UNIVERSIDADE FEDERAL FLUMINENSE

Programa de Mestrado e Doutorado em Engenharia de Produção

Multivariate Data Analysis

Binary Logistic Regression

Professor: Valdecy Pereira, D. Sc.

email: valdecy.pereira@gmail.com

Outline

1. Definition
2. Logit
3. GOF
4. Interpretation
5. Prediction
6. Bibliography

MVDA – *Binary Logistic Regression*

A regression technique that has a dichotomous dependent variable, and metric or dichotomous independent variables is known as **Binary Logistic Regression**, with the following formulation:

$$Y_i \in \{0; 1\}$$

$$Z_i = ln\left(\frac{p_i}{1 - p_i}\right) = B_0 + B_1 X_{1i} + \cdots + B_k\, X_{kn}$$

Where:

i = Each case of a sample size n;

Y_i = Dependent Variable Dichotomous (Occurrence = 1 and Non-Occurrence = 0);

Z_i = Logit;

p_i = Probability of Occurrence [$\mu(Y) = p_i$ e $\sigma^2(Y) = p_i \times (1 - p_i)$];

$1 - p_i$ = Probability of non-occurrence;

B_0 = Constant;

B_k = Regression coefficients;

X_{ki} = Independent Variable k (Predictor k).

MVDA – *Binary Logistic Regression*

The logit, which is a continuous variable, is calculated as the natural logarithm of chance, and chance is defined as the ratio between the occurrence and non-occurrence of an event. For example, a chance 3:1 means that for every 4 events, 3 events occurs and 1 do not.

$$ln\left(\frac{p_i}{1-p_i}\right) = Z_i$$

$$\frac{p_i}{1-p_i} = e^{(Z_i)}$$

$$chance_{Y_i=1} = e^{Z_i}$$

MVDA – *Binary Logistic Regression*

The output of a logit model is the probability of a case (i) to belong to an occurrence group ($Y_i = 1$) or a non-occurrence group ($Y_i = 0$).

$$\frac{p_i}{1 - p_i} = e^{(Z_i)}$$

$$p_i = \left(\frac{e^{(Z_i)}}{1 + e^{(Z_i)}}\right) = \left(\frac{1}{1 + e^{-(B_0 + B_1 X_{1i} + \cdots + B_k X_{kn})}}\right)$$

$$1 - p_i = \left(\frac{1}{1 + e^{(Z_i)}}\right) = \left(\frac{1}{1 + e^{(B_0 + B_1 X_{1i} + \cdots + B_k X_{kn})}}\right)$$

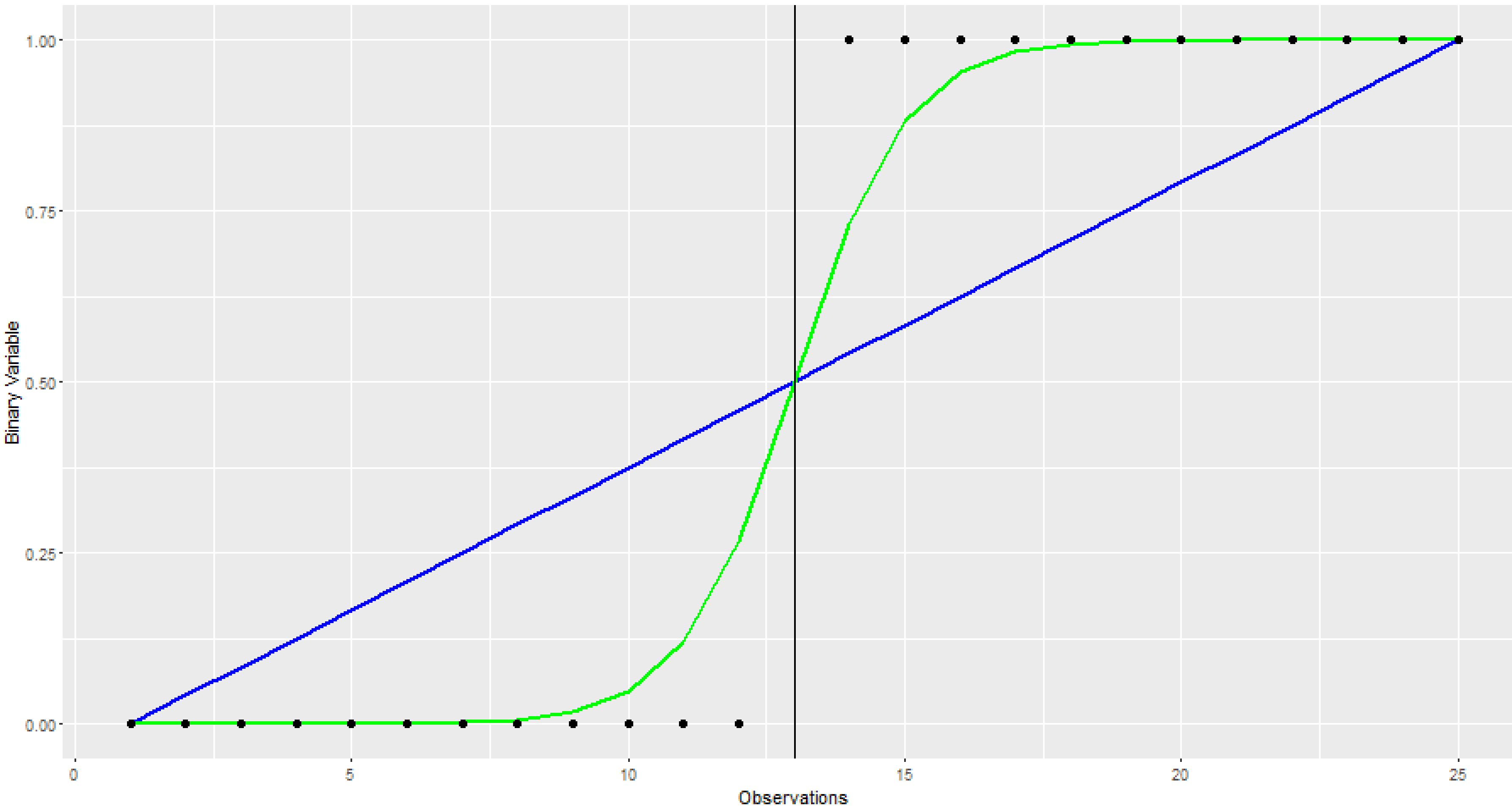
MVDA – *Binary Logistic Regression*

In order to properly model a dataset in which the dependent variable is non-metric, the multiple linear regression cannot be used, because the assumption of homoscedasticity is violated. This violation is very severe and invalidate the results of the multiple linear regression model.

But when the variables do not meet the assumptions of:

- Normality,
- Linearity,
- Homoscedasticity.

The logit model is the technique of choice, since it does not make these assumptions.


1.00
0.75
0.50
0.25
0.00
Binary Variable
0
5
10
15
20
25
Observations

Assumptions

MVDA – *Binary Logistic Regression*

ASSUMPTIONS

- The dependent variable must be dichotomous;
- The independent variables must be metric or dichotomous;
- The Ratio $\frac{n}{k} \geq 10$ $\rightarrow$ at least 10 observations (n) for each predictor (k). The higher the ratio $\frac{n}{k}$ better;
- Absence of collinearity or multicollinearity;
- Outliers Verification.

MVDA – *Binary Logistic Regression*

In order to explain a **Logit** approach, the following dataset case study will be used: A high school needs to know if students that go to school by car are more likely to arrive late ($Y_i = 1$) or not ($Y_i = 0$) in the classroom. A sample of 100 students was collected and in addition to the indication of, if the student arrived late or not, the following information was also collected:

- Distance traveled (km);
- Quantity of traffic lights (discrete variable);
- Period day (categorical variable: Morning or Afternoon*);
- Profile of the driver (categorical variable: Calm*, Moderate or Aggressive).

* Reference Category.

