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

Forecasting

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

Forecasting

A regression technique that has a dichotomous dependent variable, and metrics 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).

Forecasting

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.

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

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

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

Forecasting

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 ($Y_i = 0$).

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

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

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

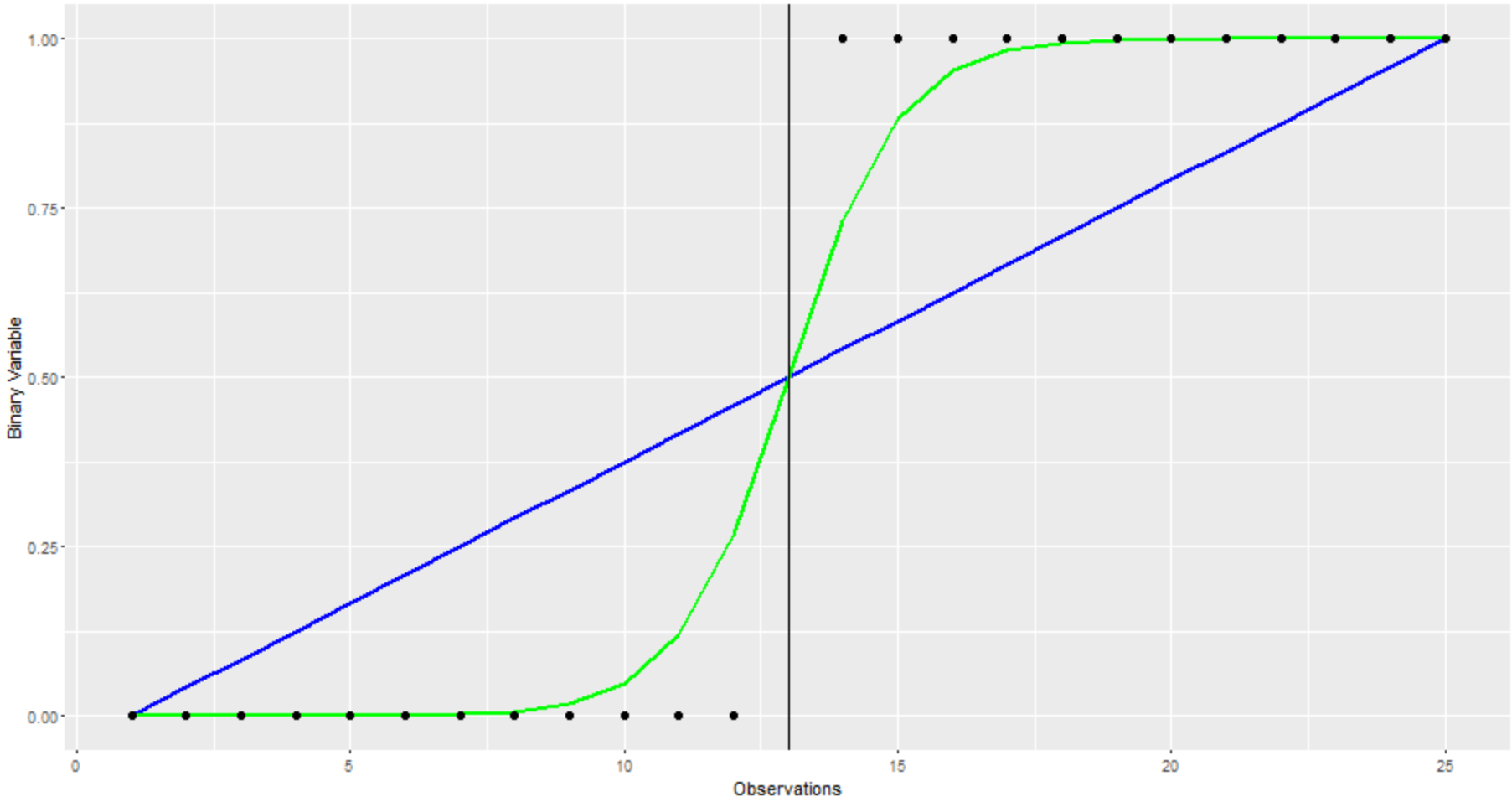
Forecasting

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

Forecasting

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.

Forecasting

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.

