan kuadrat (seperti regresi biasa). Logit juga dikenal sebagai log peluang. Fungsi logit harus berhubungan linier dengan variabel bebas. Ini dari persamaan A, dimana ruas kirinya merupakan kombinasi linier dari x. Hal ini mirip dengan asumsi OLS bahwa y berhubungan linier dengan x. Variabel b0, b1, b2…dst tidak diketahui dan harus diperkirakan berdasarkan data pelatihan yang tersedia. Dalam model regresi logistik, mengalikan b1 dengan satu unit akan mengubah logit sebesar b0. Perubahan P akibat perubahan satu satuan akan bergantung pada nilai yang dikalikan. Jika b1 positif maka P bertambah dan jika b1 negatif maka P berkurang.

1. **Latihan Praktikum**

# Pendahuluan

## C.1 Package

Berikut adalah packages yang akan digunakan dalam praktikum ini, yaitu :

| **Nama Package** | **Kegunaan** |
| --- | --- |
| broom | untuk merapikan tampilan data |
| caret | akan digunakan beberapa fungsi untuk membentuk Confusion Matrix |
| DataExplorer | akan digunakan beberapa fungsi seperti plot\_bar, plot\_histogram, plot\_boxplot, untuk membantu visualisasi data |
| grid | diperlukan agar user defined function arrange bisa digunakan |
| ISLR | Sumber Datasets default |
| pscl | akan digunakan fungsi pR2 (Pseudo R2) untuk Evaluasi Model |
| tidyverse |  |

Untuk melakukan Instalasi, Anda bisa menggunakan syntax berikut ini :

install.packages(c("broom","caret","DataExplorer","grid","ISLR","pscl","tidyverse"))

Jangan lupa untuk load package tersebut.

library(broom)

library(caret)

library(DataExplorer)

library(grid)

library(InformationValue)

library(ISLR)

library(pscl)

library(tidyverse)

Sebenarnya Anda tidak perlu untuk load keseluruhan pakcage, Anda bisa menggunakan syntax namapackage::fungsi untuk menggunakan suatu fungsi dari suatu package di R, selama package tersebut sudah Anda Install.

## C.1.2 Data

Data yang akan digunakan dalam praktikum ini ada 2, yaitu :  
- Default Datasets yang berasal dari package ISLR  
- Lending Club Datasets

## C.1.3 Tambahan

Untuk mempermudah visualisasi beberapa bagian dalam Praktikum ini, digunakan fungsi arrange yang dibuat oleh Stephen Turner

vp.layout <- **function**(x, y) viewport(layout.pos.row=x, layout.pos.col=y)

arrange <- **function**(..., nrow=NULL, ncol=NULL, as.table=FALSE) {

dots <- list(...)

n <- length(dots)

**if**(is.null(nrow) & is.null(ncol)) { nrow = floor(n/2) ; ncol = ceiling(n/nrow)}

**if**(is.null(nrow)) { nrow = ceiling(n/ncol)}

**if**(is.null(ncol)) { ncol = ceiling(n/nrow)}

## **NOTE** see n2mfrow in grDevices for possible alternative

grid.newpage()

pushViewport(viewport(layout=grid.layout(nrow,ncol) ) )

ii.p <- 1

**for**(ii.row **in** seq(1, nrow)){

ii.table.row <- ii.row

**if**(as.table) {ii.table.row <- nrow - ii.table.row + 1}

**for**(ii.col **in** seq(1, ncol)){

ii.table <- ii.p

**if**(ii.p > n) **break**

print(dots[[ii.table]], vp=vp.layout(ii.table.row, ii.col))

ii.p <- ii.p + 1

}

}

}

# C.2  Default Datasets

## C.2.1 Eksplorasi Data

data(Default)

Untuk mengetahui lebih lanjut mengenai Default, Anda bisa menggunakan syntax berikut :

help(Default)

**Deskripsi**

*A simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt.*

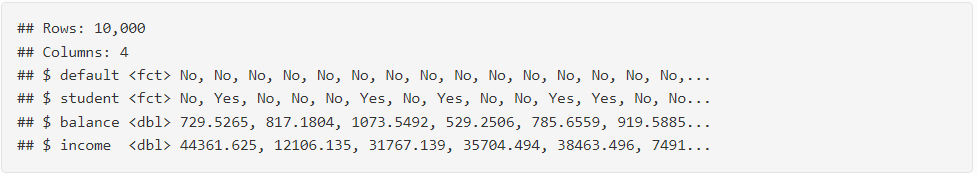
**Format**

*A data frame with 10000 observations on the following 4 variables.*

* default : A factor with levels No and Yes indicating whether the customer defaulted on their debt
* student : A factor with levels No and Yes indicating whether the customer is a student
* balance : The average balance that the customer has remaining on their credit card after making their monthly payment
* income : Income of customer

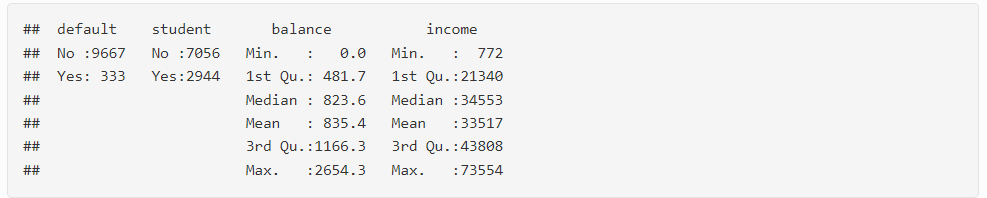
**glimpse**

Anda bisa melihatnya dengan glimpse dari dplyr yang merupakan bagian dari tidyverse

