**LOGISTIC REGRESSION USING R & MINITAB**

1. **Logistic Regression with One Continuous Predictor**

Example: Use the “Effect of Age on Disease” data set in lecture (construct an excel table named “Patient Data.xlsx” with three columns: Patient, Age, and Disease)

Load the library “readxl” to read excel file

Command:

> setwd("D:/Program Problem/EU Plus Project/WP3/Course 9 \_ Data Analytic/Power Point")

! Change the working environment to the folder where the data file is located

> PatientData <- read\_excel("Patient Data.xlsx")

! Read data file and assign a name

> RegressionResult = glm(Disease~Age, data = PatientData, family = binomial)

! Create regression model with the command “**glm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

glm(formula = Disease ~ Age, family = binomial, data = PatientData)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6136 -0.6591 -0.4310 0.7856 1.8118

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.37210 1.96555 -2.224 0.0261 \*

Age 0.06696 0.03223 2.077 0.0378 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 25.898 on 19 degrees of freedom

Residual deviance: 20.201 on 18 degrees of freedom

AIC: 24.201

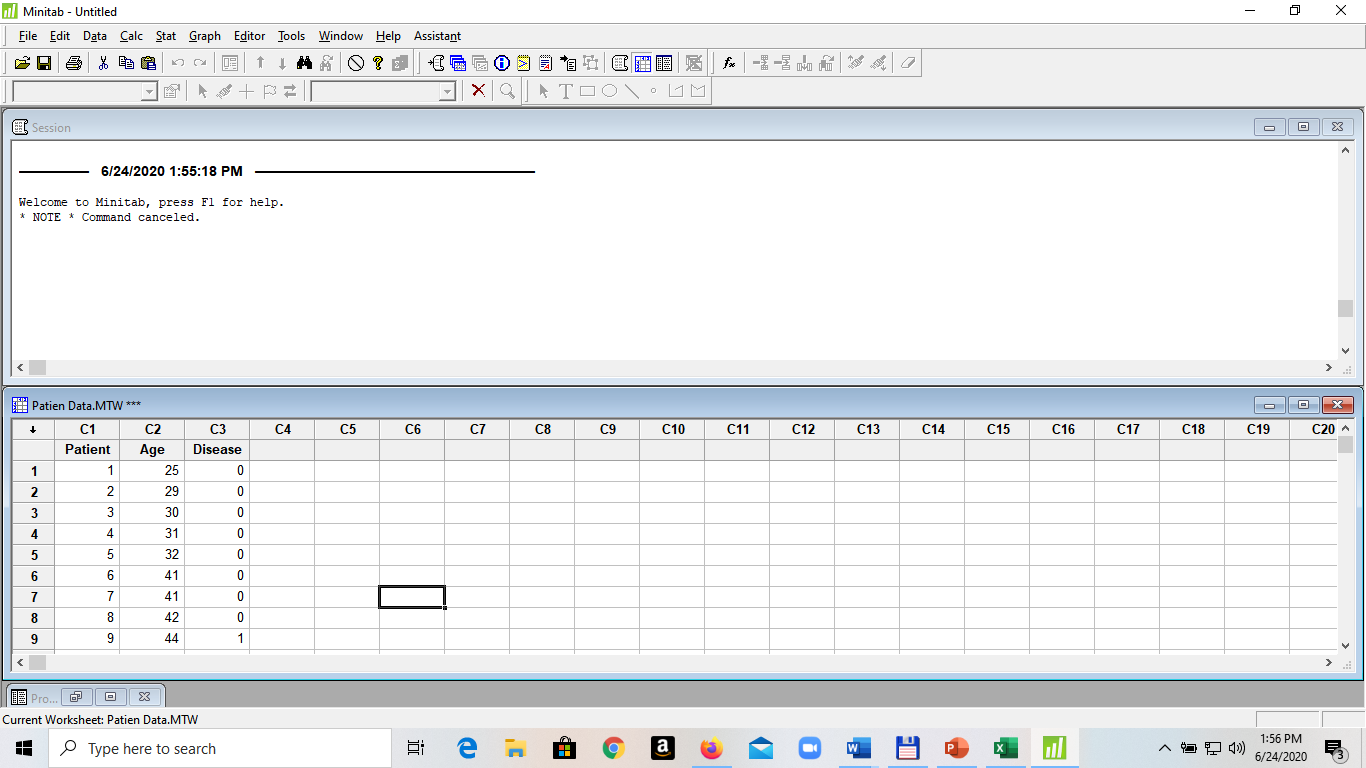
Number of Fisher Scoring iterations: 4

**Notes:**

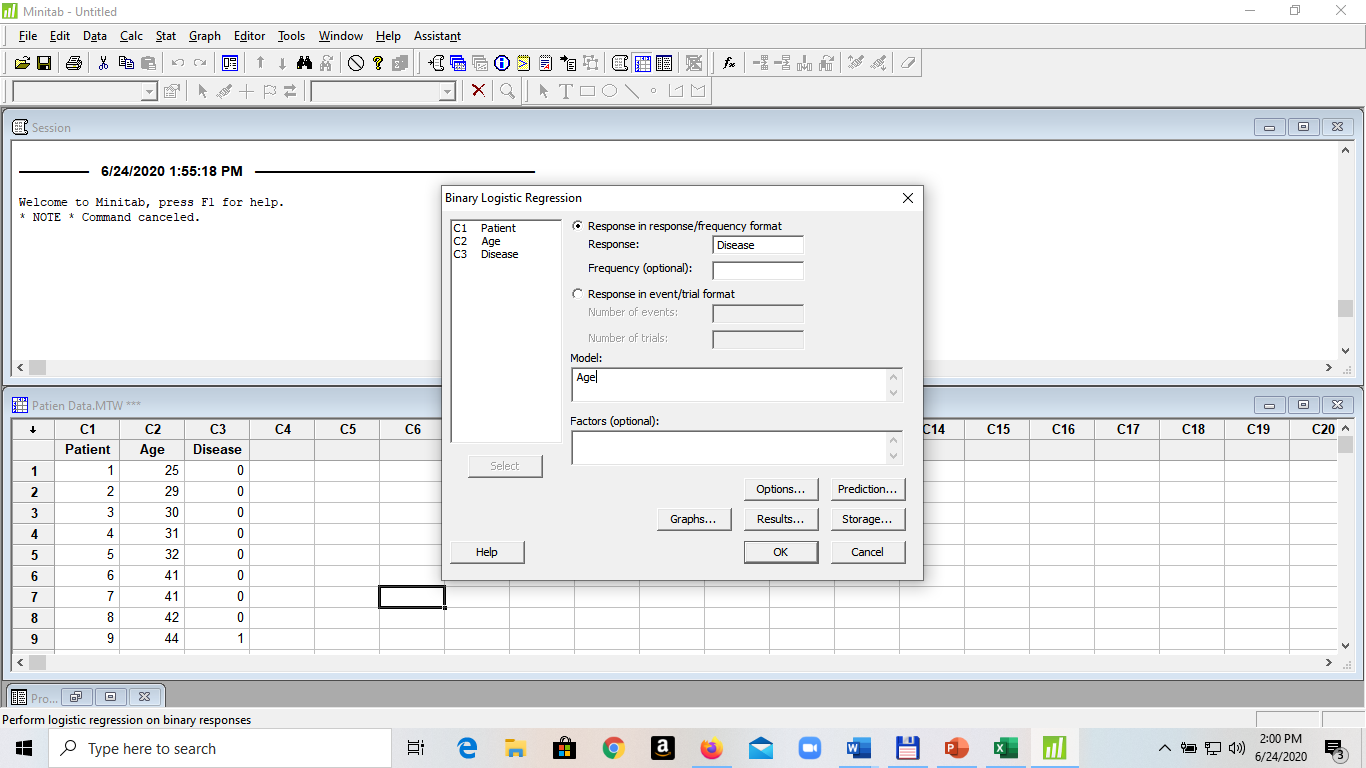
1. The null deviance is the deviance of model without predictors which shows how well the response variable is predicted by a model that includes only the intercept (grand mean). The residual deviance is the deviance of model with predictors. So: G = 25.898 – 20.201 = 5.697. In fact, the command: anova(RegressionResult) can be used to find deviance
2. Fisher’s scoring algorithm is a derivative of Newton’s method for solving maximum likelihood problems numerically. For this problem we see that Fisher’s Scoring Algorithm needed four iterations to perform the fit. This doesn’t really tell you a lot that you need to know, other than the fact that the model did indeed converge, and had no trouble doing it.
3. The Akaike Information Criterion (AIC) provides a method for assessing the quality of your model through comparison of related models.  It’s based on the “Deviance”, but penalizes you for making the model more complicated.  Much like adjusted R-squared, it’s intent is to prevent you from including irrelevant predictors. However, unlike adjusted R-squared, the number itself is not meaningful. If you have more than one similar candidate models (where all of the variables of the simpler model occur in the more complex models), then you should select the model that has the smallest AIC. So, it’s useful for comparing models, but isn’t interpretable on its own.