Forecasting

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

Forecasting

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

Binary Logistic Regression

Forecasting

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 :

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

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

Forecasting

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

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

The values of regression coefficients are found by the Newton-Raphson method:

$$B_{m+1} = B_m + (X' V_m X)^{-1} X' (Y - P_m)$$

B_m= Previous estimative, where $B_0 = 0; \forall i$

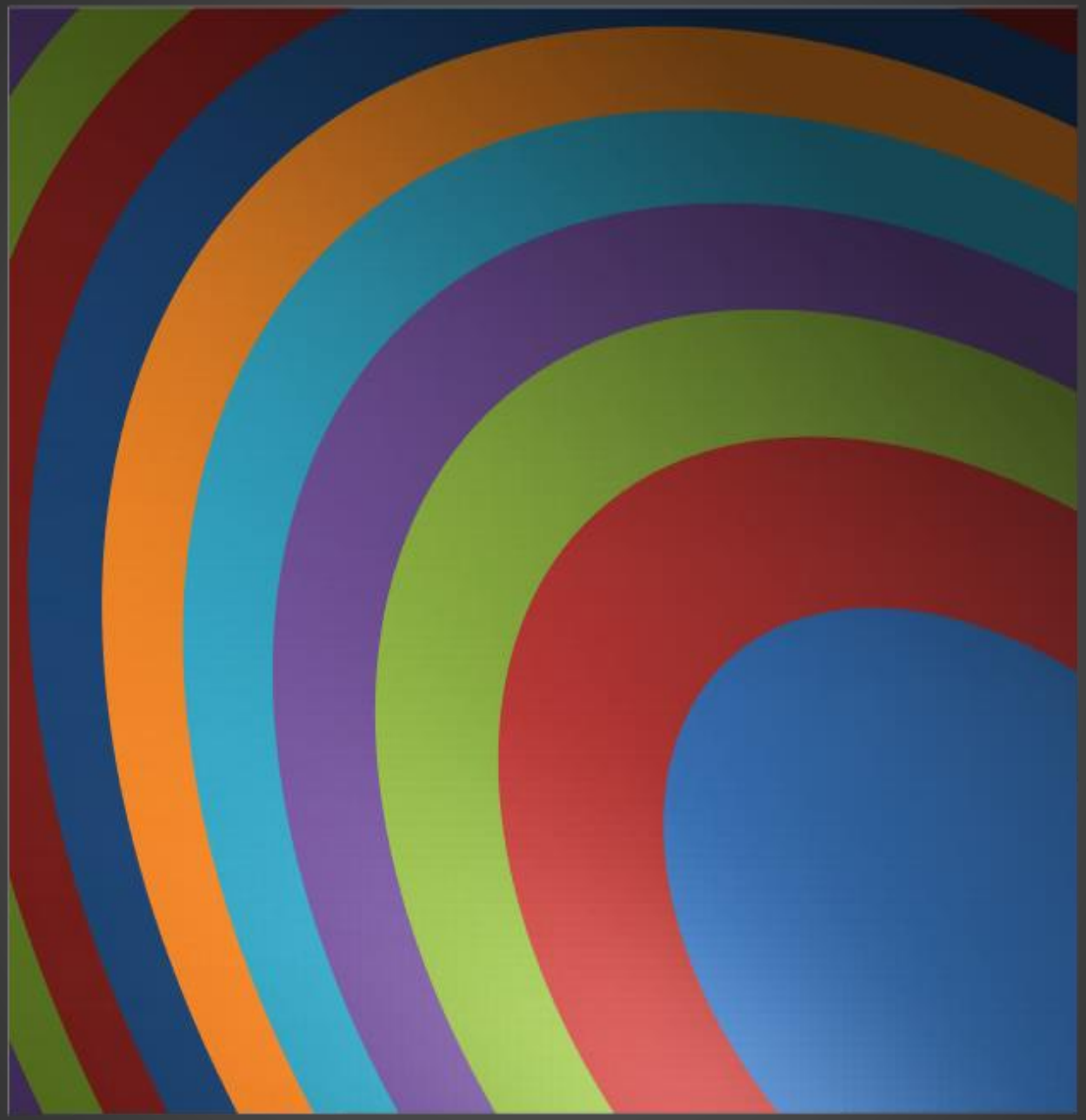
B_{m+1}= Next estimative, converge when $B_{m+1} = B_m$