****glimpse(Default)

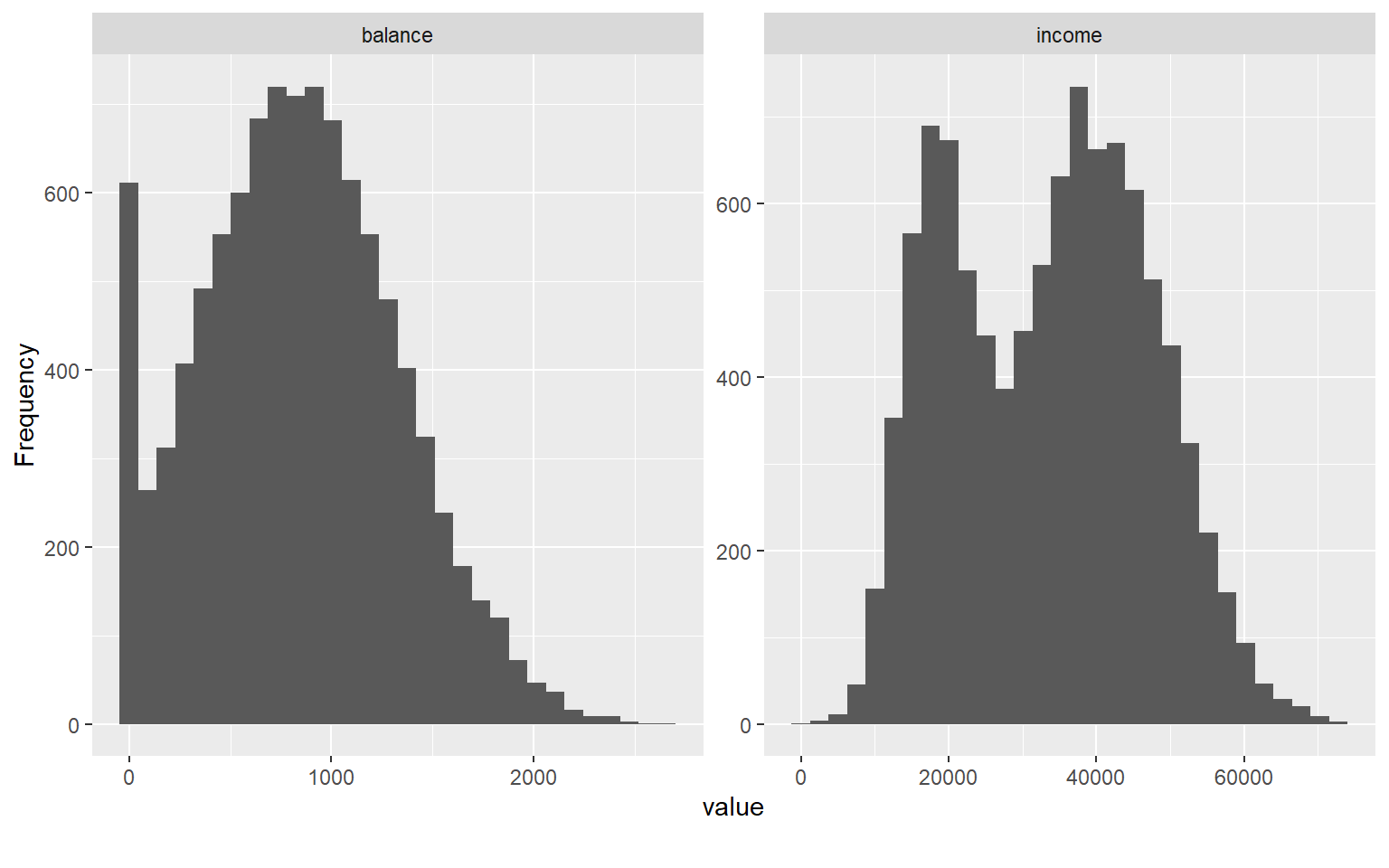
**summary**

Seandainya kita ingin mengetahui sebaran dan kecenderungan pola hubungan antar peubah, salah satu yang dapat dilakukan adalah sebagai berikut.

****summary(Default)

**plot\_histogram**

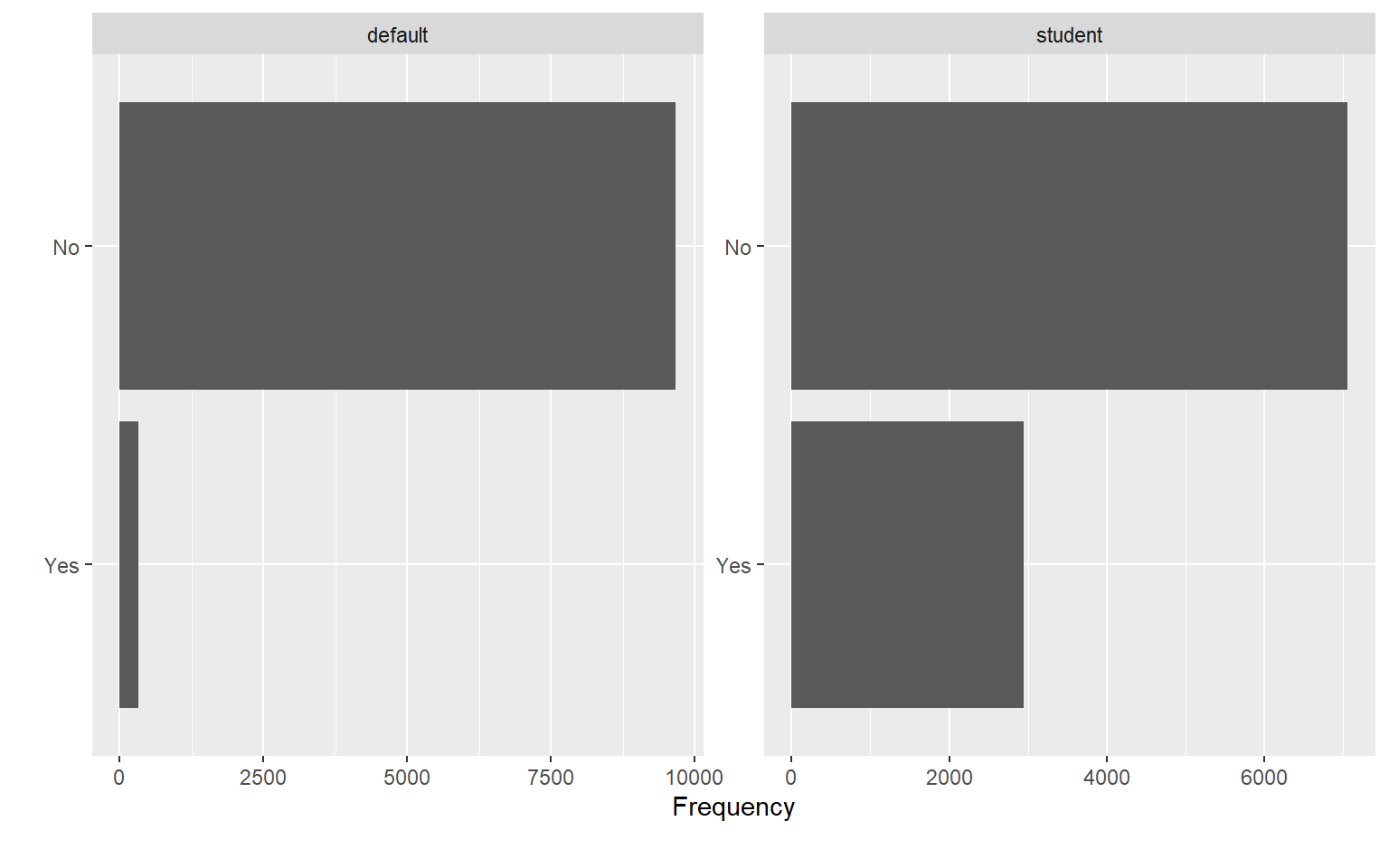
plot\_histogram(Default)