MVDA – *Binary Logistic Regression*

Id	Student	Y (Late?) Yes = 1; No = 0	Distance (X_1)	Traffic L. (X_2)	Period (X_3)	Profile (X_4)
1	Gabriela	0	12.5	7	Morning	Calm
2	Patrícia	0	13.3	10	Morning	Calm
3	Gustavo	0	13.4	8	Morning	Aggressive
4	Letícia	0	23.5	7	Morning	Calm
5	Luiz Ovídio	0	9.5	8	Morning	Calm
6	Leonor	0	13.5	10	Morning	Calm
7	Dalila	0	13.5	10	Morning	Calm
8	Antônio	0	15.4	10	Morning	Calm
9	Júlia	0	14.7	10	Morning	Calm
10	Mariana	0	14.7	10	Morning	Calm
…						
34	Cintia	0	11.5	10	Afternoon	Calm
…						
99	Leandro	1	14.2	10	Morning	Moderate
100	Estela	1	1	13	Morning	Calm

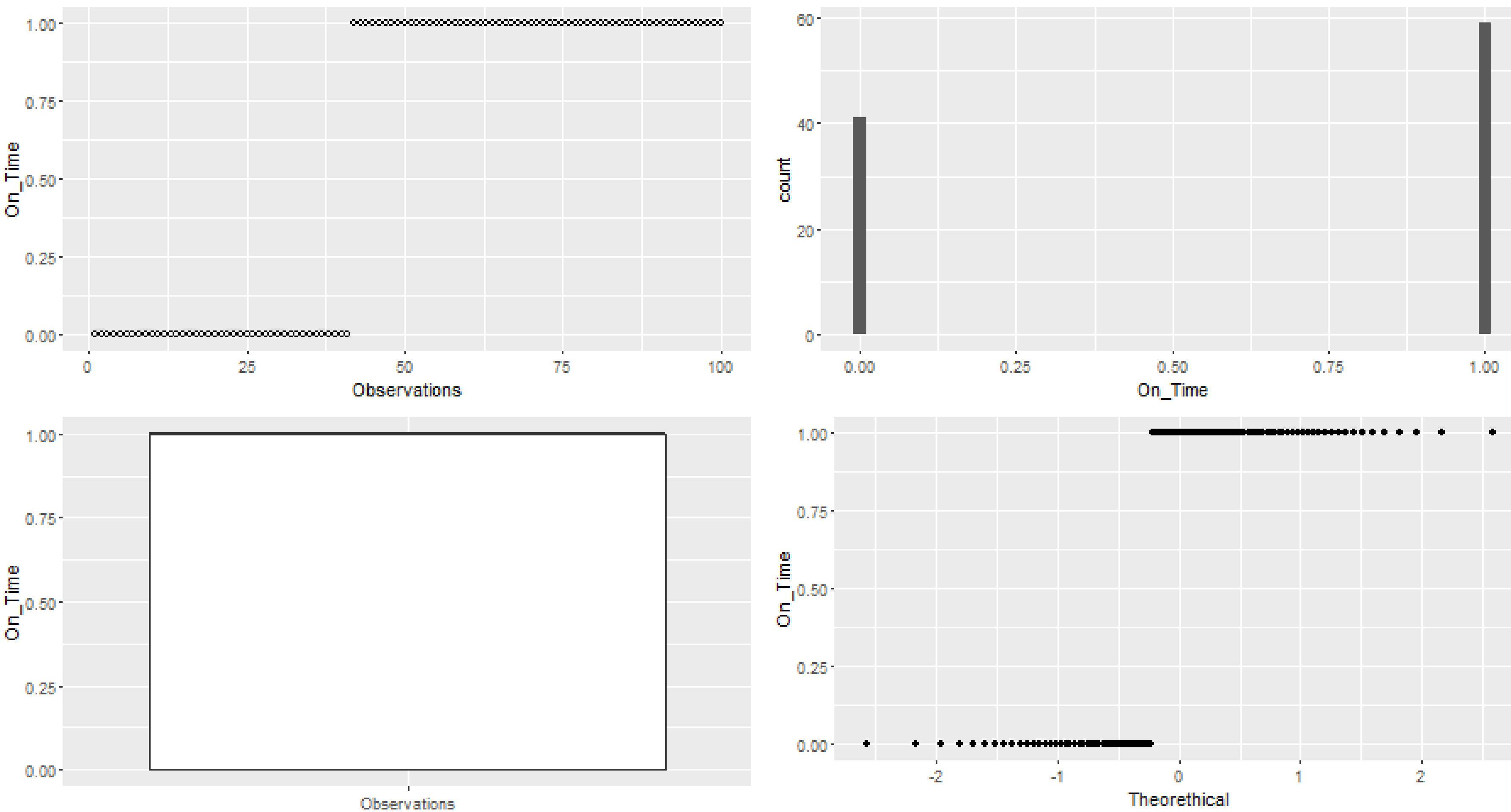
MVDA – *Binary Logistic Regression*

Id	Student	Y (Late?) Yes = 1; No = 0	Distance (X_1)	Traffic L. (X_2)	Period (X_3)	Profile A (X_4)	Profile B(X_5)
1	Gabriela	0	12.5	7	1	0	0
2	Patrícia	0	13.3	10	1	0	0
3	Gustavo	0	13.4	8	1	1	0
4	Letícia	0	23.5	7	1	0	0
5	Luiz Ovídio	0	9.5	8	1	0	0
6	Leonor	0	13.5	10	1	0	0
7	Dalila	0	13.5	10	1	0	0
8	Antônio	0	15.4	10	1	0	0
9	Júlia	0	14.7	10	1	0	0
10	Mariana	0	14.7	10	1	0	0
…							
34	Cintia	0	11.5	10	0	0	0
…							
99	Leandro	1	14.2	10	1	0	1
100	Estela	1	1	13	1	0	0