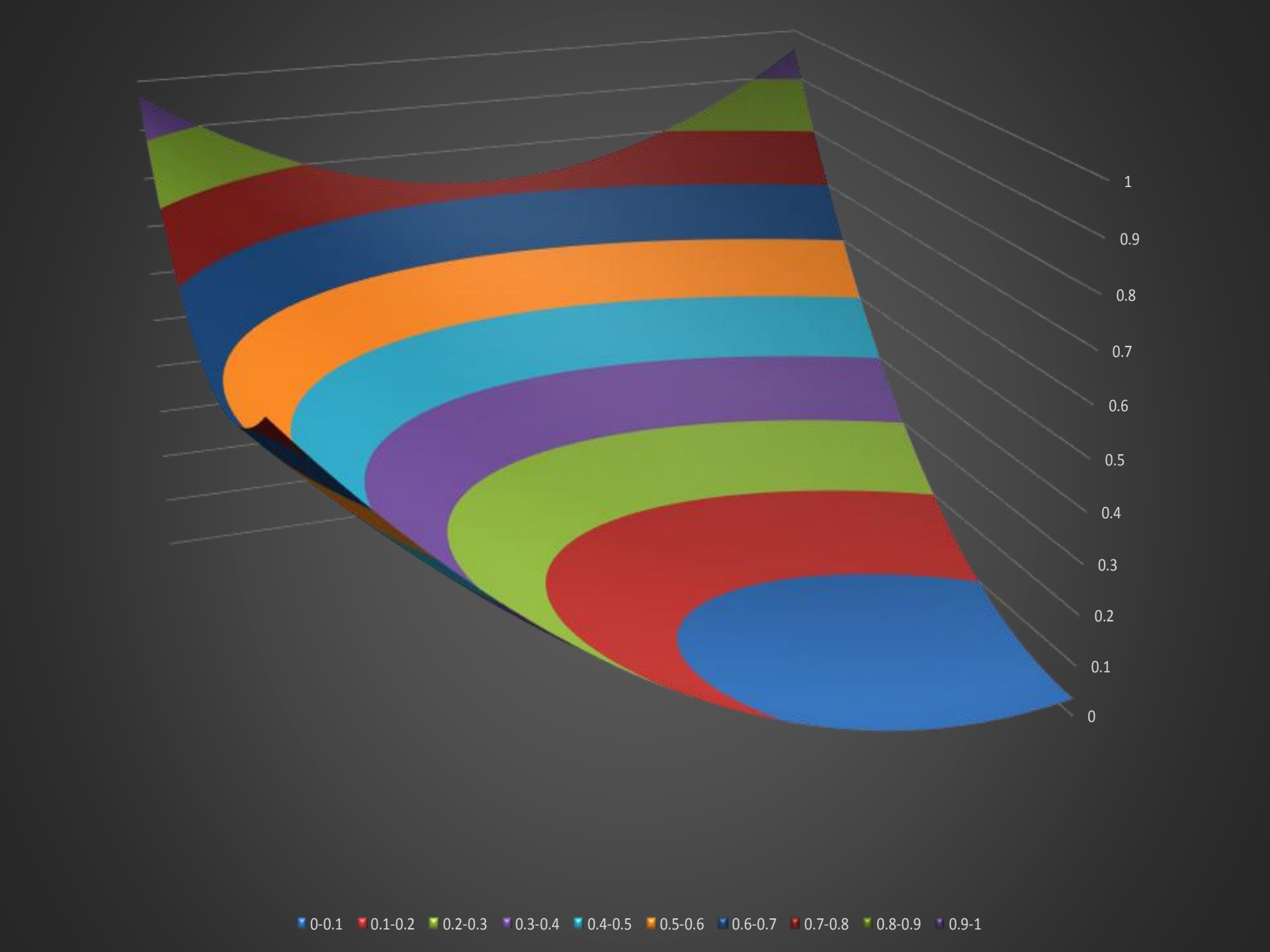
X= Design Matrix

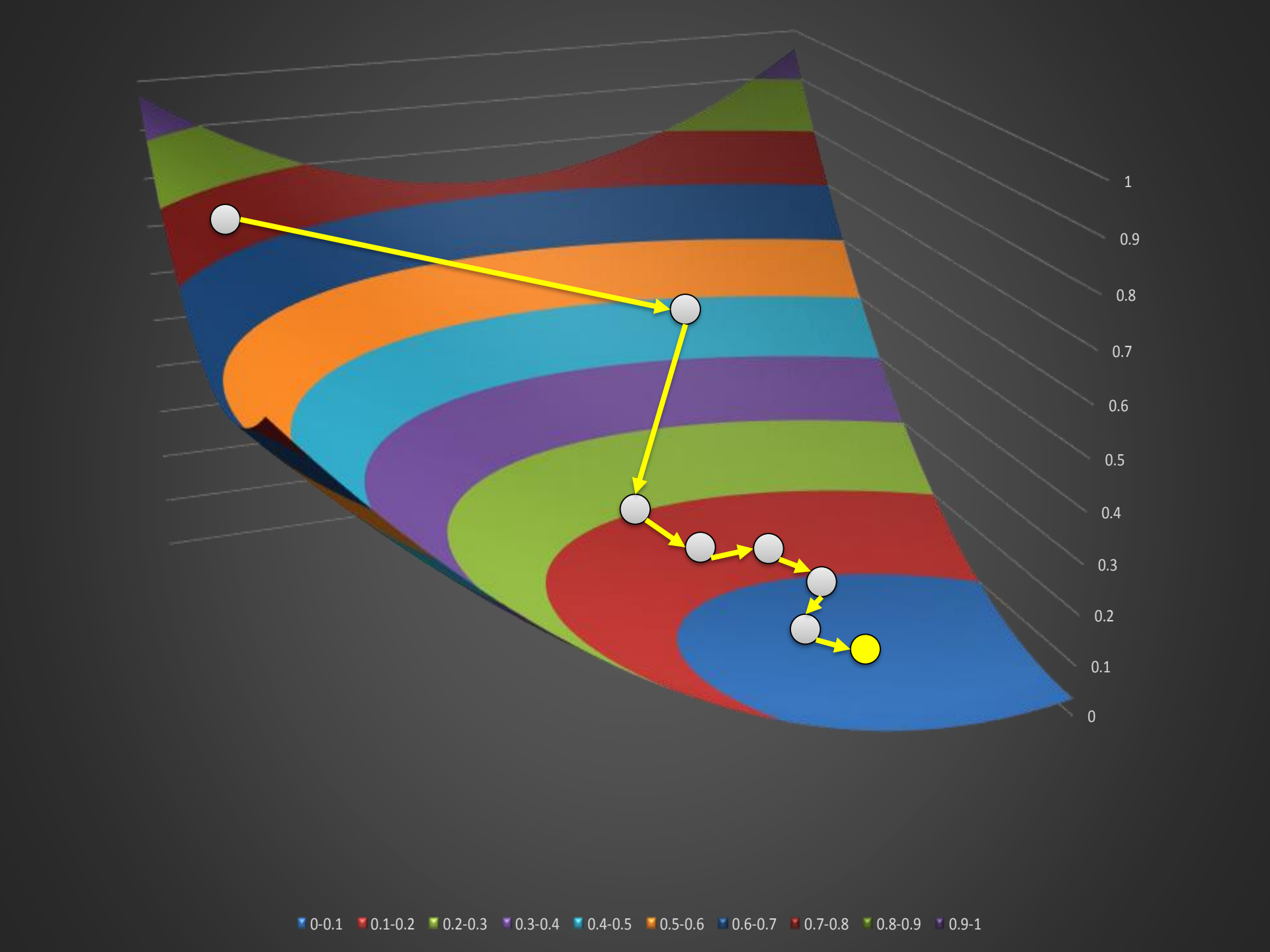
V_m= Diagonal Matrix, $V_{ii} = p_i \times (1 - p_i)$

Y= Dependent variable column matrix

P_m= Occurrence column matrix


0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1

Forecasting

$LL_{máx}$	-29.066
b_0	-30.202
b_1	0.220
b_2	2.767
b_3	-3.653
b_4	1.346
b_5	2.914

Forecasting

Once obtained the regression coefficients, each one must be validated. First we need to determine the standard error of each B_i, using the Variance-covariance Matrix S:

$$S = (X'VX)^{-1}$$

The square root of the diagonal matrix S, provides the standard error values of each B_i (except for B_0). Delete predictors that have a standard error higher than 2, because it may be an indicator of multicollinearity.

SE_b	
b_0	9.981
b_1	0.110
b_2	0.922
b_3	0.878
b_4	0.748
b_5	1.179

Forecasting