plot\_histogram ini menampilkan peubah dengan tipe numerik.

**plot\_bar**

plot\_bar(Default)

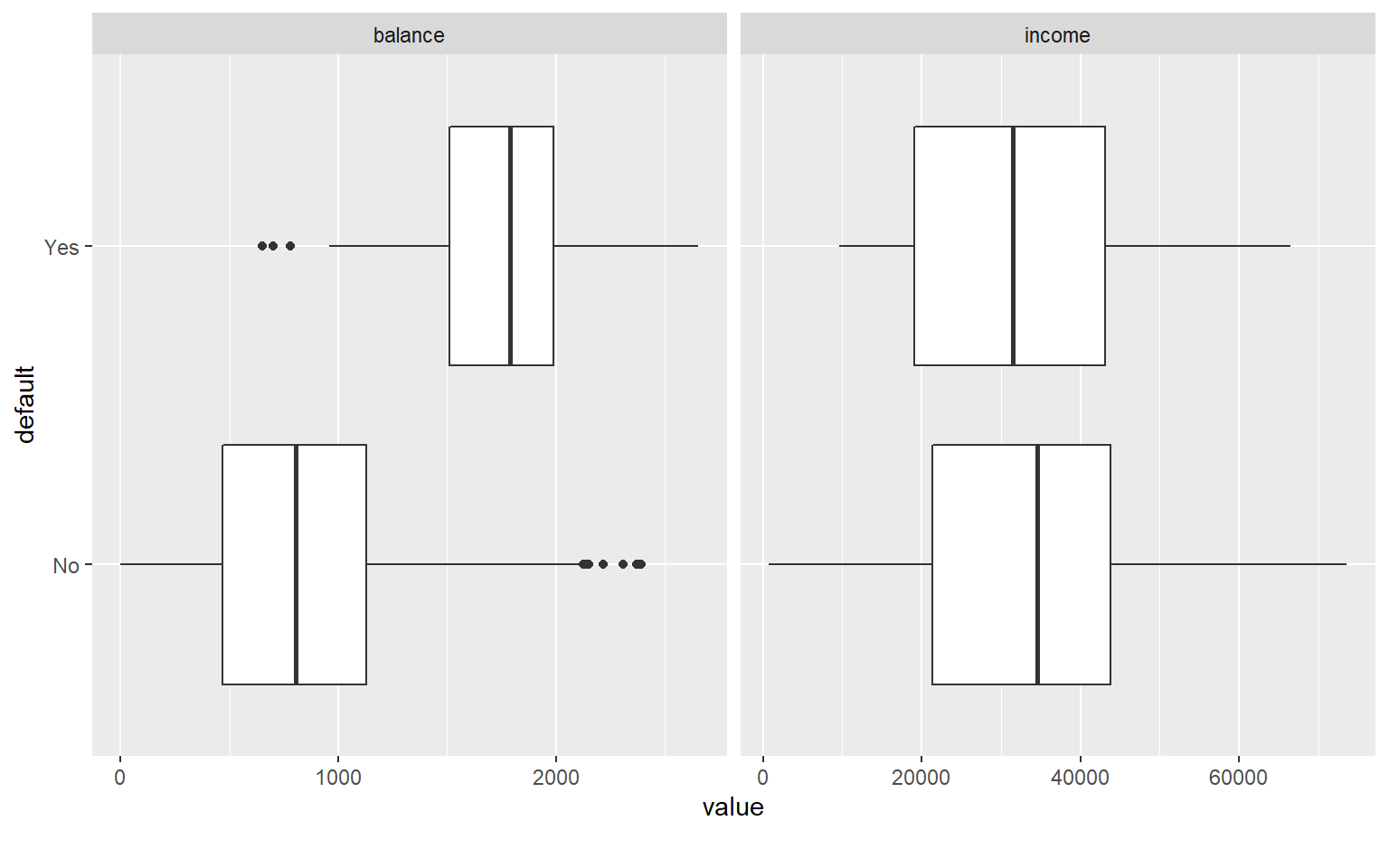


plot\_bar ini menampilkan peubah dengan tipe kategorik.

**plot\_boxplot**

Untuk melihat pola hubungan antara peubah numerik dan kategorik dapat lebih mudah diidentifikasi menggunakan plot\_boxplot.

plot\_boxplot(Default,by="default")



Output di atas memperlihatkan bahwa median pendapatan antara nasabah yang mengalami kredit macet atau tidak macet cederung mirip. Sebaliknya, terdapat perbedaan jumlah tagihan kartu kredit pada nasabah dengan kredit macet dan yang tidak. Pada kasus ini, orang yang memiliki tagihan lebih banyak, cenderung lebih mengalami kredit macet.