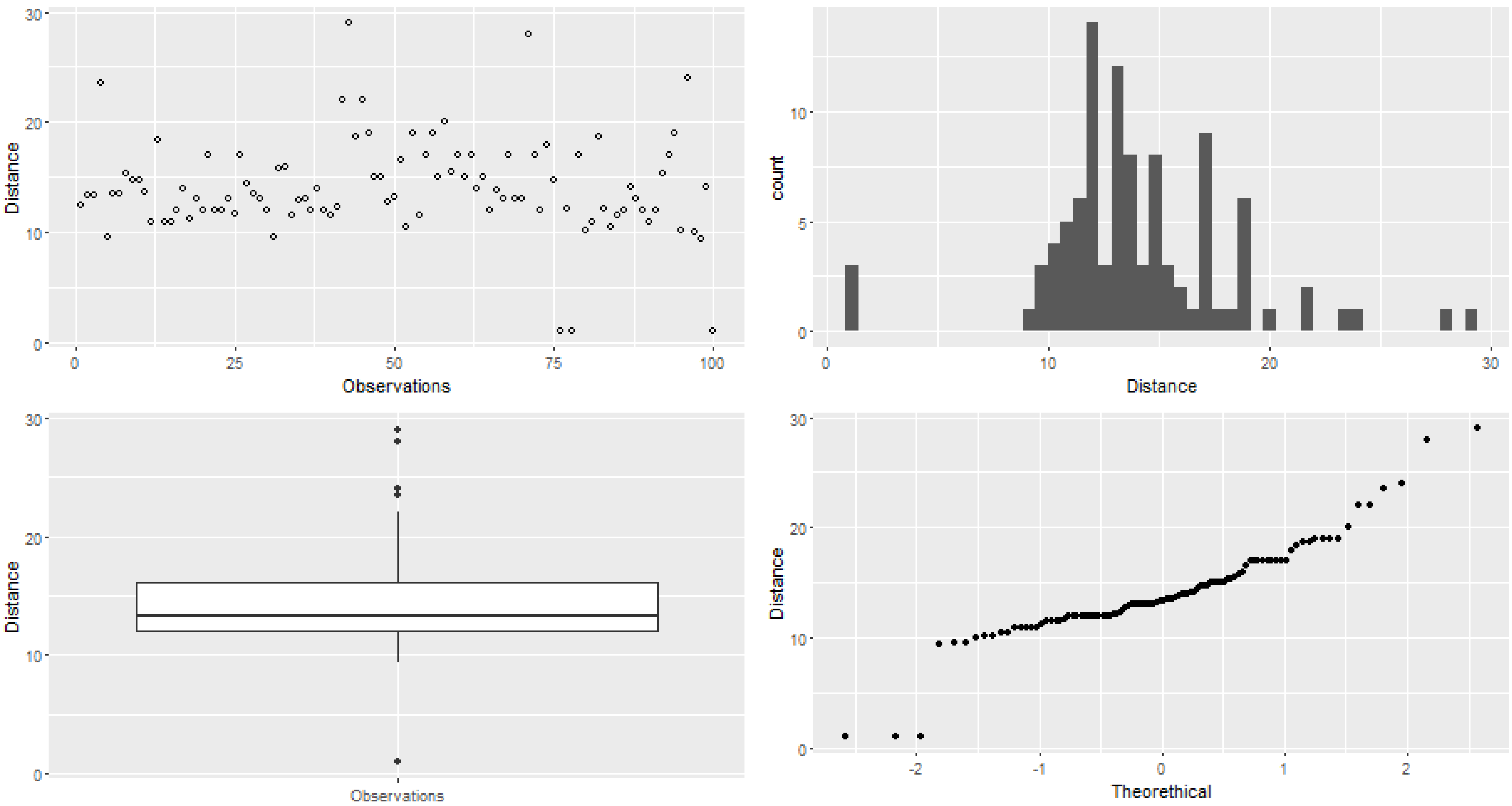
MVDA – *Binary Logistic Regression*

```
library("ggplot2")
ggplot(data = my_data, aes(x = Observations, y = On_Time)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = On_Time)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(data = my_data, aes(On_Time)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes(sample = On_Time)) + stat_qq()+ xlab("Theorethical") + ylab("On_Time")
ggplot(data = my_data, aes(x = Observations, y = Distance)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Distance)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(data = my_data, aes(Distance)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes(sample = Distance)) + stat_qq()+ xlab("Theorethical") + ylab("Distance")
ggplot(data = my_data, aes(x = Observations, y = Traffic_Lights)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Traffic_Lights)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(data = my_data, aes(Traffic_Lights)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes(sample = Traffic_Lights)) + stat_qq()+ xlab("Theorethical") + ylab("Traffic_Lights")
ggplot(data = my_data, aes(x = Observations, y = Time_Period)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Time_Period)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(data = my_data, aes(Time_Period)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes(sample = Time_Period)) + stat_qq()+ xlab("Theorethical") + ylab("Time_Period")
ggplot(data = my_data, aes(x = Observations, y =  Driver_Profile1)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y =  Driver_Profile1)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes( Driver_Profile1)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes(sample = Driver_Profile1)) + stat_qq()+ xlab("Theorethical") + ylab("Driver_Profile1")
ggplot(data = my_data, aes(x = Observations, y =  Driver_Profile2)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y =  Driver_Profile2)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes( Driver_Profile2)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes(sample = Driver_Profile2)) + stat_qq()+ xlab("Theorethical") + ylab("Driver_Profile2")
```