**USING MINITAB**

1. Declare data and save as “Patient Data.MTW”



1. Open **Stat > Regression > Binary Logistic Regression** and input variables for the model



Result:

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant -4.37210 1.96557 -2.22 0.026

Age 0.0669561 0.0322335 2.08 0.038 1.07 1.00 1.14

Log-Likelihood = -10.101

Test that all slopes are zero: G = 5.696, DF = 1, P-Value = 0.017

1. **Logistic Regression with Two Continuous Predictors**

Example: An experiment consisting of numeracy test scores (numeracy), scores on an anxiety test (anxiety), and a binary outcome variable (success) that records whether or not the students eventually succeeded in gaining admission to a university through an admissions test

Construct an excel table named “Admission Test.xlsx” with the data given below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Numeracy | Anxiety | Success | Numeracy | Anxiety | Success |
| 6.6 | 13.8 | 0 | 10.6 | 16.6 | 0 |
| 7.1 | 14.6 | 0 | 10.6 | 16.9 | 0 |
| 7.3 | 17.4 | 0 | 10.7 | 15.4 | 0 |
| 7.5 | 14.9 | 1 | 10.8 | 13.1 | 1 |
| 7.9 | 13.4 | 0 | 11 | 17.3 | 0 |
| 7.9 | 13.5 | 1 | 11.1 | 13.1 | 1 |
| 8 | 13.8 | 0 | 11.2 | 14 | 0 |
| 8.2 | 16.6 | 0 | 11.3 | 17.7 | 0 |
| 8.3 | 13.5 | 1 | 12 | 10.6 | 1 |
| 8.3 | 15.7 | 0 | 12.3 | 14.7 | 1 |
| 8.4 | 13.6 | 1 | 12.4 | 10.1 | 1 |
| 8.4 | 14 | 1 | 12.8 | 11.6 | 1 |
| 8.6 | 16.1 | 0 | 12.8 | 14.2 | 1 |
| 8.7 | 10.5 | 1 | 12.9 | 12.1 | 1 |
| 8.8 | 16.9 | 0 | 13.4 | 13.9 | 1 |
| 8.8 | 17.4 | 0 | 13.5 | 11.4 | 1 |
| 9.1 | 13.9 | 0 | 13.6 | 15.1 | 1 |
| 9.1 | 15.8 | 0 | 13.8 | 13 | 1 |
| 9.1 | 16.4 | 0 | 14.2 | 11.3 | 1 |
| 9.3 | 14.7 | 1 | 14.3 | 11.4 | 1 |
| 9.5 | 15 | 0 | 14.5 | 10.4 | 1 |
| 9.8 | 13.3 | 0 | 14.6 | 14.4 | 1 |
| 10.1 | 10.9 | 1 | 15 | 11 | 1 |
| 10.5 | 12.4 | 1 | 15.1 | 14 | 1 |
| 10.6 | 12.9 | 1 | 15.7 | 13.4 | 1 |

Load the library “readxl” to read excel file

Command:

> setwd("D:/Program Problem/EU Plus Project/WP3/Course 9 \_ Data Analytic/Power Point")

! Change the working environment to the folder where the data file is located

> TestScore <- read\_excel("Admission Test.xlsx")

! Read data file and assign a name

> RegressionResult = glm(Success~Numeracy + Anxiety, data = TestScore, family = binomial)

! Create regression model with the command “**glm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

glm(formula = Success ~ Numeracy + Anxiety, family = binomial,

data = TestScore)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.83958 -0.30510 0.04823 0.35431 2.08545

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 14.2386 6.7985 2.094 0.03623 \*

Numeracy 0.5774 0.2481 2.327 0.01995 \*

Anxiety -1.3841 0.4804 -2.881 0.00396 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 68.029 on 49 degrees of freedom