**student - balance**

a1<-Default %>% filter(student=="Yes") %>%

select(balance,default) %>%

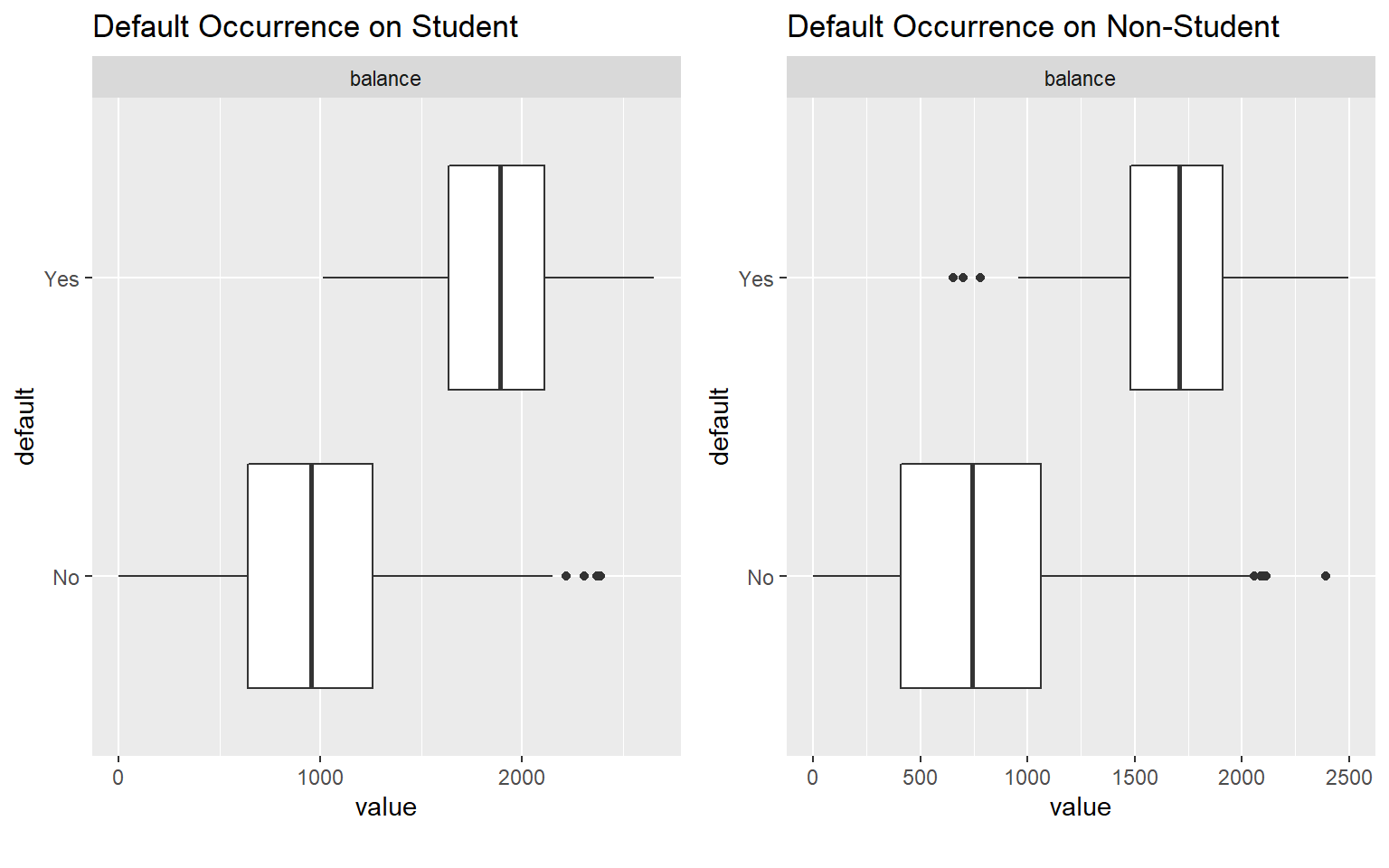
plot\_boxplot(by="default", title="Default Occurrence on Student")

a2<-Default %>% filter(student=="No") %>%

select(balance,default) %>%

plot\_boxplot(by="default", title="Default Occurrence on Non-Student")

arrange(a1,a2)



## $page\_1

##

## $page\_1

**student - income**

b1<-Default %>% filter(student=="Yes") %>%

select(income,default) %>%

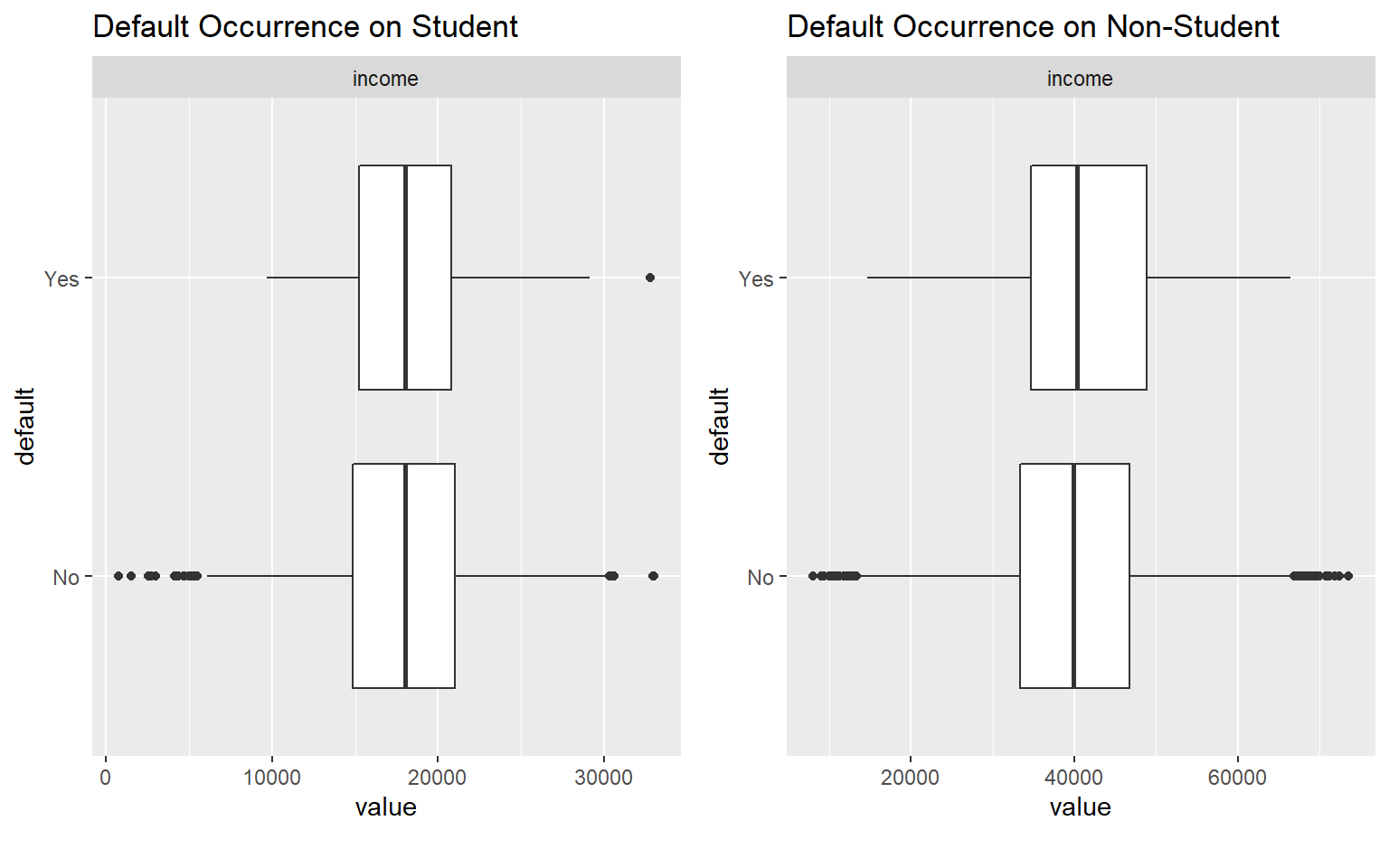
plot\_boxplot(by="default", title="Default Occurrence on Student")