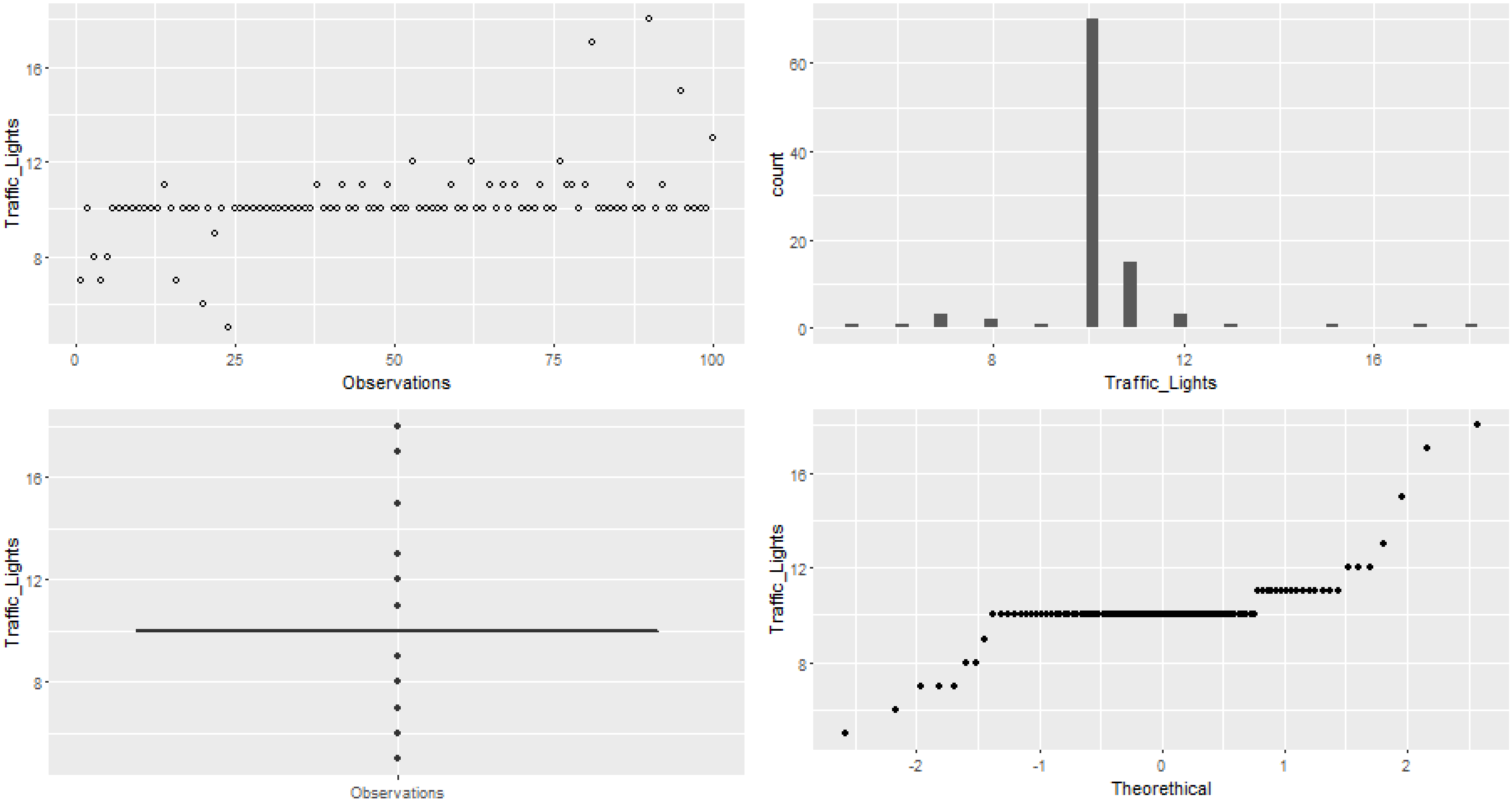
MVDA – *Binary Logistic Regression*



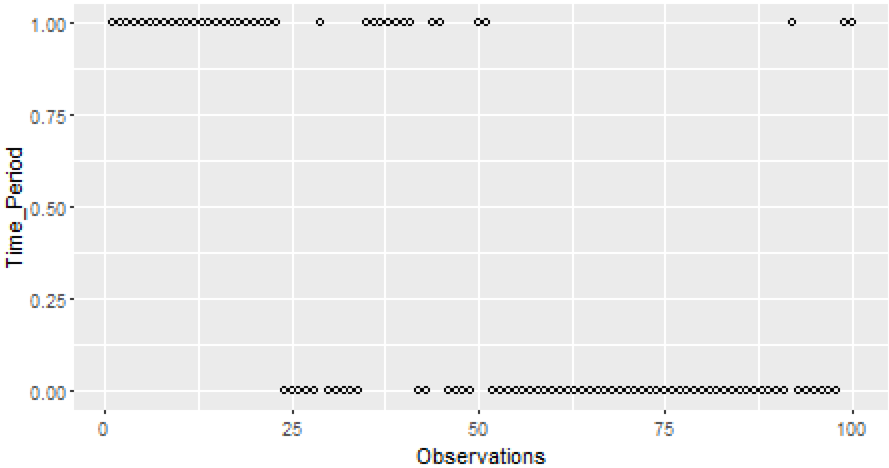
MVDA – *Binary Logistic Regression*

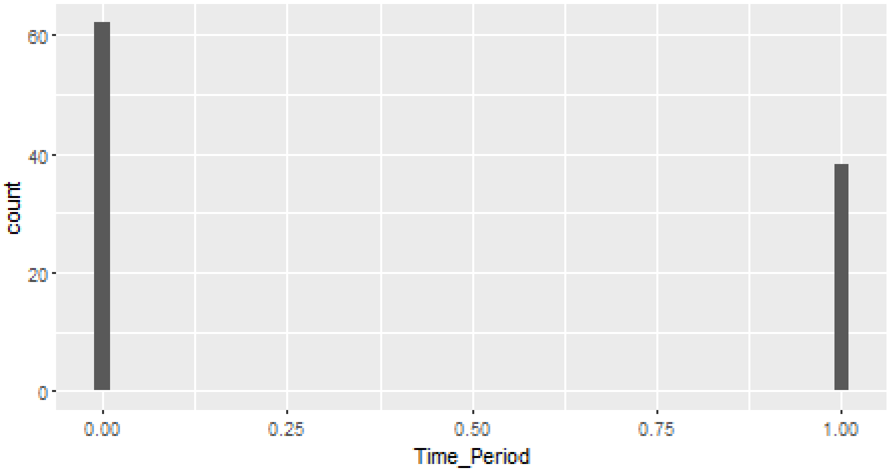


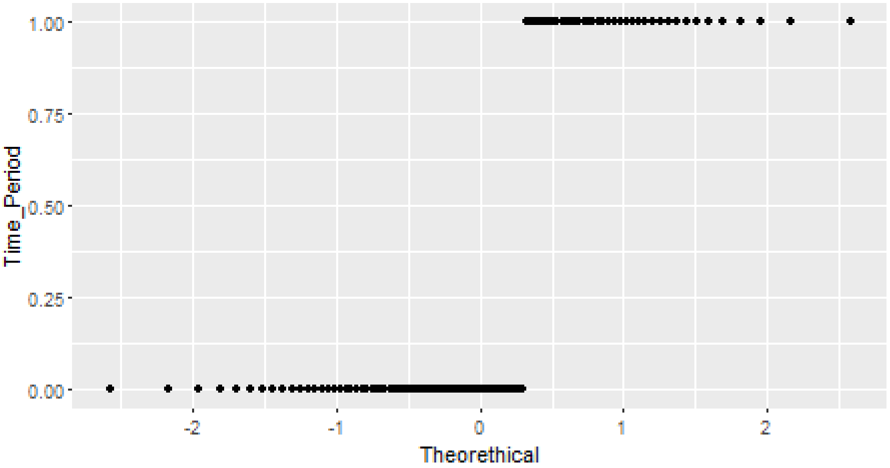
MVDA – *Binary Logistic Regression*



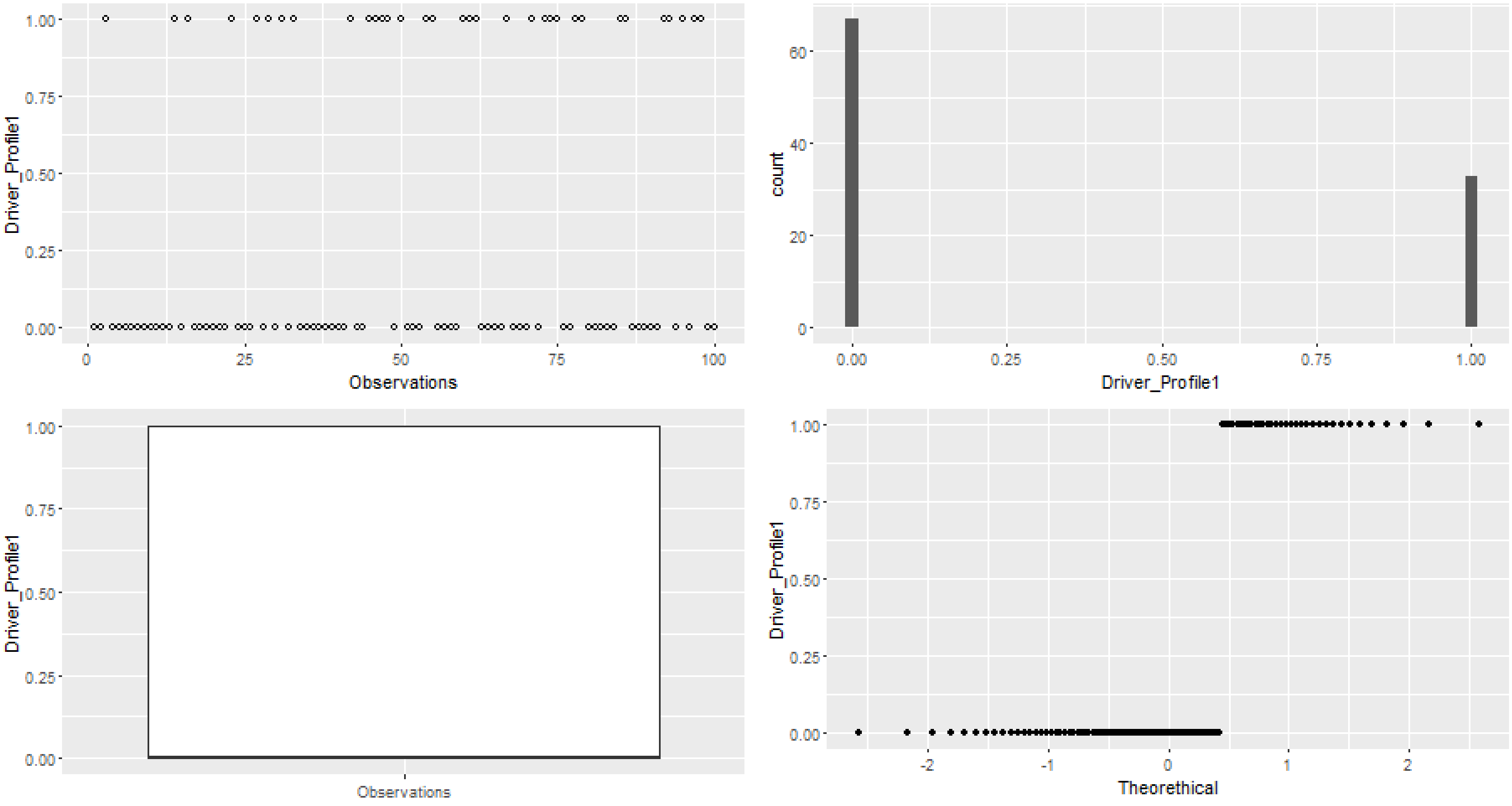
MVDA – *Binary Logistic Regression*



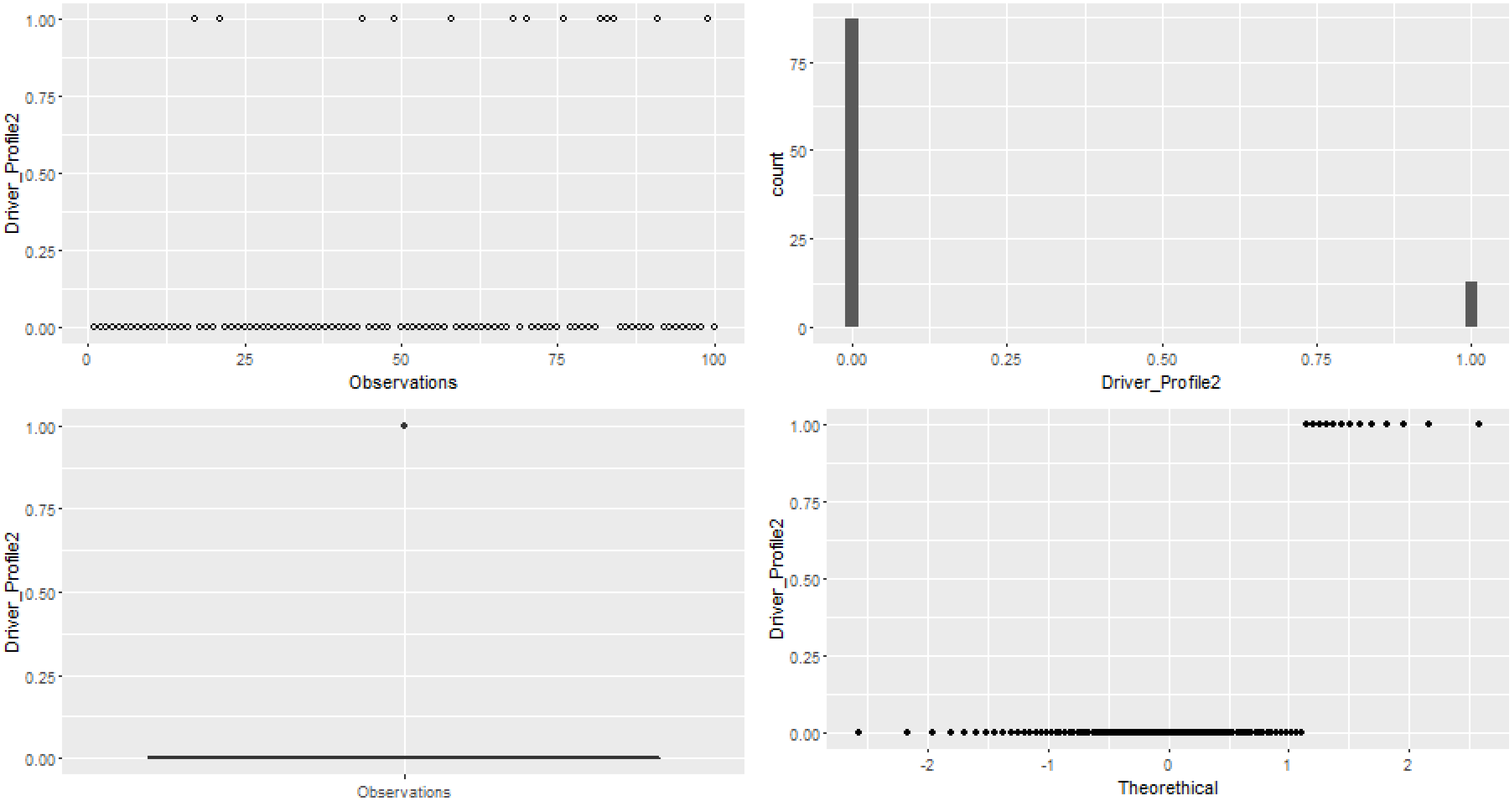




MVDA – *Binary Logistic Regression*

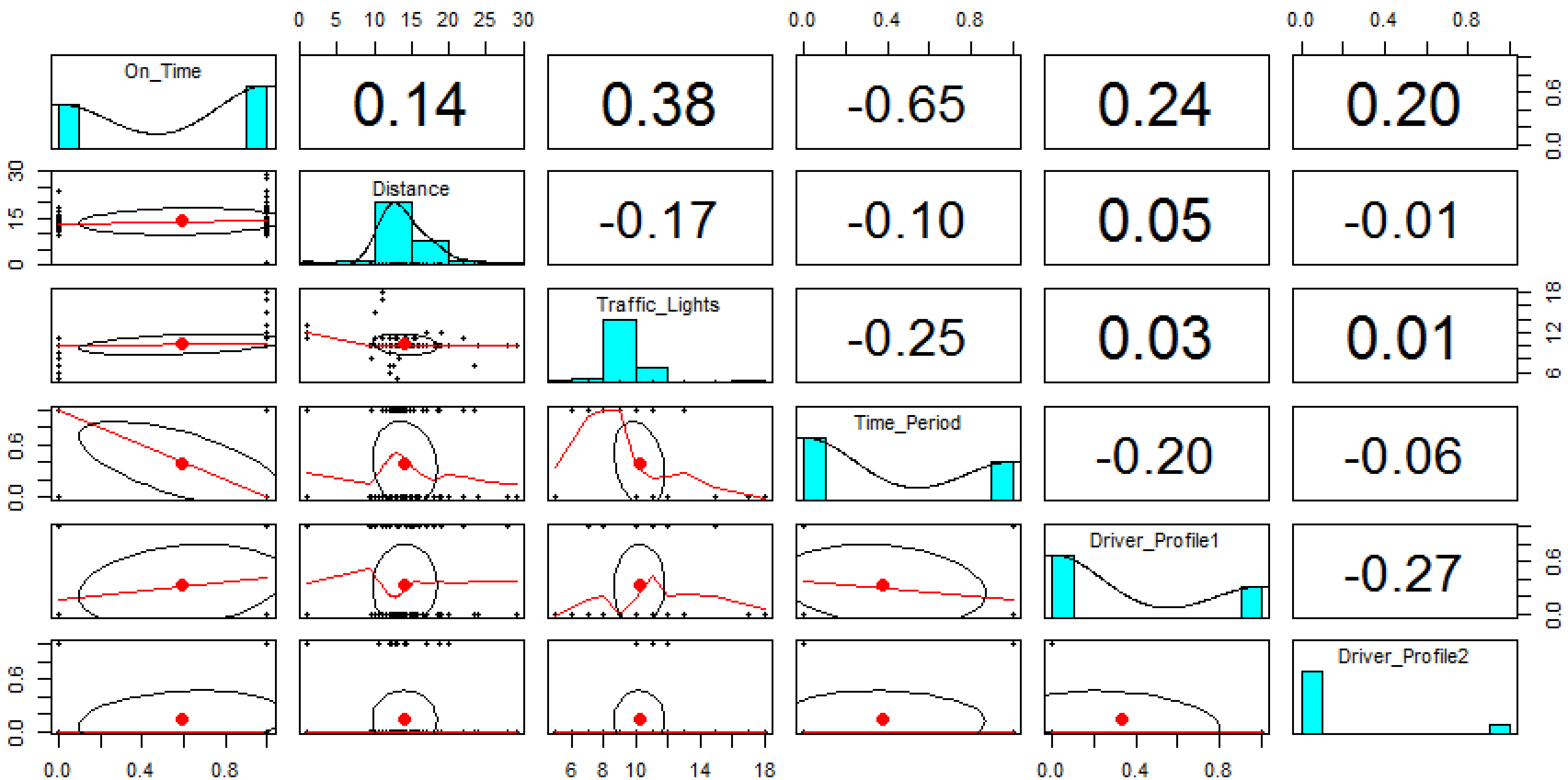


MVDA – *Binary Logistic Regression*



MVDA – *Binary Logistic Regression*

```
library ("psych")
pairs.panels(my_data)
```



Binary Logistic Regression

MVDA – *Binary Logistic Regression*

```
# Logit
logit_01 <- glm(On_Time ~ ., data = my_data, family = "binomial")
summary(logit_01)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2037  -0.2638   0.1231   0.4419   2.2928

Coefficients:
                Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)     -30.2003     9.9806     -3.026    0.00248 **
Distance          0.2202     0.1097      2.007    0.04474 *
Traffic_Lights    2.7667     0.9216      3.002    0.00268 **
Time_Period      -3.6534     0.8781     -4.160    3.18e-05 ***
Driver_Profile1   1.3460     0.7477      1.800    0.07184 .
Driver_Profile2   2.9145     1.1788      2.472    0.01342 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:     135.372  on 99  degrees of freedom
Residual deviance:  58.131  on 94  degrees of freedom
AIC: 70.131
```

