Nama : Ibnu Fajar Setiawan **NIM :** 065002000006	UNIVERSITAS TRISAKTI **Praktikum Data Analitik**	**Modul 7** **Nama Dosen: Syandra Sari, S.Kom, M.Kom**
Hari/Tanggal : Kamis, 10 November 2022		**Nama Aslab :** **1. Azzahra Nuranisa (065001900044)** 2. **Ida Jubaidah (065001900037)**

Praktikum 7 – REGRESI LOGISTIK

DESKRIPSI MODUL : Melakukan pengujian Regresi Logistik

No	Elemen Kompetensi	Indikator Kinerja	Jml Jam	hlm
1	Mampu melakukan pengujian regresi Logistik	Mampu melakukan pengujian regresi Logistik	2	

TEORI SINGKAT

Berbeda dengan regresi lininer yang digunakan untuk memprediksi variabel Y yang bersifat kontinu, regresi logistic digunakan untuk memprediksi variabel Y yang bersifat kategorik. Kasus regresi logistic dengan Y yang terdiri dari hanya dua kelas dinamakan *binary classification problems (**binomial logistic regression)***. Prediktor dapat bersifat kontinu, kategorik maupun gabungan keduanya.

ELEMEN KOMPETENSI I

Deskripsi : Dapat melakukan pengujian regresi Logistik

Kompetensi Dasar : Mampu melakukan pengujian regresi Logistik

Kasus 1 : Titanic Dataset

Gunakan data berikut ini untuk membangun model prediktif. Berikan interpretasi atas setiap output yang dihasilkan. Mulailah analisis dengan membuat tabulasi silang setiap predictor yang bersifat kategorik dengan respon (Y).

DATA DICTIONARY

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

REGRESI LOGISTIK

```
> databaru=read.delim("clipboard")

> str(databaru)

>  sampel1<-sample(1:nrow(databaru),0.75*nrow(databaru))
>  traininglogistik<-data.frame(databaru)[sampel1,]
>  testinglogistik<-data.frame(databaru)[-sampel1,]
>  modellogistik=glm(Survived~.,data=traininglogistik,family = binomial)
>  summary(modellogistik)
```

```
> head(dataibnu)
  PassengerId Survived Pclass    Sex Age SibSp Parch    Fare
1           1        0      3   male  22     1     0   7.250
2           2        1      1 female  38     1     0 712.833
3           3        1      3 female  26     0     0   7.925
4           4        1      1 female  35     1     0  53.100
5           5        0      3   male  35     0     0   8.050
6           7        0      1   male  54     0     0 518.625
> dataibnu$Fare = as.numeric(as.character(dataibnu$Fare))
> str(dataibnu)
'data.frame':   680 obs. of  8 variables:
 $ PassengerId: int  1 2 3 4 5 7 8 9 10 11 ...
 $ Survived   : int  0 1 1 1 0 0 0 1 1 1 ...
 $ Pclass     : int  3 1 3 1 3 1 3 3 2 3 ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 54 2 27 14 4 ...
 $ SibSp      : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch      : int  0 0 0 0 0 0 1 2 0 1 ...
 $ Fare       : num  7.25 712.83 7.92 53.1 8.05 ...
> sampel1<-sample(1:nrow(dataibnu),0.75*nrow(dataibnu))
> traininglogistik<-data.frame(dataibnu)[sampel1,]
> testinglogistik<-data.frame(dataibnu)[-sampel1,]
> modellogistik=glm(Survived~.,data=traininglogistik,family = binomial)
> summary(modellogistik)

Call:
glm(formula = Survived ~ ., family = binomial, data = traininglogistik)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7608  -0.6448  -0.3754   0.6483   2.5444

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.0927622  0.7040583   7.233 4.71e-13 ***
PassengerId  0.0008389  0.0004844   1.732  0.08330 .
Pclass      -1.2482175  0.1803344  -6.922 4.46e-12 ***
Sexmale     -2.6332211  0.2716164  -9.695  < 2e-16 ***
Age         -0.0492221  0.0099593  -4.942 7.72e-07 ***
SibSp       -0.3374033  0.1480474  -2.279  0.02267 *
Parch       -0.1122407  0.1434612  -0.782  0.43399
Fare         0.0023378  0.0008321   2.810  0.00496 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 663.77  on 494  degrees of freedom
Residual deviance: 437.33  on 487  degrees of freedom
  (15 observations deleted due to missingness)
AIC: 453.33

Number of Fisher Scoring iterations: 5
```