b2<-Default %>% filter(student=="No") %>%

select(income,default) %>%

plot\_boxplot(by="default", title="Default Occurrence on Non-Student")

arrange(b1,b2)



## $page\_1

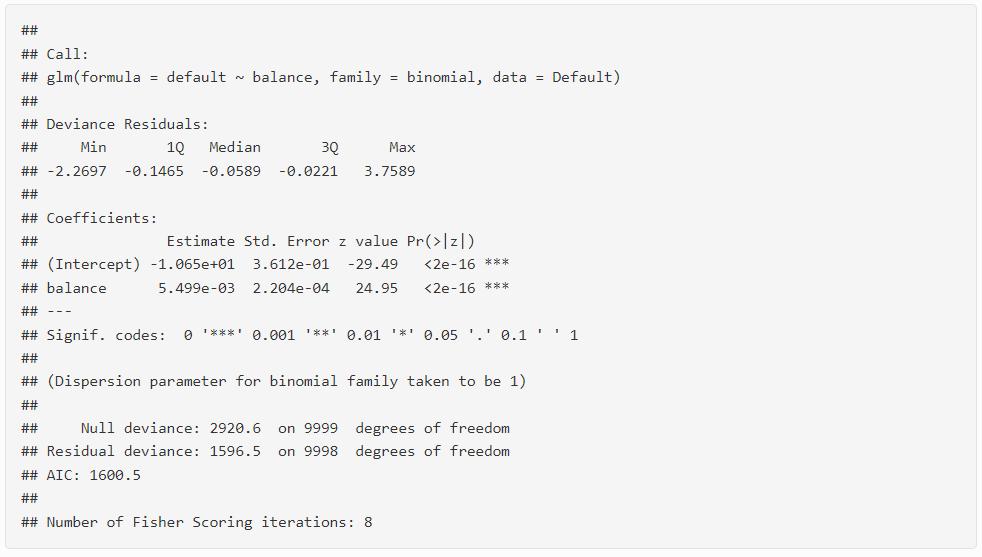
##

## $page\_1

## C.3 The Model, Estimation and Inference

Sebagai awal dari ilustrasi yang akan dibahas pada modul kali ini, akan diperlihatkan pemodelan regresi logistik dengan menggunakan satu peubah penjelas terlebih dulu.

reglog<-glm(default~balance,data=Default,family=binomial)

summary (reglog)

Default %>%

mutate(prob = ifelse(default == "Yes", 1, 0)) %>%

ggplot(aes(balance, prob)) +

geom\_point(alpha = .15) +

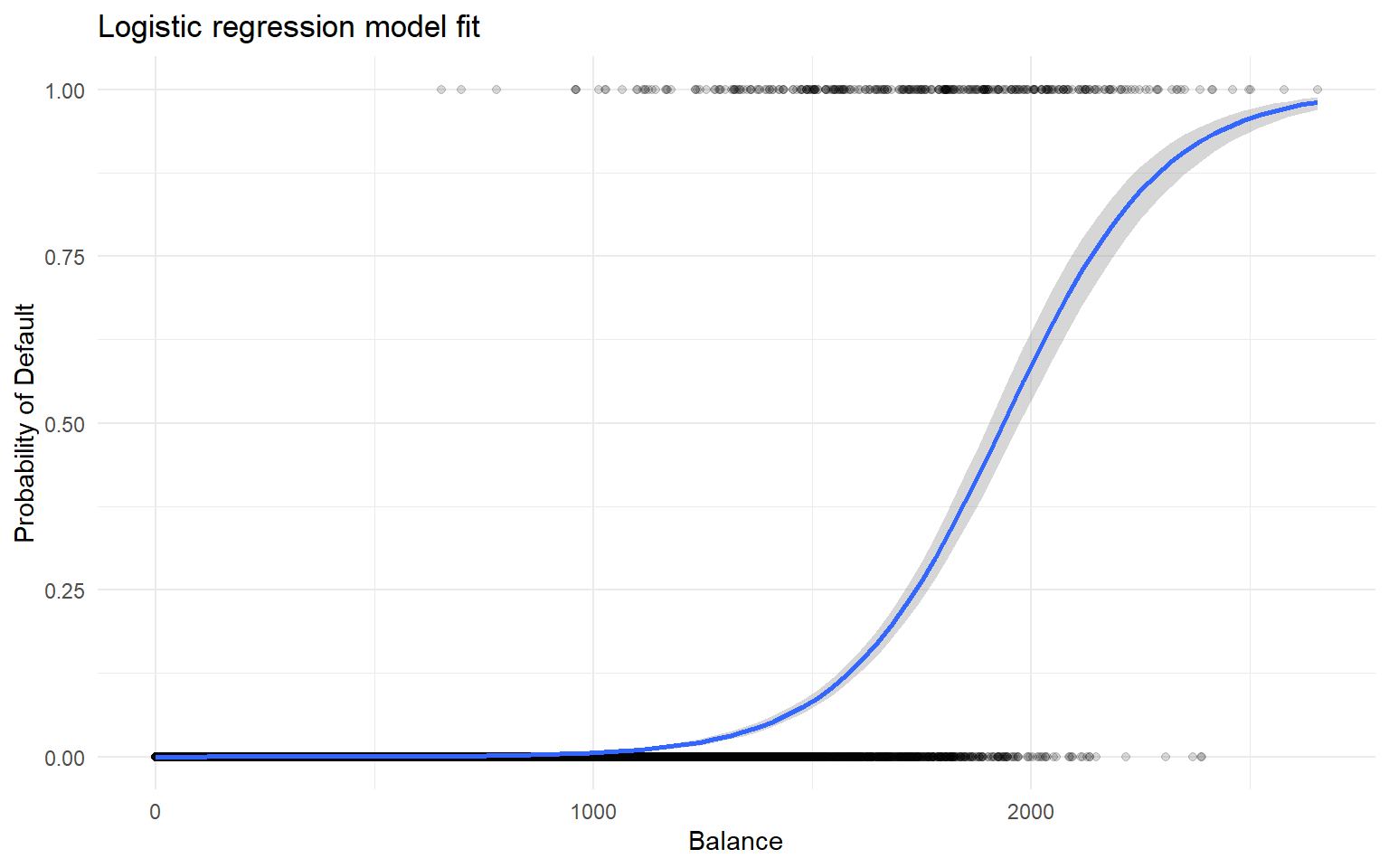
geom\_smooth(formula=y~x, method = "glm", method.args = list(family = "binomial")) +