We need to check the signicance of each B_i using the z Wald hypothesis test:

$$H_0: B_i = 0$$
$$H_1: B_i \neq 0$$

$$W_{test} = \frac{B_i}{SE_{B_i}}$$
$$W_{critical} = z_{\alpha/2}$$

Reject the null hypothesis H_0 if $W_{test} > W_{critical}$ or $W_{test} < -W_{critical}$

Forecasting

Therefore for a $W_{critical} = z_{\alpha/2} = 1.96$:

$$W_0 = \frac{B_0}{SE_{B_0}} = \frac{-30.202}{9.981} = -3.026$$ Reject the null hypothesis.

$$W_1 = \frac{B_1}{SE_{B_1}} = \frac{0.202}{0.110} = 2.000$$ Reject the null hypothesis.

$$W_2 = \frac{B_2}{SE_{B_2}} = \frac{2.767}{0.922} = 3.001$$ Reject the null hypothesis.

$$W_3 = \frac{B_3}{SE_{B_3}} = \frac{-3.653}{0.878} = -4.161$$ Reject the null hypothesis.

$$W_4 = \frac{B_4}{SE_{B_4}} = \frac{1.346}{0.748} = 1.799$$ Accept the null hypothesis.

$$W_5 = \frac{B_5}{SE_{B_5}} = \frac{2.914}{1.179} = 2.472$$ Reject the null hypothesis.