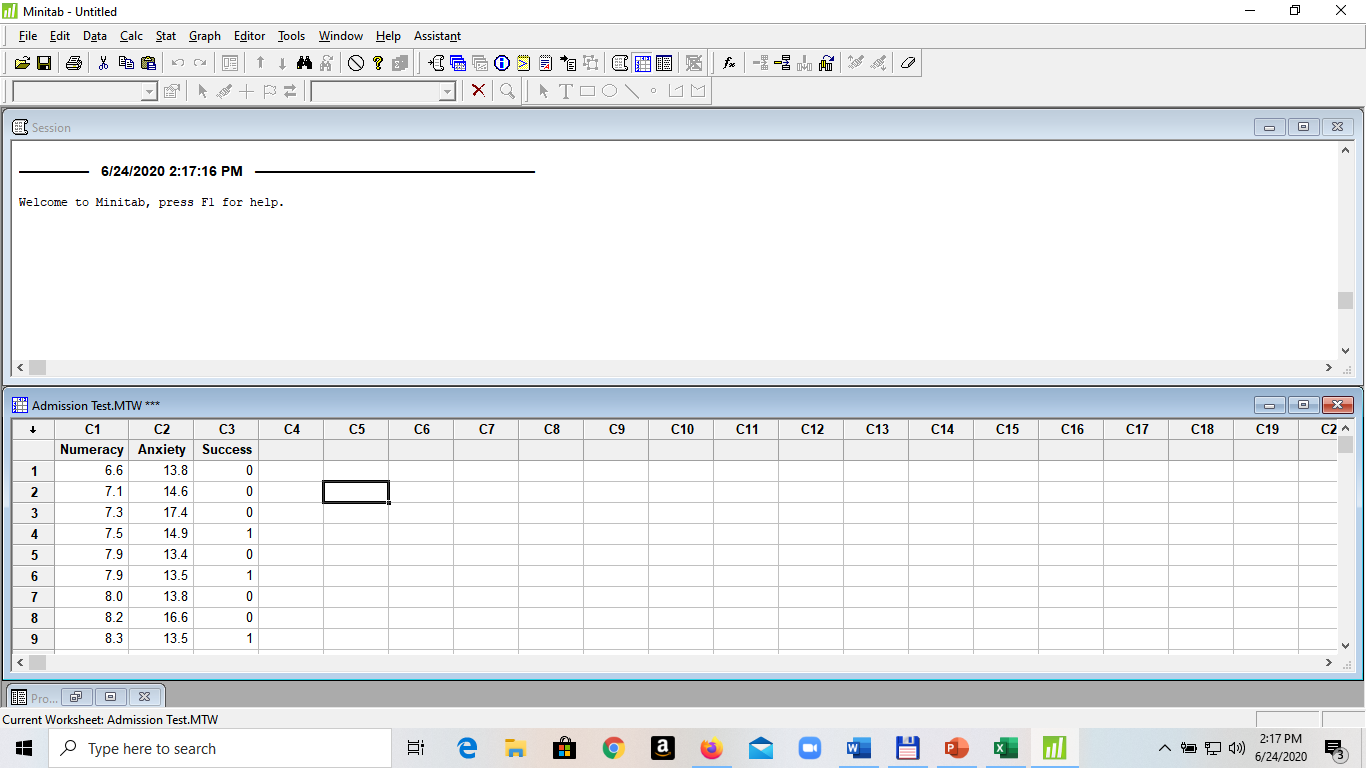
Residual deviance: 28.286 on 47 degrees of freedom