ggtitle("Logistic regression model fit") +

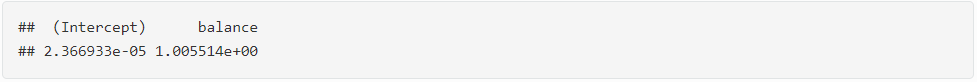
xlab("Balance") +

ylab("Probability of Default")+

theme\_minimal()



### B.2.1 Interpreting coefficients

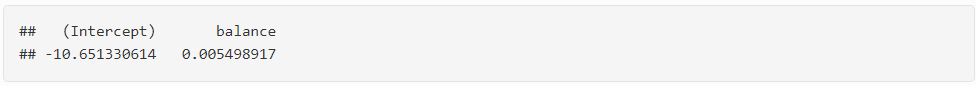
reglog$coefficients

Seperti yang telah dibahas pada kelas perkuliahan,

*hasil*β1β1*=0.0055 (positif) berarti membesarnya tagihan akan meningkatkan peluang gagal bayar.*

*Pengguna kartu kredit yang lebih besar tagihannya memiliki risiko 0.0055 kali lebih besar untuk gagal bayar. Lebih tepat lagi, peningkatan satu unit tagihan (dollar) berhubungan dengan meningkatnya log-odds gagal bayar sebesar 0.0055 unit.*

Untuk memudahkan interpretasi, seringkali digunakan rasio odds, yaitu rasio antara dua odds.

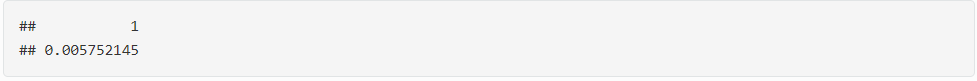
exp(reglog$coefficients)

Pada ilustrasi ini, terlihat bahwa nilai rasio odds untuk peubah balance adalah 1.00551. Hal ini dapat diartikan bahwa ketika tagihan kartu kredit meningkat $ 1 maka kemungkinan kejadian kredit macet meningkat sebesar 1.00551 kali dibadingkan tagihannya tetap

confint(reglog)

### B.2.2 Prediction

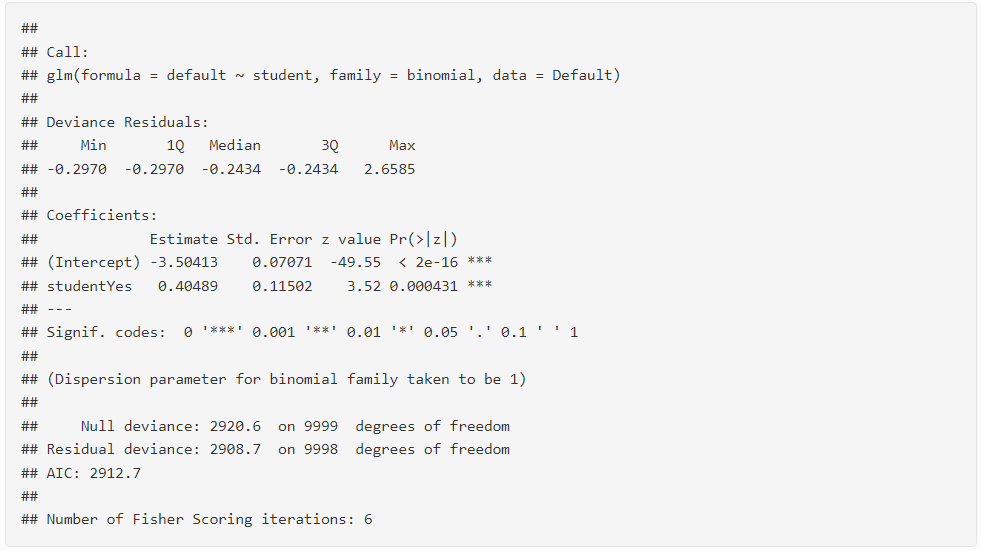
Seandainya kita ingin memprediksi peluang kredit macet jika besarnya tagihan seorang nasabah adalah sebesar $ 1,000

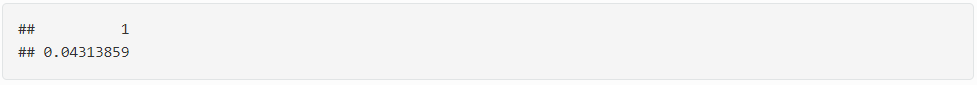
predict(reglog, newdata = data.frame(balance=1000), type="response")

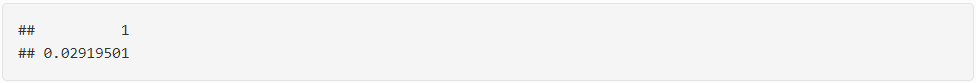
Artinya, seseorang yang memiliki jumlah tagihan $ 1,000 diprediksi memiliki peluang kredit macet sebesar 0.005752145.