Forecasting

Excluding the variable x_4:

LL_{max}	-30.800
b_0	-30.933
b_1	0.204
b_2	2.920
b_3	-3.776
b_5	2.459

Forecasting

Therefore for a $W_{critical} = z_{\alpha/2} = 1.96$:

$$W_0 = \frac{B_0}{SE_{B_0}} = \frac{-30.933}{10.636} = -2.909$$ Reject the null hypothesis.

$$W_1 = \frac{B_1}{SE_{B_1}} = \frac{0.204}{0.101} = 2.020$$ Reject the null hypothesis.

$$W_2 = \frac{B_2}{SE_{B_2}} = \frac{2.920}{1.011} = 2.888$$ Reject the null hypothesis.

$$W_3 = \frac{B_3}{SE_{B_3}} = \frac{-3.776}{0.847} = -4.458$$ Reject the null hypothesis.

$$W_5 = \frac{B_5}{SE_{B_5}} = \frac{2.459}{1.139} = 2.159$$ Reject the null hypothesis.

Forecasting

Variance-covariance Matrix S:

$$S = (X'VX)^{-1}$$

SE_b	
b_0	10.636
b_1	0.101
b_2	1.011
b_3	0.847
b_5	1.139

Forecasting

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:

$$B_0 \pm z_{2.5\%} \times SE_{B0} = [-51.782; -10.088]$$

$$B_1 \pm z_{2.5\%} \times SE_{B1} = [0.006; 0.402]$$

$$B_2 \pm z_{2.5\%} \times SE_{B2} = [0.938; 4.902]$$

$$B_3 \pm z_{2.5\%} \times SE_{B3} = [-5.436; -2.116]$$

$$B_5 \pm z_{2.5\%} \times SE_{B5} = [0.227; 4.691]$$

Goodness of Fit

Forecasting

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:

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

LL0	-67.686
b0	0.364

Forecasting

The likelihood ratio test is calculated by:

$$\chi^2{}_{test} = -2(LL_0 - LL_{max})$$

$$\chi^2{}_{critical} = \chi^2{}_{k;\alpha}$$

Where:

$\chi^2{}_{k;\alpha}$ = One-sided chi-square test for k predictors at a certain level of significance α.

The test can also be used to compare different models:

$$\chi^2{}_{test} = -2(LL_{Final\ Model} - LL_{Other\ Model})$$

χ^2 Table

	α		
G.L	10%	5%	1%
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566

χ^2 Table

	α		
G.L	10%	5%	1%
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566

Forecasting

So the following hypothesis can be tested:

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

$$\chi^2{}_{test} = -2(LL_0 - LL_{max})$$

Reject the null hypothesis H_0 if $\chi^2{}_{test} > \chi^2{}_{k;\alpha}$. Therefore:

$$\chi^2{}_{k;\alpha} = \chi^2{}_{4;5\%} = 9.488$$
$$\chi^2{}_{test} = -2(LL_0 - LL_{max})$$

$$\chi^2{}_{test} = -2[-67.686 - (-30.800)] = 74.136$$

Rejec the null hypothesis.

Forecasting

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.

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

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

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

Forecasting

$$pseudo(r^2)_{MacFadden} = \left(\frac{-2 \times (-67.686) + 2 \times (-30.800)}{-2 \times (-67.686)}\right) = 0.545$$

$$pseudo(r^2)_{Cox\ \&\ Snell} = 1 - \left(\frac{e^{-67.686}}{e^{-30.800}}\right)^{\frac{2}{100}} = 0.522$$

$$pseudo(r^2)_{Nagelkerke} = \frac{1 - \left(\frac{e^{-67.686}}{e^{-30.800}}\right)^{\frac{2}{100}}}{1 - (e^{-67.686})^{\frac{2}{100}}} = 0.703$$

Forecasting

The AIC (Akaike Information Criterion) and BIC (Bayesian information criterion) are both information criteria that serve only to compare different models. The lower the value, the better the model:

$$AIC = -2LL_0 + 2(p+q) \rightarrow AIC = -2(-67.686) + 2(4+1) = 145.372$$
$$BIC = -2LL_0 + (p+q)\ln n \rightarrow BIC = -2(-67.686) + (4+1)\ln 100 = 158.398$$

$$AIC = -2LL_{max} + 2(p+q) \rightarrow AIC = -2(-30.800) + 2(4+1) = 71.600$$
$$BIC = -2LL_{max} + (p+q)\ln n \rightarrow BIC = -2(-30.800) + (4+1)\ln 100 = 84.626$$

Where:

p = Quantity of betas (excluding the intercept);

q = Quantity of intercepts.