AIC: 34.286

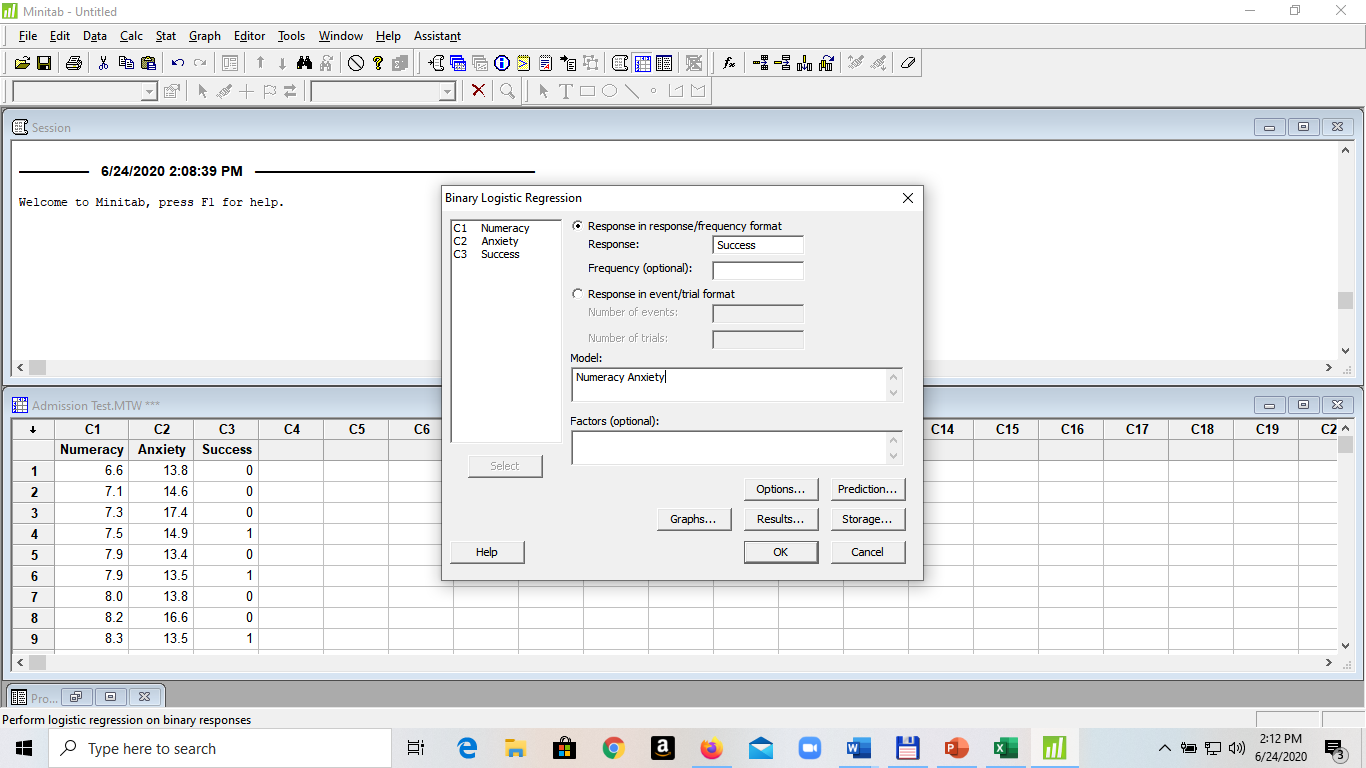
Number of Fisher Scoring iterations: 6

**USING MINITAB**

1. Declare data and save as “Admission Test.MTW”



1. Open **Stat > Regression > Binary Logistic Regression** and input variables for the model



Result:

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant 14.2386 6.79923 2.09 0.036

Numeracy 0.577352 0.248098 2.33 0.020 1.78 1.10 2.90

Anxiety -1.38407 0.480465 -2.88 0.004 0.25 0.10 0.64

Log-Likelihood = -14.143

Test that all slopes are zero: G = 39.744, DF = 2, P-Value = 0.000

**If interaction term is incorporated:**

> RegressionResult = glm(Success~Numeracy\*Anxiety, data = TestScore, family = binomial)

! Create regression model with the command “**glm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

glm(formula = Success ~ Numeracy \* Anxiety, family = binomial,

data = TestScore)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.85712 -0.33055 0.02531 0.34931 2.01048

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.87883 46.45256 0.019 0.985

Numeracy 1.94556 4.78250 0.407 0.684

Anxiety -0.44580 3.25151 -0.137 0.891

Numeracy:Anxiety -0.09581 0.33322 -0.288 0.774

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 68.029 on 49 degrees of freedom

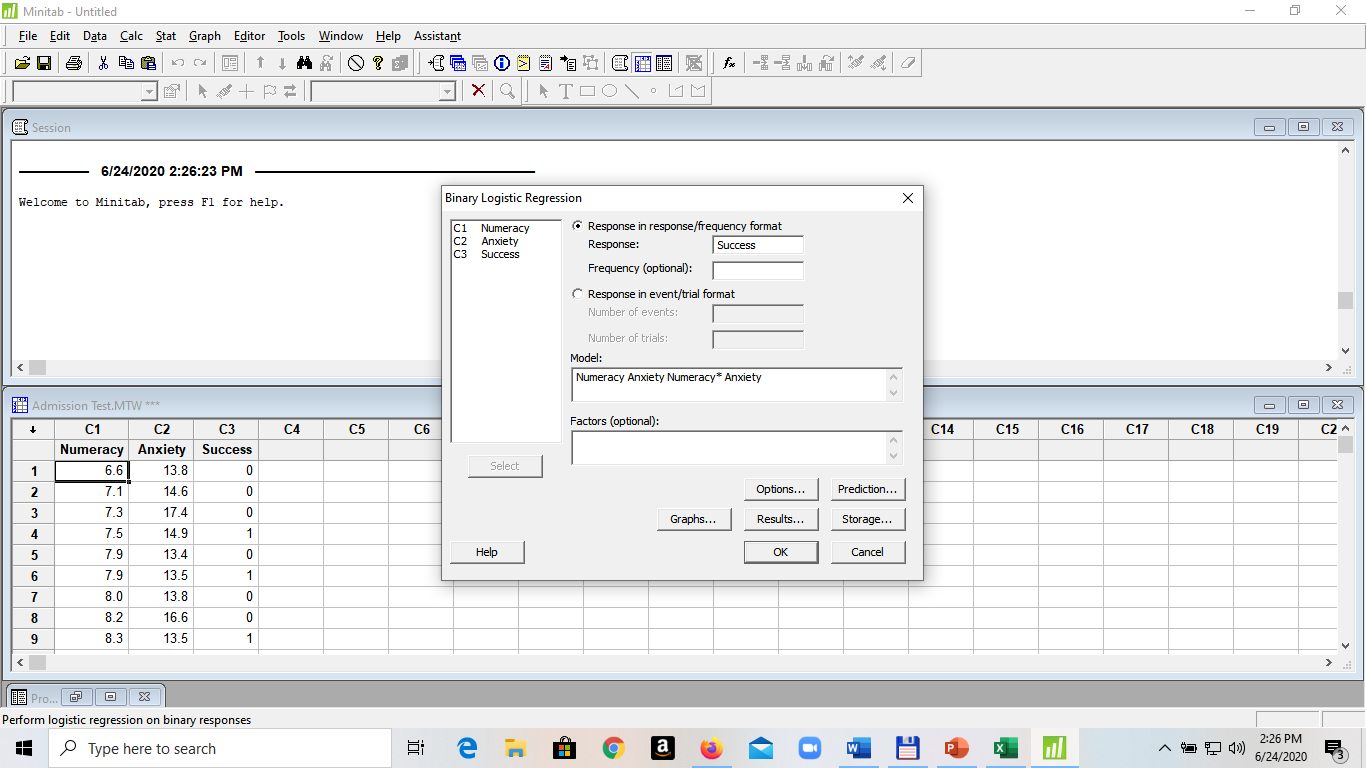
Residual deviance: 28.201 on 46 degrees of freedom