### B.2.3 Categorical Independent Variable

reglog2<-glm(default~student,data=Default,family=binomial)

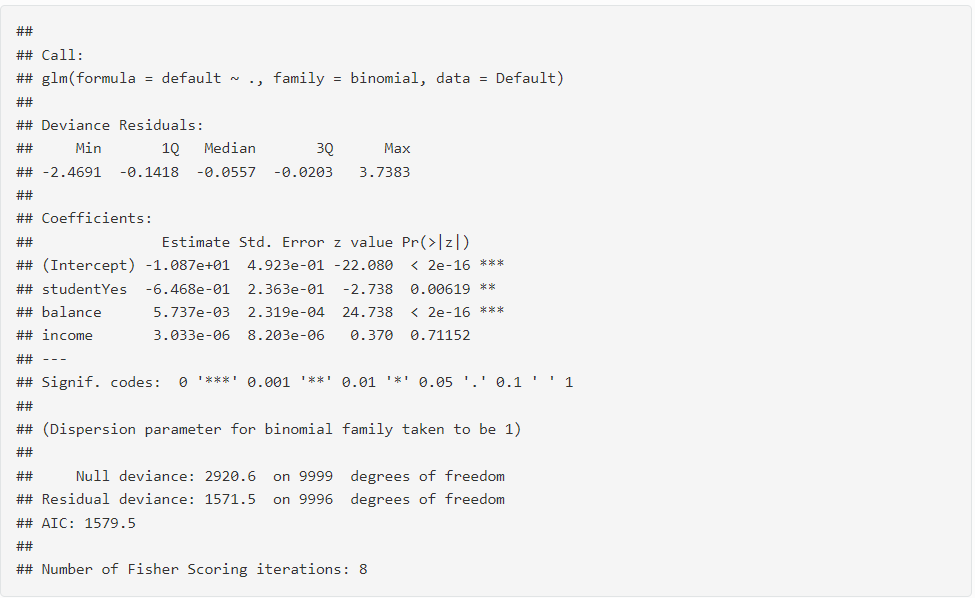
summary(reglog2)

predict(reglog2, newdata = data.frame(student="Yes"), type="res")

predict(reglog2, newdata = data.frame(student="No"), type="res")

### B.2.4 Multiple Logistic Regression

full.mod = glm(default~., data=Default,family=binomial)

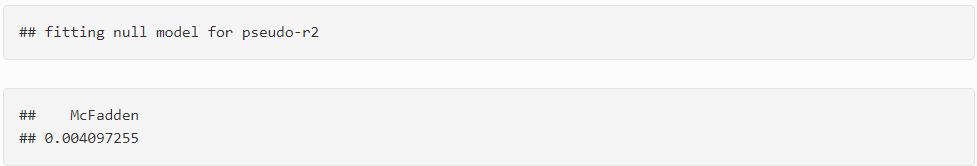
summary(full.mod)

### Perhatikan bahwa pola pengaruh peubah penjelas tidak sama seperti yang dihasilkan pada model sebelumnya yang hanya menggunakan satu peubah penjelas. Dimana sebelumnya studentYes bernilai positif.

## B.3 Model Evaluation

### B.3.1 Pseudo **R2**

pscl::pR2(reglog)["McFadden"]

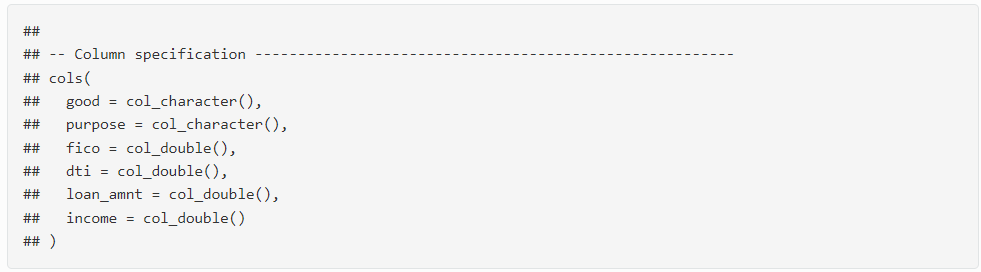
pscl::pR2(reglog2)["McFadden"]

pscl::pR2(full.mod)["McFadden"]

# C. Lending Club Datasets

## C.1 Eksplorasi Data

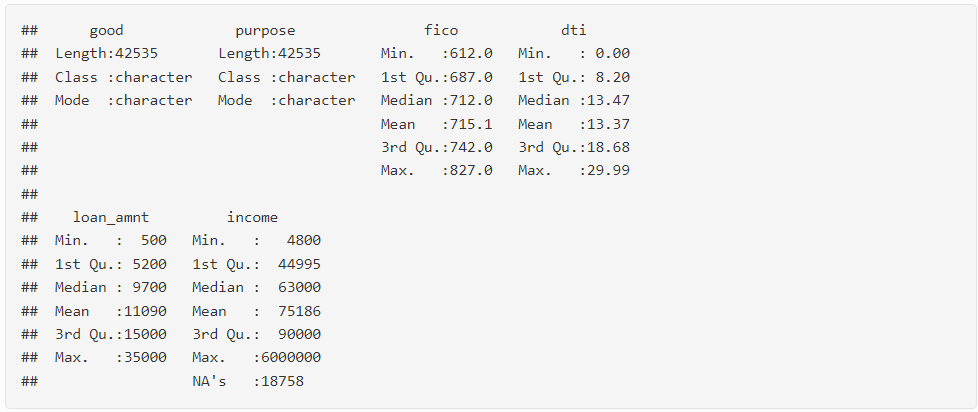
Pada ilustrasi ini akan digunakan data Lending Club Data.

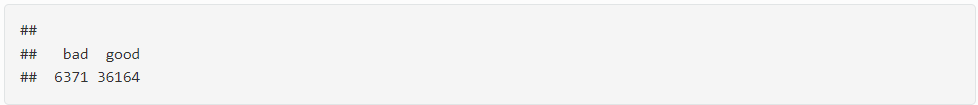
loan <- read\_csv("https://www.dropbox.com/s/89g1yyhwpcqwjn9/lending\_club\_cleaned.csv?raw=1")

* Peubah good merupakan indikator apakah pinjaman tersebut termasuk baik (good) atau buruk (bad);
* Peubah fico merupakan skor kredit dimana skor yang lebih tinggi adalah yang lebih baik;
* Peubah dti menunjukkan rasio debt-to-income,
* Peubah loan\_amnt menunjukkan jumlah pinjaman dan
* Peubah income menunjukkan pendapatan. Sebagai catatan, peubah income bernilai NA jika sumber pendapatan tidak terverifikasi.