#MELAKUKAN PREDIKSI

```
> prediksilogistik=predict(modellogistik,testinglogistik)
> pred_logreg<-as.numeric(prediksilogistik>.5)
> tabel_logreg<-table(pred_logreg,testinglogistik$Survived)
```

```
> tabel_logreg
> prediksilogistik=predict(modellogistik,testinglogistik)
> pred_logreg<-as.numeric(prediksilogistik>.5)
> tabel_logreg<-table(pred_logreg,testinglogistik$Survived)
> tabel_logreg

pred_logreg  0  1
          0 96 22
          1  6 42
> confusionMatrix(table(pred_logreg,testinglogistik$Survived),positive = "1")
Confusion Matrix and Statistics


pred_logreg  0  1
          0 96 22
          1  6 42

               Accuracy : 0.8313
                 95% CI : (0.7655, 0.8849)
    No Information Rate : 0.6145
    P-Value [Acc > NIR] : 1.039e-09

                  Kappa : 0.6266

 Mcnemar's Test P-Value : 0.004586

            Sensitivity : 0.6562
            Specificity : 0.9412
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.8136
             Prevalence : 0.3855
         Detection Rate : 0.2530
   Detection Prevalence : 0.2892
      Balanced Accuracy : 0.7987

       'Positive' Class : 1
```

ELEMEN KOMPETENSI II

Deskripsi : Dapat melakukan pengujian regresi Logistik dengan dataset Iris

Kompetensi Dasar : Mampu melakukan pengujian regresi Logistik dengan Dataset Iris

Kasus 2. The iris data set (*species virginica and versicolor only)*

Gunakan data berikut ini untuk membangun model prediktif. Berikan interpretasi atas setiap output yang dihasilkan. Mulailah analisis dengan membuat tabulasi silang setiap predictor yang bersifat kategorik dengan respon (Y).

make a reduced iris data set that only contains virginica and versicolor species

```
> library(dplyr)

> iris.small <- filter(iris, Species %in% c("virginica", "versicolor"))
```

logistic regression

```
> glm.out <- glm(Species ~ Sepal.Width + Sepal.Length + Petal.Width + Petal.Length,
+            data = iris.small,
+            family = binomial) # family = binomial required for logistic regression
> summary(glm.out)

> exp(coef(glm.out))



> glm.out <- glm(Species ~ Sepal.Width + Petal.Width + Petal.Length,
+            data = iris.small,
+            family = binomial)

> exp(coef(glm.out))
```