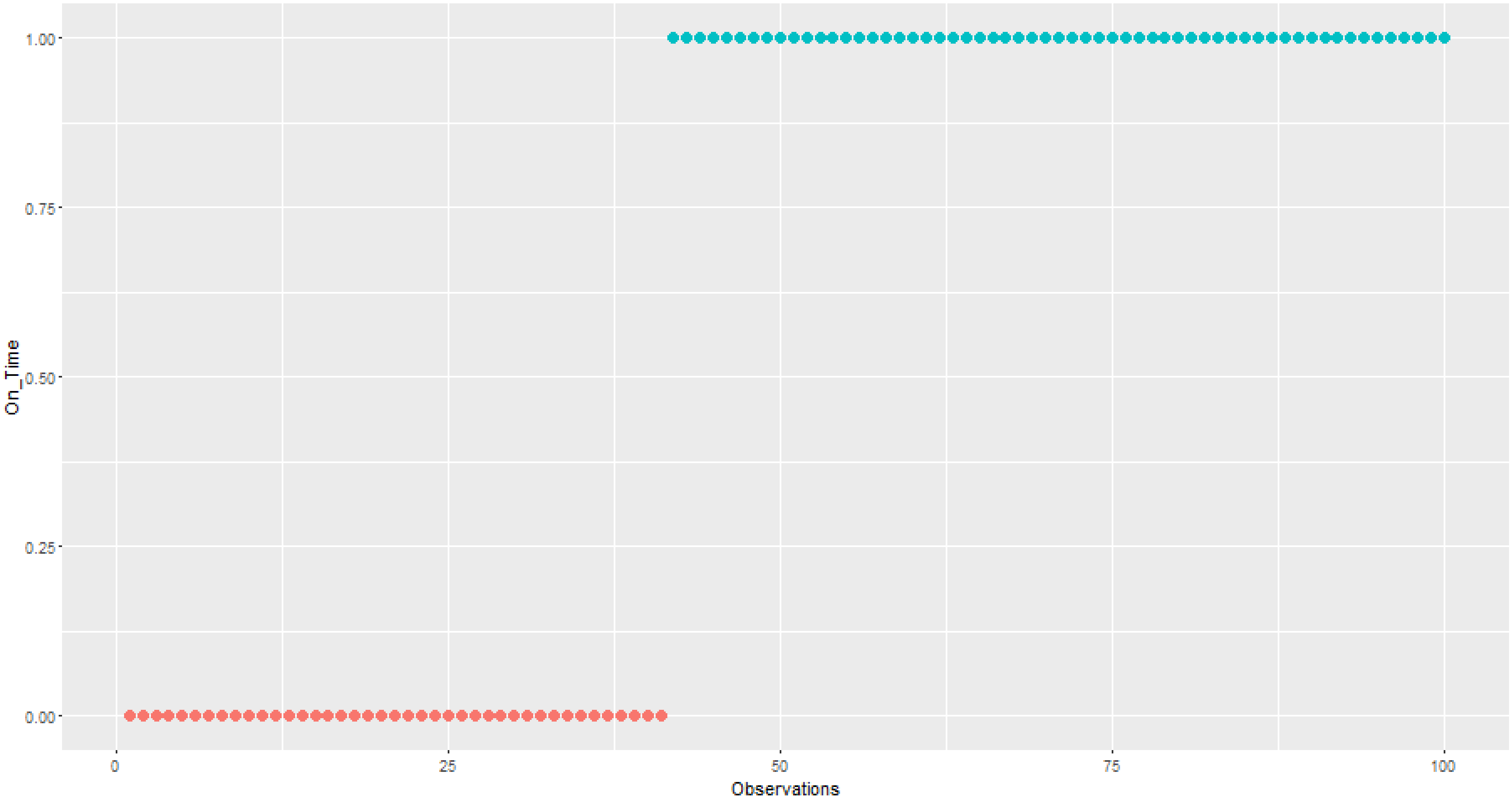
MVDA – *Binary Logistic Regression*

```
logit_02 <- glm(On_Time ~ Distance + Traffic_Lights + Time_Period + Driver_Profile2, data = my_data, family = "binomial")
summary(logit_02)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9495  -0.3169   0.1364   0.5692   2.3967

Coefficients:
                Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)     -30.9333   10.6344      -2.909    0.00363 **
Distance          0.2041    0.1012       2.018    0.04358 *
Traffic_Lights    2.9201    1.0106       2.890    0.00386 **
Time_Period      -3.7763    0.8466      -4.461    8.18e-06 ***
Driver_Profile2   2.4591    1.1394       2.158    0.03091 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:      135.372  on 99  degrees of freedom
Residual deviance:   61.602  on 95  degrees of freedom
AIC: 71.602
```

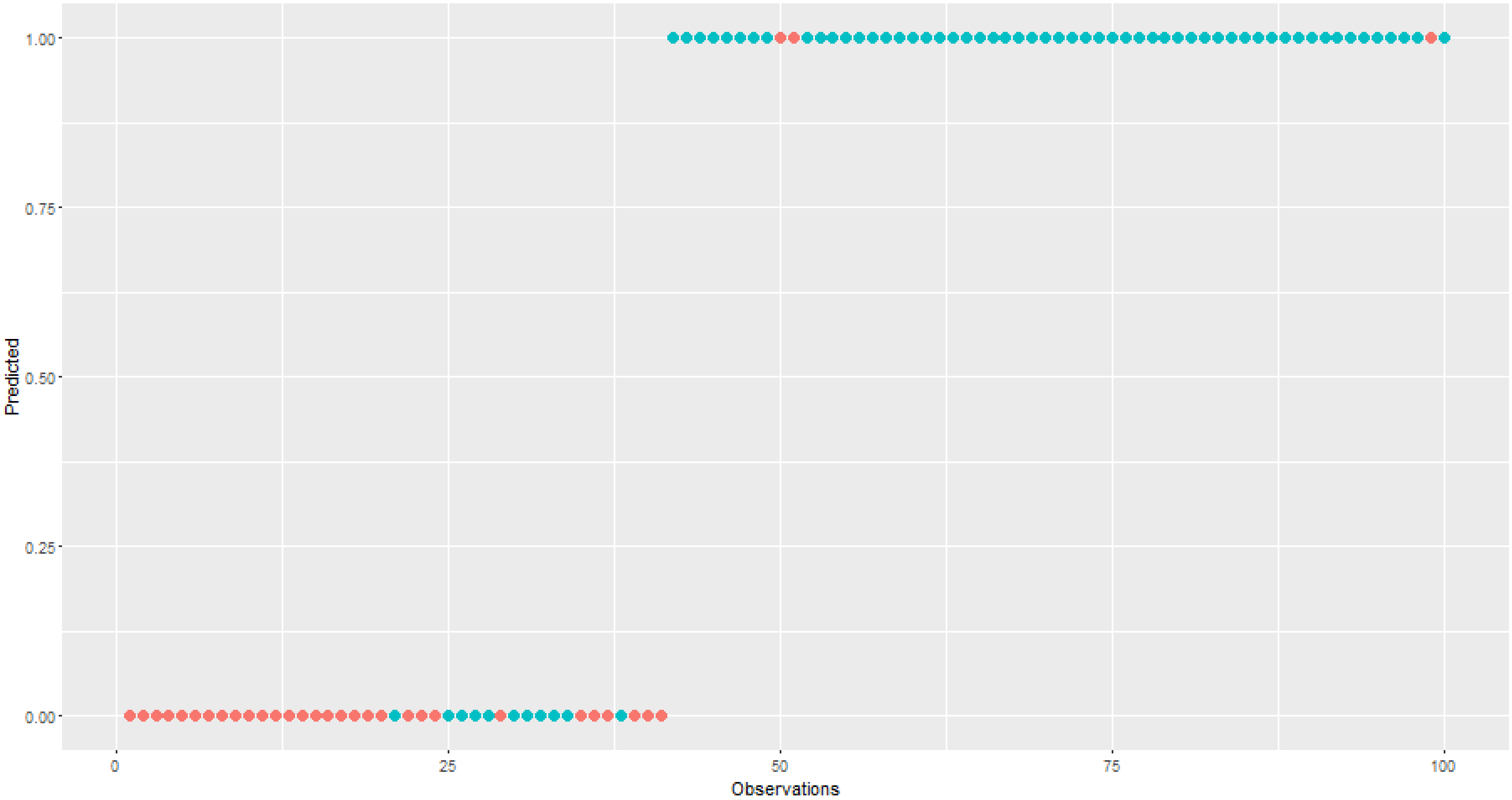
MVDA – *Binary Logistic Regression*

```
# Prediction
prob  <- predict(logit_02,type = c("response"))

ggplot(my_data, aes(x = 1:100, y = On_Time)) + geom_point(aes(colour = ifelse(my_data$On_Time >= 0.5, "red", "blue")), size = 3) + theme(legend.position = "none") + xlab ("Observations") + ylab("On_Time")

ggplot(my_data, aes(x = 1:100, y = On_Time)) + geom_point(aes(colour = ifelse(prob >= 0.5, "red", "blue")), size = 3) + theme(legend.position = "none") + xlab ("Observations") + ylab("Predicted")
```