AIC: 36.201

Number of Fisher Scoring iterations: 7

**USING MINITAB**

Input variables for the model:



Result:

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant 0.878828 46.4542 0.02 0.985

Numeracy 1.94556 4.78272 0.41 0.684 7.00 0.00 82415.23

Anxiety -0.445803 3.25162 -0.14 0.891 0.64 0.00 375.13

Numeracy\*Anxiety -0.0958116 0.333237 -0.29 0.774 0.91 0.47 1.75

Log-Likelihood = -14.100

Test that all slopes are zero: G = 39.828, DF = 3, P-Value = 0.000

**Remarks: The model without interaction term is better**

**(look at p-values of coefficients and also AIC)**

Note: Data can be input using data frame as follows

TestScore <- structure(list(numeracy = c(6.6, 7.1, 7.3, 7.5, 7.9, 7.9, 8,

8.2, 8.3, 8.3, 8.4, 8.4, 8.6, 8.7, 8.8, 8.8, 9.1, 9.1, 9.1, 9.3,

9.5, 9.8, 10.1, 10.5, 10.6, 10.6, 10.6, 10.7, 10.8, 11, 11.1,

11.2, 11.3, 12, 12.3, 12.4, 12.8, 12.8, 12.9, 13.4, 13.5, 13.6,

13.8, 14.2, 14.3, 14.5, 14.6, 15, 15.1, 15.7), anxiety = c(13.8,

14.6, 17.4, 14.9, 13.4, 13.5, 13.8, 16.6, 13.5, 15.7, 13.6, 14,

16.1, 10.5, 16.9, 17.4, 13.9, 15.8, 16.4, 14.7, 15, 13.3, 10.9,

12.4, 12.9, 16.6, 16.9, 15.4, 13.1, 17.3, 13.1, 14, 17.7, 10.6,

14.7, 10.1, 11.6, 14.2, 12.1, 13.9, 11.4, 15.1, 13, 11.3, 11.4,

10.4, 14.4, 11, 14, 13.4), success = c(0L, 0L, 0L, 1L, 0L, 1L,

0L, 0L, 1L, 0L, 1L, 1L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L,

1L, 1L, 1L, 0L, 0L, 0L, 1L, 0L, 1L, 0L, 0L, 1L, 1L, 1L, 1L, 1L,

1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L)), .Names = c("Numeracy",

"Anxiety", "Success"), row.names = c(NA, -50L), class = "data.frame")

(Declaration L: means integer value)

1. **Logistic Regression with A Dichotomous Predictor**

Example: Use the “Churn by membership in the Voice Mail Plan” data set in lecture (Slide 55, construct an excel table named “Churn\_Voice Mail.xlsx” with three columns: Vmail, Churn, Frequency)

|  |  |  |
| --- | --- | --- |
| Vmail | Churn | Frequency |
| 0 | 0 | 2008 |
| 0 | 1 | 403 |
| 1 | 0 | 842 |
| 1 | 1 | 80 |

Load the library “readxl” to read excel file

Command:

> setwd("D:/Program Problem/EU Plus Project/WP3/Course 9 \_ Data Analytic/Power Point")

! Change the working environment to the folder where the data file is located

> ChurnData <- read\_excel("Churn\_Voice Mail.xlsx")

! Read data file and assign a name

> RegressionResult = glm(Churn~Vmail, data=ChurnData, family=binomial, weights=Frequency)

! Create regression model with the command “**glm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

glm(formula = Churn ~ Vmail, family = binomial, data = ChurnData,

weights = Frequency)

Deviance Residuals:

1 2 3 4

-27.10 37.97 -12.36 19.78

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.60596 0.05458 -29.422 < 2e-16 \*\*\*

Vmail -0.74780 0.12907 -5.794 6.89e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2758.3 on 3 degrees of freedom