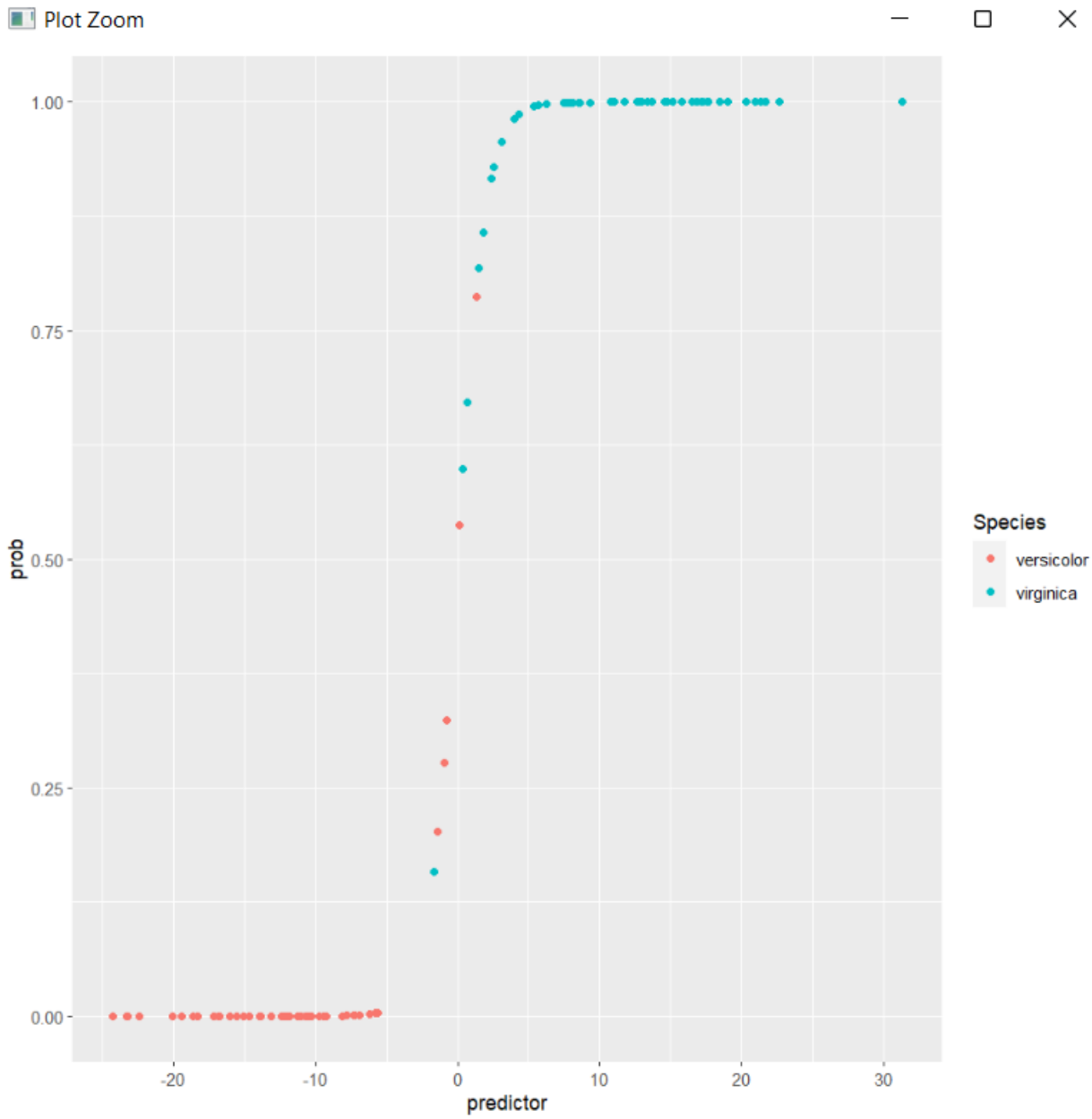
```
> library(dplyr)
> iris.small <- filter(iris, Species %in% c("virginica", "versicolor"))
> glm.out <- glm(Species ~ Sepal.Width + Sepal.Length + Petal.Width + Petal.Length,data = iris.small,family = binomial) # f
amily = binomial required for logistic regression
> summary(glm.out)

Call:
glm(formula = Species ~ Sepal.Width + Sepal.Length + Petal.Width +
    Petal.Length, family = binomial, data = iris.small)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01105  -0.00541  -0.00001   0.00677   1.78065

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -42.638     25.707  -1.659   0.0972 .
Sepal.Width    -6.681      4.480  -1.491   0.1359
Sepal.Length   -2.465      2.394  -1.030   0.3032
Petal.Width    18.286      9.743   1.877   0.0605 .
Petal.Length    9.429      4.737   1.991   0.0465 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  11.899  on 95  degrees of freedom
AIC: 21.899

Number of Fisher Scoring iterations: 10
```

```
> exp(coef(glm.out))
 (Intercept)  Sepal.Width Sepal.Length  Petal.Width Petal.Length
3.038345e-19 1.254665e-03 8.499013e-02 8.741145e+07 1.244887e+04
> glm.out <- glm(Species ~ Sepal.Width + Petal.Width + Petal.Length,data = iris.small,family = binomial)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> exp(coef(glm.out))
 (Intercept)  Sepal.Width  Petal.Width Petal.Length
1.138872e-22 2.303132e-04 2.026532e+09 2.629495e+03
```