MVDA – *Binary Logistic Regression*



MVDA – *Binary Logistic Regression*



MVDA – *Binary Logistic Regression*

The probability of Y_i is given by:

$$p(Y_i) = (p_i)^{Y_i} \times (1 - p_i)^{1-Y_i}$$

For a sample with n cases, we can define the likelihood function as :

$$L = \prod_{i=1}^{n} [(p_i)^{Y_i} \times (1 - p_i)^{1-Y_i}]$$

$$L = \prod_{i=1}^{n} \left[\left(\frac{e^{(Z_i)}}{1 + e^{(Z_i)}} \right)^{Y_i} \times \left(\frac{1}{1 + e^{(Z_i)}} \right)^{1-Y_i} \right]$$

MVDA – *Binary Logistic Regression*

In practice it is more convenient to work with the maximum estimation of the log likelihood function:

$$LL = \sum_{i=1}^{n} \left\{ \left[(Y_i) \ln\left(\frac{e^{(Z_i)}}{1 + e^{(Z_i)}} \right) \right] + \left[(1 - Y_i) \ln\left(\frac{1}{1 + e^{(Z_i)}} \right) \right] \right\} = máx$$

```
# Betas
logit_02$coefficients

   (Intercept)          Distance     Traffic_Lights      Time_Period    Driver_Profile2
 -30.9333453         0.2041463          2.9201140        -3.7763010          2.4590675

# Maximum LogLikelyhood
LLmax <- logLik(logit_02)

'log Lik.' -30.80079 (df=5)
```

MVDA – *Binary Logistic Regression*

The confidence interval is given by $B_i \pm z_{\alpha/2} \times SE_{B_i}$. Therefore for a 95% $(z = 1.96)$ confidence interval:

```
# CIs Using Standard Errors
confint.default(logit_02,level = 0.95)

                     2.5 %          97.5 %
(Intercept)     -51.776290960   -10.0903996
Distance          0.005886105     0.4024066
Traffic_Lights    0.939388292     4.9008396
Time_Period      -5.435617700    -2.1169843
Driver_Profile2   0.225909956     4.6922250
```

Goodness of Fit

MVDA – *Binary Logistic Regression*

We need to calculate the adequacy of the model. To measure the adequacy we use the null model (LL_0) and compared with our final model ($LL_{máx}$) through the likelihood ratio test. The null model has only the intercept (B_0). Therefore the following hypothesis can be tested:

$$H_0: The\ model\ is\ not\ adequate$$
$$H_1: The\ model\ is\ adequate$$

```
# Likelihood Ratio Test
with(logit_02, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

[1] 3.6265e-15
```

MVDA – *Binary Logistic Regression*

There are several measures of association designed to mimic the r^2 analysis of, like the $pseudo\ r^2$. Values between 0.2 and 0.4 are considered highly satisfactory.

$$pseudo(r^2)_{MacFadden} = \left(\frac{-2LL_0 + 2LL_{max}}{-2LL_0}\right)$$

$$pseudo(r^2)_{Cox\ \&\ Snell} = 1 - \left(\frac{e^{LL_0}}{e^{LL_{max}}}\right)^{\frac{2}{N}}$$

$$pseudo(r^2)_{Nagelkerke} = \frac{1 - \left(\frac{e^{LL_0}}{e^{LL_{max}}}\right)^{\frac{2}{N}}}{1 - \left(e^{LL_0}\right)^{\frac{2}{N}}}$$

MVDA – *Binary Logistic Regression*

```
# Pseudo-r2
LL0 <- logLik(glm(On_Time ~ 1, data = my_data, family = "binomial"))
LLmax <- logLik(logit_02)

psd_r_McFadden      = (-2*LL0 + 2*LLmax)/(-2*LL0)

[1] 'log Lik.' 0.544945 (df=1)

psd_r_Cox_and_Snell = 1 - (exp(LL0)/exp(LLmax))^(2/nrow(my_data))

[1] 'log Lik.' 0.5217881 (df=1)

psd_r_Nagelkerke    = (1 - (exp(LL0)/exp(LLmax))^(2/nrow(my_data)))/(1 - (exp(LL0))^(2/nrow(my_data)))

[1] 'log Lik.' 0.7034824 (df=1)
```

Interpretation

MVDA – *Binary Logistic Regression*

$$Z_i = -30.933 + 0.204X_{1i} + 2.920X_{2i} - 3.776X_{3i} + 2.459X_{5i}$$

$$p_i = \left(\frac{e^{-30.933+0.204X_{1i}+2.920X_{2i}-3.776X_{3i}+2.459X_{5i}}}{1 + e^{-30.933+0.204X_{1i}+2.920X_{2i}-3.776X_{3i}+2.459X_{5i}}}\right)$$

$e^{B_i}; i \neq 0 \rightarrow$ Average change in the chance of arriving late ($Y = 1$) all other conditions remain constant.

```
# Odds Ratios
exp_coef <- exp(coef(logit_02))

  (Intercept)      Distance      Traffic_Lights      Time_Period      Driver_Profile2
 3.679754e-14   1.226478e+00     1.854340e+01       2.290727e-02      1.169390e+01
```

MVDA – *Binary Logistic Regression*

$e^{B_1} = e^{0.204} = 1.226 \therefore$ Chance to arrive late increases 22.6% if the distance increase in 1 km.

$e^{B_2} = e^{2.920} = 18.543 \therefore$ Chance to arrive late increases 1754.3% if the quantity of traffic ligths increases in 1 unity.

$e^{B_3} = e^{-3.776} = 0.023 \therefore$ Chance to arrive late decreases 97.7% in the morning period.

$e^{B_5} = e^{2.459} = 11.693 \therefore$ Chance to arrive late increases 1069.3% if the driver has an aggressive profile

MVDA – *Binary Logistic Regression*

$$Z_i = -30.933 + 0.204\boldsymbol{X_{1i}} + 2.920\boldsymbol{X_{2i}} - 3.776\boldsymbol{X_{3i}} + 2.459\boldsymbol{X_{5i}}$$

$$p_i = \left(\frac{e^{-30.933+0.204\boldsymbol{X_{1i}}+2.920\boldsymbol{X_{2i}}-3.776\boldsymbol{X_{3i}}+2.459\boldsymbol{X_{5i}}}}{1 + e^{-30.933+0.204\boldsymbol{X_{1i}}+2.920\boldsymbol{X_{2i}}-3.776\boldsymbol{X_{3i}}+2.459\boldsymbol{X_{5i}}}}\right)$$

$$p_{Gabriela} = \left(\frac{e^{-30.933+0.204(\mathbf{12.5})+2.920(\mathbf{7})-3.776(\mathbf{1})+2.459(\mathbf{0})}}{1 + e^{-30.933+0.204(\mathbf{12.5})+2.920(\mathbf{7})-3.776(\mathbf{1})+2.459(\mathbf{0})}}\right)$$

$$p_{Gabriela} = \left(\frac{e^{-13.819}}{1 + e^{-13.819}}\right) = \left(\frac{0.00000996}{1.00000996}\right) = 0.0009\%$$

Gabriela has a change of 0.009% to arrive late ($Y = 1$)

MVDA – *Binary Logistic Regression*

Cutoff de 0.5

≠ *Classification*

Id	Studant	Y (Late?) Yes = 1; No = 0	Probability i	Prediction
1	Gabriela	0	8.01978E-06	0
2	Patrícia	0	0.037039567	0
3	Gustavo	0	0.000597068	0
4	Letícia	0	9.03594E-05	0
5	Luiz Ovídio	0	6.58771E-05	0
6	Leonor	0	0.038642641	0
7	Dalila	0	0.038642641	0
8	Antônio	0	0.057559687	0
9	Júlia	0	0.049747133	0
10	Mariana	0	0.049747133	0
...				
34	Cintia	0	0.499731557	0
...				
99	**Leandro**	**1**	**0.463704**	**0**
100	Estela	1	0.911589482	1