Interpretation

Forecasting

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

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

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

Forecasting

$e^{B_1} = e^{0.204} = 1.226 \therefore$ 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 a moderate profile

Forecasting

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

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

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

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

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

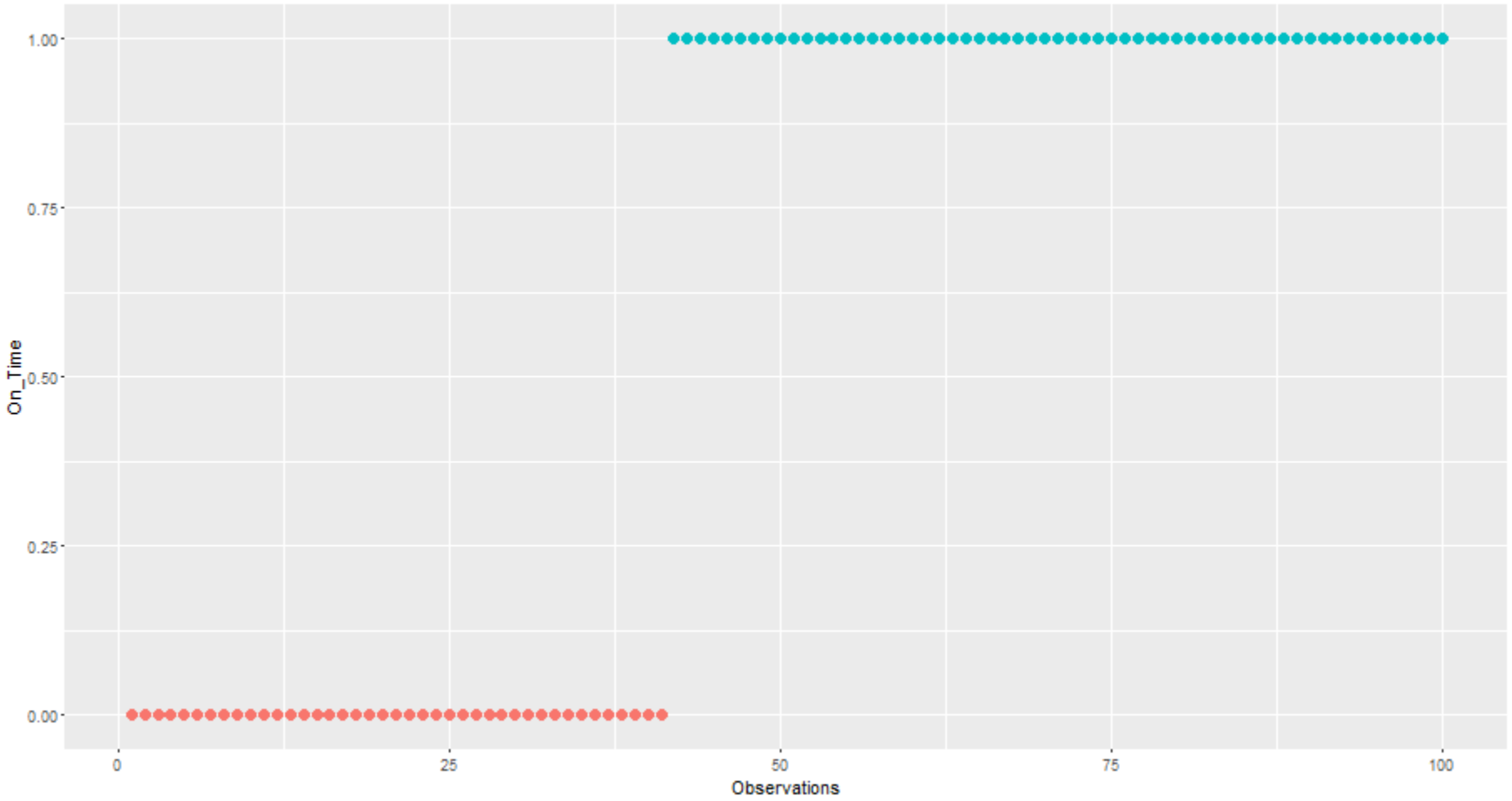