TUGAS

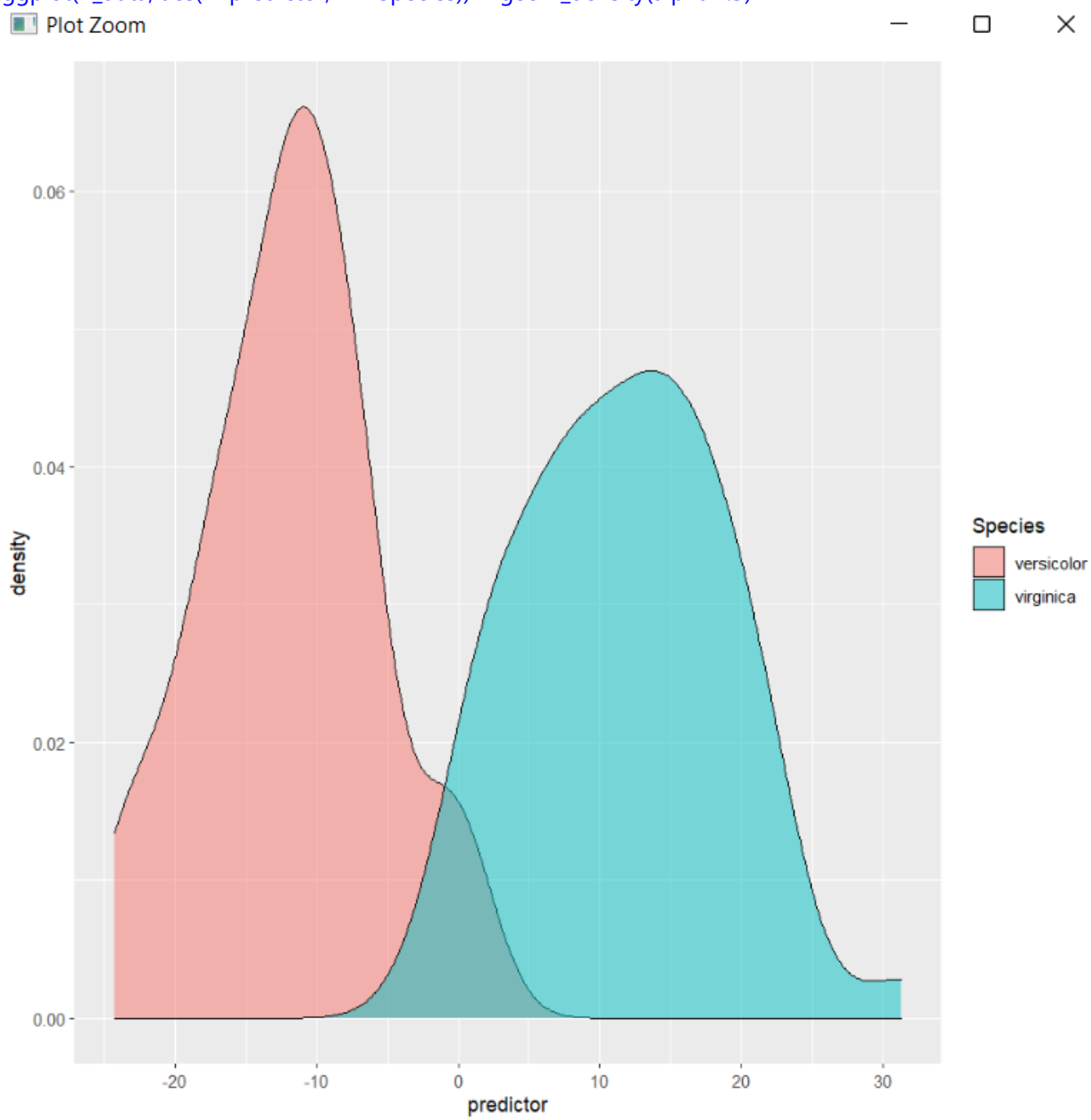
Make a plot of the fitted probability as a function of the linear predictor, colored by species identity. Hint: you will have to make a new data frame combining data from the fitted model with data from the `iris.small` *data frame.*

```
> lr_data <- data.frame(predictor=glm.out$linear.predictors, prob=glm.out$fitted.values,
Species=iris.small$Species)
> ggplot(lr_data, aes(x=predictor, y=prob, color=Species)) + geom_point()
```