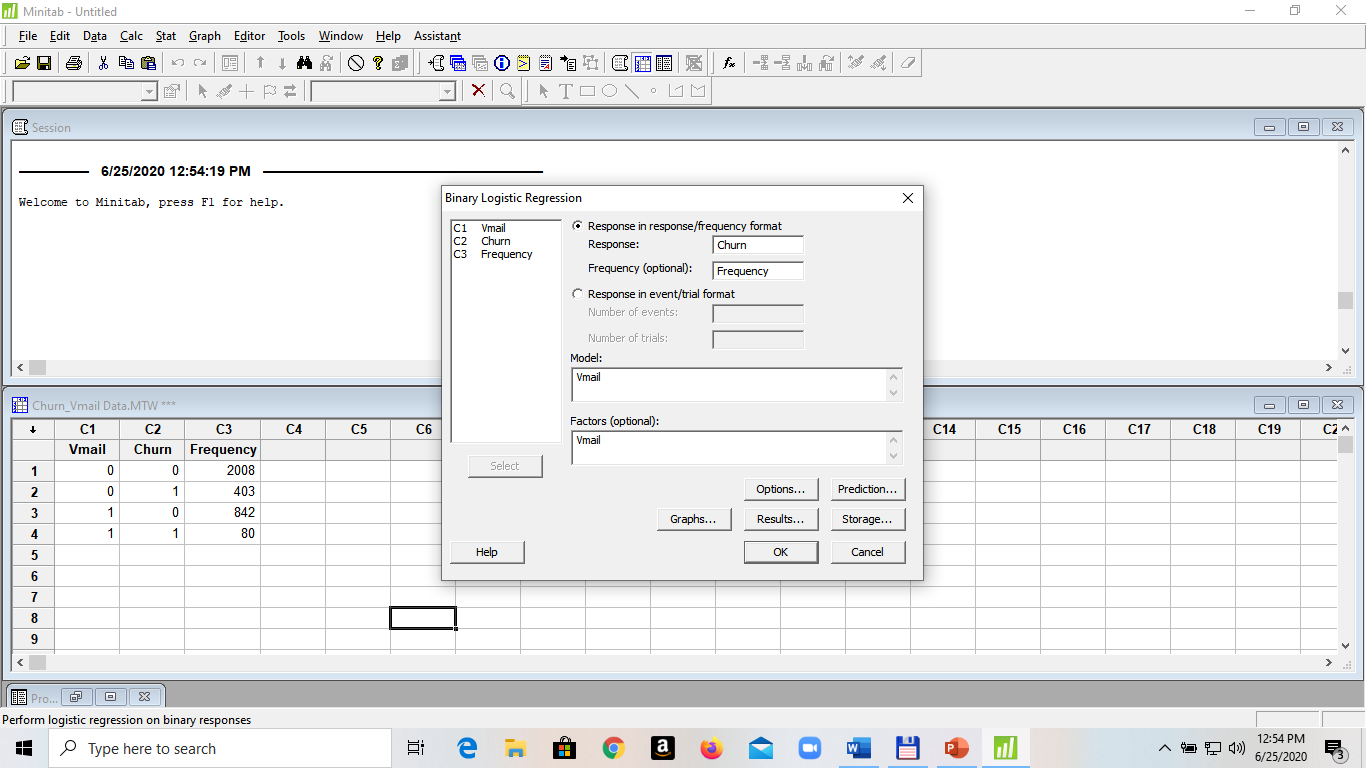
Residual deviance: 2720.3 on 2 degrees of freedom

AIC: 2724.3

Number of Fisher Scoring iterations: 5

**USING MINITAB**

Input variables for the model:



Result:

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant -1.60596 0.0545839 -29.42 0.000

Vmail

1 -0.747795 0.129101 -5.79 0.000 0.47 0.37 0.61

Log-Likelihood = -1360.165

Test that all slopes are zero: G = 37.964, DF = 1, P-Value = 0.000

1. **Logistic Regression with A Polychotomous Predictor**

Example: Use the “Churn – Customer Service Call” data set in lecture (Slide 65, construct an excel table named “Churn\_Service Call.xlsx” with four columns: CSC\_Medium, CSC\_High, Churn, Frequency)

|  |  |  |  |
| --- | --- | --- | --- |
| CSC\_Medium | CSC\_High | Churn | Frequency |
| 0 | 0 | 0 | 1664 |
| 0 | 0 | 1 | 214 |
| 1 | 0 | 0 | 1057 |
| 1 | 0 | 1 | 131 |
| 0 | 1 | 0 | 129 |
| 0 | 1 | 1 | 138 |

Load the library “readxl” to read excel file

Command:

> setwd("D:/Program Problem/EU Plus Project/WP3/Course 9 \_ Data Analytic/Power Point")

! Change the working environment to the folder where the data file is located

> ChurnData <- read\_excel("Churn\_Service Call.xlsx")