MVDA – *Binary Logistic Regression*

Confusion Matrix:

TN (*True Negative*)	FN (*False Negative*)	PN (*Predicted Negative*)
FP (*False Positive*)	TP (*True Positive*)	PP (*Predicted Positive*)
ON (*Observed Negative*)	OP (*Observed Positive*)	n = TN+FP+FN+TP

TN = Case: 0 & Prediction: 0;

TP = Case: 1 & Prediction: 1;

FN = Case: 0 & Prediction: 1; Type II Error

FP = Case: 1 & Prediction: 0; Type I Error

$PN = TN + FN;$

$PP = TP + FP;$

$ON = TN + FP;$

$OP = TP + FN;$

MVDA – *Binary Logistic Regression*

- $\boldsymbol{TPR}$ (*True Positive Rate*) = $\frac{TP}{OP}$
- $\boldsymbol{TNR}$ (*True Negative Rate*) = $\frac{TN}{ON}$
- $\boldsymbol{ACC}$ (*Accuracy*) = $\frac{TP+TN}{n}$;
- $\boldsymbol{FPR}$ (*False Positive Rate*) = $1 - TNR$ ou $\frac{FP}{ON}$
- $\boldsymbol{PPV}$ (*Positive Prediticed Value = Sensitivity*) = $\frac{TP}{PP}$;
- $\boldsymbol{NPV}$ (*Negative Prediticed Value = Specificity*) = $\frac{TN}{PN}$;

MVDA – *Binary Logistic Regression*

```
# Confusion Matrix
library("SDMTools")
confusion.matrix(my_data$On_Time, fitted(logit_02), threshold = 0.5)

         obs
pred     0     1
   0    30     3
   1    11    56

# Accuracy
acc <- (30+56)/100

[1] 0.86
```

MVDA – *Binary Logistic Regression*

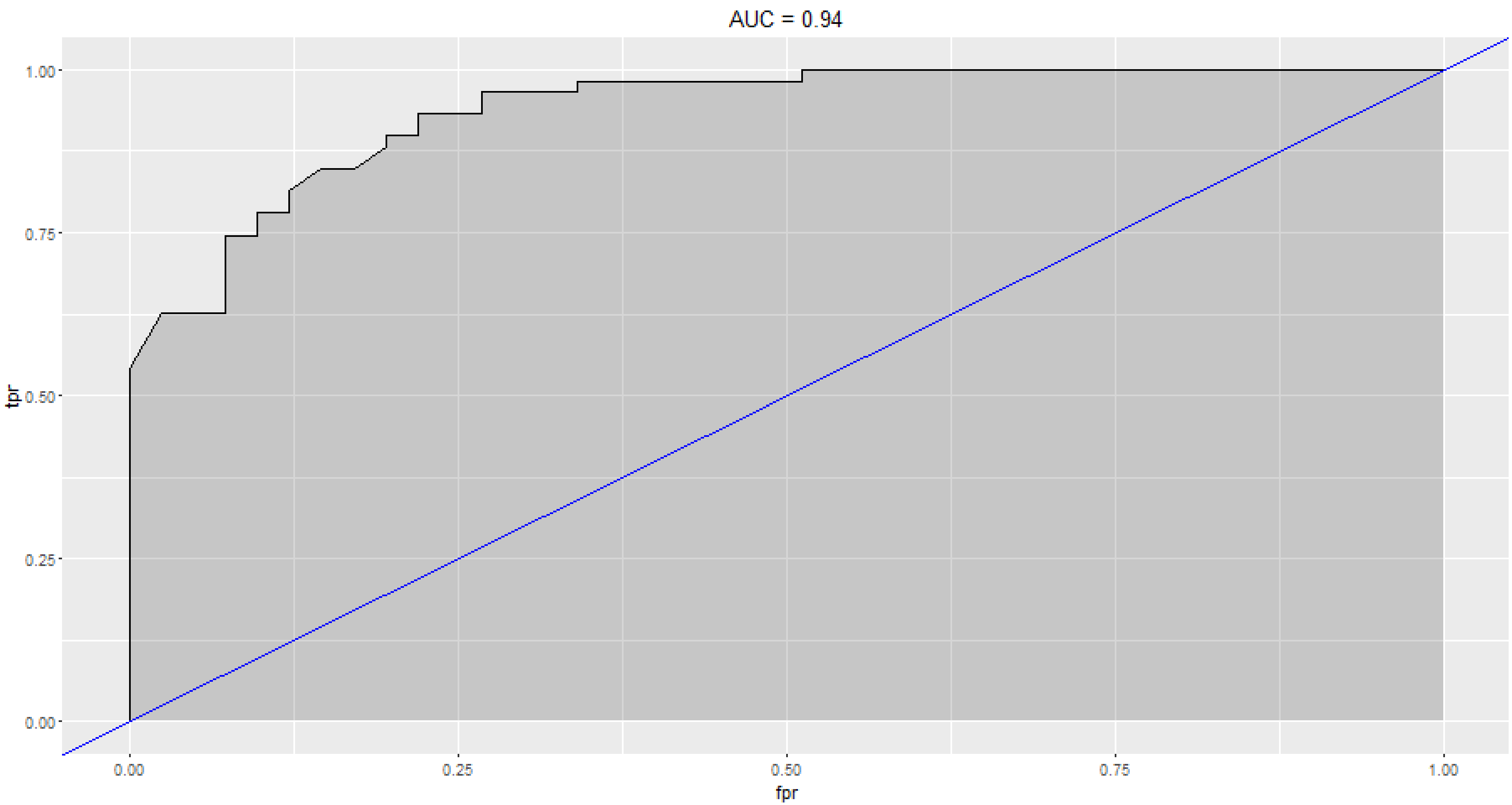
TN = 30	FN = 3	PN = 41
FP = 11	TP = 56	PP = 59
ON = 44	OP = 59	n = 100

MVDA – *Binary Logistic Regression*

The values of *Sensitivity* and 1 - *Specificity* are used to plot a ROC curve (Receiver Operating Characteristic). The more distant the ROC curve is in relation to a reference curve, the better. A curve very close to the reference shows that the model's ability to discriminate between the occurrence and non-occurrence is due to chance.

```
# ROC Curve
library("ggplot2")
library("ROCR")
pred <- prediction(prob, my_data$On_Time)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
auc <- performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
roc.data <- data.frame(fpr = unlist(perf@x.values), tpr = unlist(perf@y.values), model = "GLM")
ggplot(roc.data, aes(x = fpr, ymin = 0, ymax = tpr)) + geom_ribbon(alpha = 0.2) + geom_line(aes(y = tpr)) +
geom_abline(colour = "blue", intercept = 0, slope = 1) + ggtitle(paste0("AUC = ", round(auc, digits = 2)))
```

MVDA – *Binary Logistic Regression*



Prediction

MVDA – *Binary Logistic Regression*

```
# Prediction
new_data <- as.data.frame(cbind(9.5, 8, 1, 0))
colnames(new_data) <- c("Distance", "Traffic_Lights", "Time_Period", "Driver_Profile2")
predict(logit_02, new_data, type = c("response"))
```

MVDA

https://github.com/Valdecy/Multivariate_Data_Analysis

##
##

Created by: Prof. Valdecy Pereira. D.Sc.
UFF - Universidade Federal Fluminense (Brazil)
email: valdecypereira@yahoo.com.br
Course: Multivariate Data Analysis
Lesson: Binary Logistic Regression (Logit)

Citation:
PEREIRA. V. (2016). Project: Multivariate Data Analysis. File: R-MVDA-08-LR-B.pdf. GitHub repository: <https://github.com/Valdecy/Multivariate_Data_Analysis>

##
##

Bibliography

CORRAR, L.J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada para Cursos de Administração, Ciências Contábeis e Economia**. ATLAS, 2009.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. CAMPUS, 2009.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. BOOKMAN, 2009.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. CENGAGE Learning, 2011.

LEVINE, D. M.; STEPHAN, D. F.; KREHBIEL, T. C.; BERENSON, M. L. **Estatística - Teoria e Aplicações - Usando Microsoft Excel**. LTC, 2012.