Make a density plot that shows how the two species are separated by the linear predictor.

```
ggplot(lr_data, aes(x=predictor, fill=Species)) + geom_density(alpha=.5)
```



Assume you have obtained samples from three plants, with measurements as listed below. Predict the likelihood that each of these plants belongs to the species virginica.

```
> plant1 <- data.frame(Sepal.Length=6.4, Sepal.Width=2.8, Petal.Length=4.6, Petal.Width=1.8)
> plant2 <- data.frame(Sepal.Length=6.3, Sepal.Width=2.5, Petal.Length=4.1, Petal.Width=1.7)
> plant3 <- data.frame(Sepal.Length=6.7, Sepal.Width=3.3, Petal.Length=5.2, Petal.Width=2.3)
> predict(glm.out, plant1, type="response")
> predict(glm.out, plant2, type="response")
> predict(glm.out, plant3, type="response")
```

```
> predict(glm.out, plant1, type="response")
        1
0.6934611
> predict(glm.out, plant2, type="response")
         1
0.06002675
> predict(glm.out, plant3, type="response")
        1
0.9999943
>
```

Pick a cutoff predictor value at which you would decide that a specimen belongs to virginica rather than versicolor. Calculate how many virginicas you call correctly and how many incorrectly given that choice.

```
> cutoff <- 0
> virg_true <- sum(lr_data$predictor > cutoff & lr_data$Species=="virginica")
> virg_false <- sum(lr_data$predictor <= cutoff & lr_data$Species=="virginica")
> virg_true
> virg_false
```

```
> cutoff <- 0
> virg_true <- sum(lr_data$predictor > cutoff & lr_data$Species=="virginica")
> virg_false <- sum(lr_data$predictor <= cutoff & lr_data$Species=="virginica")
> virg_true
[1] 49
> virg_false
[1] 1
>
```

Now do the same calculation for versicolor.

```
> vers_true <- sum(lr_data$predictor <= cutoff & lr_data$Species=="versicolor")
> vers_false <- sum(lr_data$predictor > cutoff & lr_data$Species=="versicolor")
> vers_true
> vers_false
```

```
> vers_true <- sum(lr_data$predictor <= cutoff & lr_data$Species=="versicolor")
> vers_false <- sum(lr_data$predictor > cutoff & lr_data$Species=="versicolor")
> vers_true
[1] 48
> vers_false
[1] 2
>
```

If we define a call of virginica as a positive and a call of versicolor as a negative, what are the true positive rate (sensitivity, true positives divided by all possible positives) and the true negative rate (specificity, true negatives divided by all possible negatives) in your analysis?

```
> tp <- virg_true/(virg_true + virg_false)
> tn <- vers_true/(vers_true + vers_false)
> tp
> tn
```

```
> tp <- virg_true/(virg_true + virg_false)
> tn <- vers_true/(vers_true + vers_false)
> tp
[1] 0.98
> tn
[1] 0.96
>
>
```

Sumber : **http://wilkelab.org/classes/SDS348/2015_spring_worksheets/class11_solutions.html**

1. **Cek List**

	Elemen Kompetensi	**Penyelesaian**	
		Selesai	**Tidak**
	Elemen Kompetensi I Dapat melakukan pengujian regresi Logistik.	✓	
	Elemen Kompetensi II Dapat melakukan pengujian regresi Logistik dengan dataset Iris.	✓	

2. **Form Umpan Balik**

Elemen Kompetensi	**Waktu Pengerjaan**	**Kriteria**
Elemen Kompetensi I Dapat melakukan pengujian regresi Logistik.	30	1
Elemen Kompetensi II Dapat melakukan pengujian regresi Logistik dengan dataset Iris	30	1

Kriteria

1.Sangat Menarik

2.Cukup Menarik

3.Kurang Menarik

4.Sangat Kurang Menarik