! Read data file and assign a name

> RegressionResult = glm(Churn~CSC\_Medium + CSC\_High, data=ChurnData, family=binomial, weights=Frequency)

! Create regression model with the command “**glm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

glm(formula = Churn ~ CSC\_Medium + CSC\_High, family = binomial,

data = ChurnData, weights = Frequency)

Deviance Residuals:

1 2 3 4 5 6

-20.07 30.49 -15.72 24.03 -13.70 13.50

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.05100 0.07262 -28.244 <2e-16 \*\*\*

CSC\_Medium -0.03699 0.11769 -0.314 0.753

CSC\_High 2.11844 0.14238 14.879 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2758.3 on 5 degrees of freedom

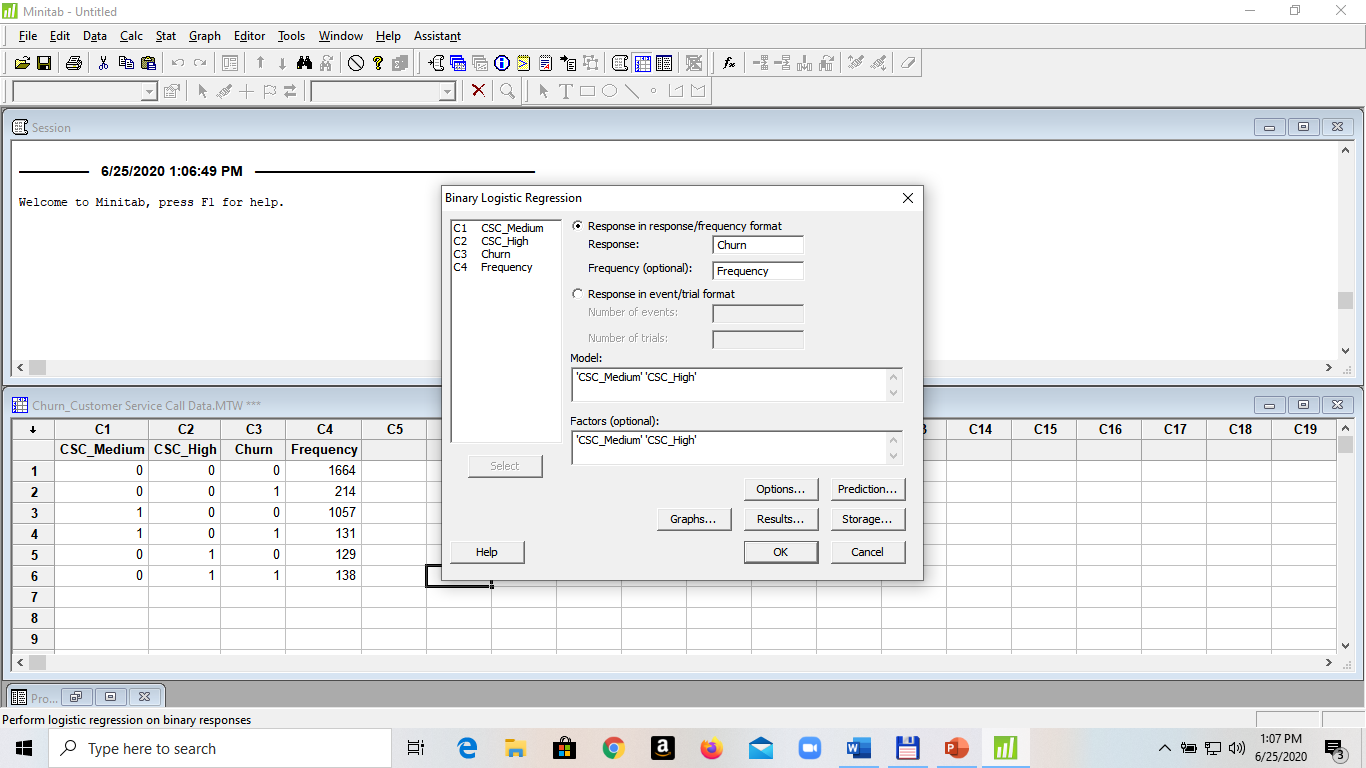
Residual deviance: 2526.7 on 3 degrees of freedom

AIC: 2532.7

Number of Fisher Scoring iterations: 5

**USING MINITAB**

Input variables for the model:



Result:

Logistic Regression Table

Odds 95% CI

Predictor Coef SE Coef Z P Ratio Lower Upper

Constant -2.05100 0.0726213 -28.24 0.000

CSC\_Medium

1 -0.0369891 0.117701 -0.31 0.753 0.96 0.77 1.21

CSC\_High

1 2.11844 0.142380 14.88 0.000 8.32 6.29 11.00

Log-Likelihood = -1263.368

Test that all slopes are zero: G = 231.557, DF = 2, P-Value = 0.000