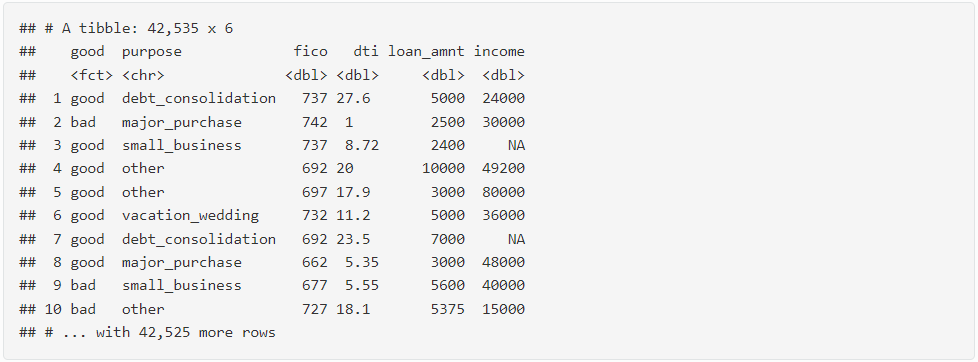
Loan

## 

summary(loan)

table(loan$good)

loan$good<-as.factor(loan$good)

loan

## C.2 Membagi Data

Untuk kepentingan prediksi model, data akan dibagi dua menjadi data train dan data test. Karena pembagian data ini menggunakan sampling, maka perlu dilakukan pengaturan seed untuk randomisasi.

set.seed(5815)

sample <- sample(nrow(loan),floor(nrow(loan)\*0.8))

train <- loan[sample,]

test <- loan[-sample,]

Sehingga, kita mempunyai dua data, train untuk membentuk model, dan test untuk menguji model yang telah dibuat.

## C.3 Model

Pada ilustrasi ini, peubah fico, dti, loan\_amnt, dan purpose akan digunakan sebagai peubah penjelas

logit <- glm(good ~ fico + dti+ loan\_amnt + purpose, data = train, family = "binomial")

summary(logit)

## 

## C.4 Evaluasi Model

## C.4.1 Confusion Matrix

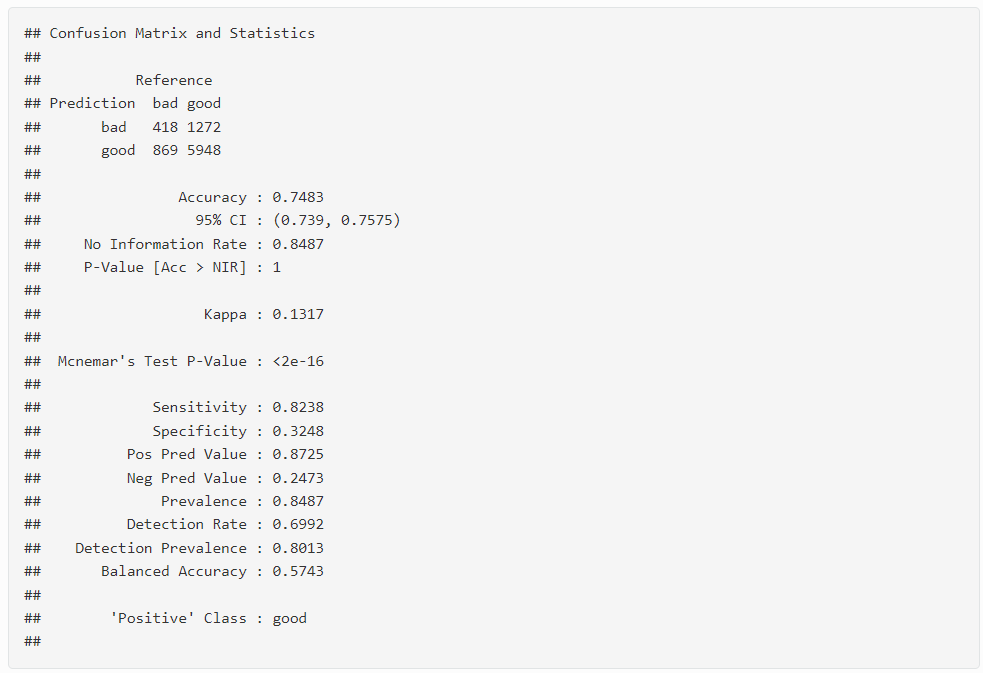
Sensitivitas (atau True Positive Rate) adalah persentase pengamatan (aktual) yang diprediksi dengan benar oleh model, sedangkan spesifisitas adalah persentase dari 0 (aktual) yang diprediksi dengan benar.

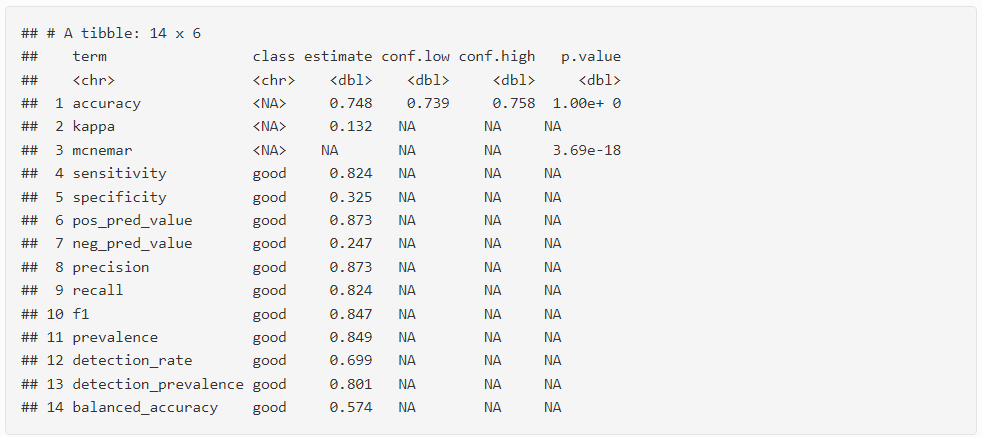
test$pred <- predict(logit, test, type="response")

test$good\_pred <- ifelse(test$pred > 0.80, "good", "bad")

test$good\_pred <- as.factor(test$good\_pred)

(conf.mat<-caret::confusionMatrix(test$good\_pred, test$good, positive="good"))

****

****broom::tidy(conf.mat)