Forecasting

Cutoff de 0.5

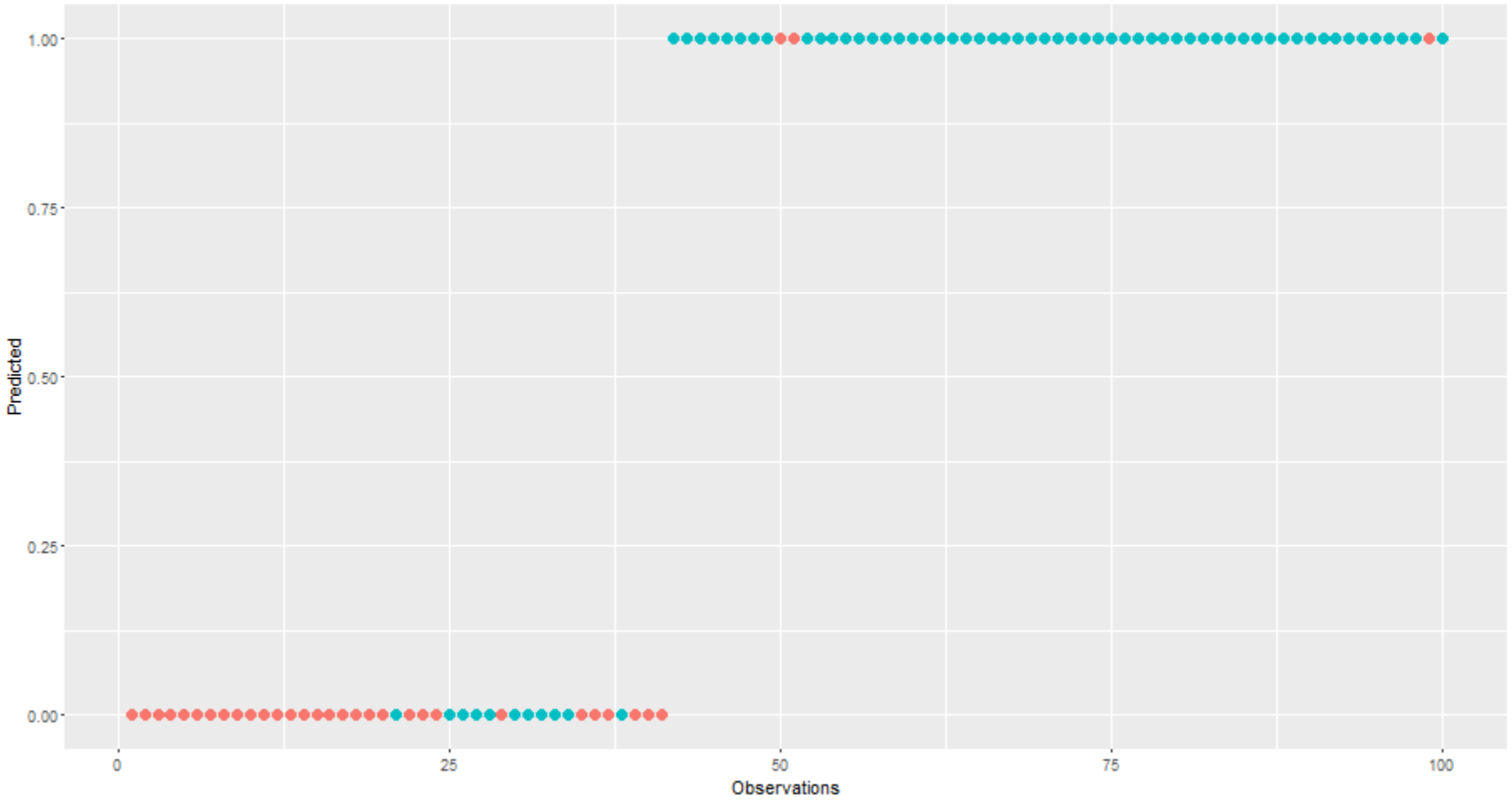
Id	Studant	Y (Late?) Yes = 1; No = 0	Probability i	Prediction
1	Gabriela	0	8.01978E-06	0
2	Patrícia	0	0.037039567	0
3	Gustavo	0	0.000597068	0
4	Letícia	0	9.03594E-05	0
5	Luiz Ovídio	0	6.58771E-05	0
6	Leonor	0	0.038642641	0
7	Dalila	0	0.038642641	0
8	Antônio	0	0.057559687	0
9	Júlia	0	0.049747133	0
10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

≠ *Classification*

Forecasting



Forecasting



Forecasting

Confusion Matrix:

	0	1	
0	TN (*True Negative*)	FN (*False Negative*)	PN (*Predicted Negative*)
1	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

$PN = TN + FN;$

$PP = TP + FP;$

$ON = TN + FP;$

$OP = TP + FN;$

Forecasting

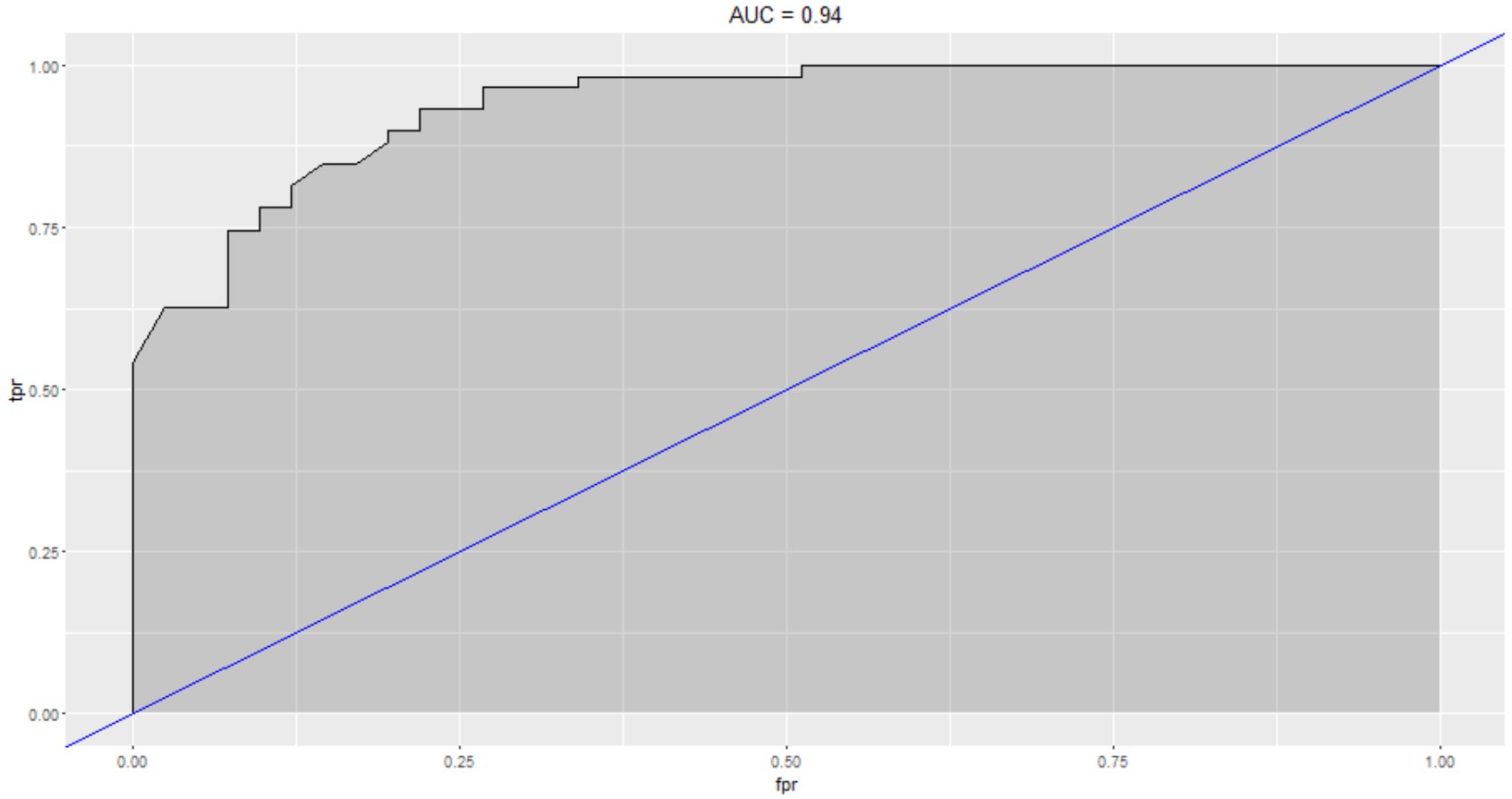
- $\boldsymbol{TPR}$ (*True Positive Rate*) = $\frac{TP}{OP}$
- $\boldsymbol{TNR}$ (*True Negative Rate*) = $\frac{TN}{ON}$
- $\boldsymbol{ACC}$ (*Accuracy*) = $\frac{TP+TN}{n}$;
- $\boldsymbol{FPR}$ (*False Positive Rate*) = $1 - TNR$ ou $\frac{FP}{ON}$
- $\boldsymbol{PPV}$ (*Positive Prediticed Value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

Forecasting

	0	1	
0	TN = 30	FN = 3	PN = 33
1	FP = 11	TP = 56	PP = 67
	ON = 41	OP = 59	n = 100

Forecasting

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.



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.