

SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

MODULE 2 – PRACTICAL 14

AIM:- Performing logistic regression using glm() (R).

OUTPUT:-

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R - R452 - ~/bhl/
> # 1. Load Dataset
> data <- read.csv("C:\\Users\\info\\downloads\\house_prices_Texas_counties.csv")
> print("Dataset Loaded Successfully")
[1] "Dataset Loaded Successfully"
>
> # 2. Remove unwanted unnamed columns
> data <- data[, !grep("Unnamed", names(data))]
>
> # 3. Dataset overview
> head(data)
  Location Bedroom Total.Sqft Bathroom Price X X.1 X.2
1 Anderson      1         650         1 110000 NA NA NA
2 Andrews      1         700         1 153200 NA NA NA
3 Angelina     1         800         1 110500 NA NA NA
4 Aransas      1         650         1 183200 NA NA NA
5 Archer       1         700         1 156100 NA NA NA
6 Armstrong    1         800         1 106000 NA NA NA
> str(data)
'data.frame':   972 obs. of  8 variables:
 $ Location : chr  "Anderson" "Andrews" "Angelina" "Aransas" ...
 $ Bedroom  : int   1 1 1 1 1 1 1 1 ...
 $ Total.Sqft: int   650 700 800 650 700 800 650 700 800 800 ...
 $ Bathroom : int   1 1 1 1 1 1 1 1 ...
 $ Price    : num  110000 153200 110500 183200 156100 ...
 $ X        : logi  NA NA NA NA NA NA ...
 $ X.1      : logi  NA NA NA NA NA NA ...
 $ X.2      : logi  NA NA NA NA NA NA ...
> summary(data)
  Location      Bedroom      Total.Sqft      Bathroom      Price      X      X.1      X.2
Length:972      Min.   :1.00      Min.   : 650      Min.   :1.00      Min.   : 62560      Mode:logical      Mode:logical      Mode:logical
Class :character 1st Qu.:1.75      1st Qu.: 950      1st Qu.:1.75      1st Qu.: 180000      NA's:972      NA's:972      NA's:972
Mode :character  Median :2.50      Median :1560      Median :2.50      Median : 270000
Mean   :2.50      Mean   :1576      Mean   :2.50      Mean   : 302922
3rd Qu.:3.25      3rd Qu.:2010      3rd Qu.:3.25      3rd Qu.: 393565
Max.   :4.00      Max.   :2640      Max.   :4.00      Max.   :1350000
>
> # 4. Create Binary Outcome Variable (High vs Low Price)
> median_price <- median(data$Price, na.rm = TRUE)
> data$HighPrice <- ifelse(data$Price > median_price, 1, 0)
>
> # Convert Location to Factor
> data$Location <- as.factor(data$Location)
>
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R - R452 - ~/bhl/
>
> # 4. Create Binary Outcome Variable (High vs Low Price)
> median_price <- median(data$Price, na.rm = TRUE)
> data$HighPrice <- ifelse(data$Price > median_price, 1, 0)
>
> # Convert Location to Factor
> data$Location <- as.factor(data$Location)
>
> # 5. Logistic Regression Model (FIXED column name)
> model <- glm(
+   HighPrice ~ Bedroom + Bathroom + Total.Sqft + Location,
+   data = data,
+   family = binomial
+ )
warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
>
> # 6. Model Summary
> summary(model)

Call:
glm(formula = HighPrice ~ Bedroom + Bathroom + Total.Sqft + Location,
    family = binomial, data = data)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.693e+02  7.469e+04  -0.002  0.998
Bedroom      4.932e+01  2.117e+04   0.002  0.998
Bathroom     NA         NA         NA     NA
Total.Sqft   -1.749e-03  3.447e+01   0.000  1.000
LocationAndrews  9.676e+01  1.030e+05   0.001  0.999
LocationAngelina  5.031e-02  9.193e+04   0.000  1.000
LocationAransas  9.674e+01  1.014e+05   0.001  0.999
LocationArcher   9.679e+01  1.024e+05   0.001  0.999
LocationArmstrong 1.248e-01  9.138e+04   0.000  1.000
LocationAtascosa  9.677e+01  1.007e+05   0.001  0.999
LocationAustin   9.682e+01  1.017e+05   0.001  0.999
LocationBailey   -4.796e+01  1.295e+05   0.000  1.000
LocationBandera  9.691e+01  1.037e+05   0.001  0.999
LocationBastrop  1.457e+02  1.363e+05   0.001  0.999
```

SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

```
Source
Console Terminal Background Jobs
R - R 4.5.2 - ~/bhi/
Total.Sqft -1.749e-03 3.447e+01 0.000 1.000
LocationAndrews 9.676e+01 1.030e+05 0.001 0.999
LocationAngelina 5.031e-02 9.193e+04 0.000 1.000
LocationAransas 9.674e+01 1.034e+05 0.001 0.999
LocationArcher 9.679e+01 1.024e+05 0.001 0.999
LocationArmstrong 1.248e-01 9.138e+04 0.000 1.000
LocationAtascosa 9.677e+01 1.007e+05 0.001 0.999
LocationAustin 9.682e+01 1.017e+05 0.001 0.999
LocationBailey -4.796e+01 1.295e+05 0.000 1.000
LocationBandera 9.691e+01 1.037e+05 0.001 0.999
LocationBastrop 1.457e+02 1.363e+05 0.001 0.999
LocationBaylor 4.840e+01 1.109e+05 0.000 1.000
LocationBee 4.842e+01 1.109e+05 0.000 1.000
LocationBell 9.688e+01 1.004e+05 0.001 0.999
LocationBexar 9.696e+01 1.027e+05 0.001 0.999
LocationBlanco 1.457e+02 1.364e+05 0.001 0.999
LocationBosque 9.698e+01 1.022e+05 0.001 0.999
LocationBowie 4.253e-01 8.954e+04 0.000 1.000
LocationBrazoria 9.693e+01 9.943e+04 0.001 0.999
LocationBrazos 1.456e+02 1.344e+05 0.001 0.999
LocationBrewster 1.456e+02 1.345e+05 0.001 0.999
LocationBriscoe 2.581e-01 9.555e+04 0.000 1.000
LocationBrooks 2.827e-01 9.536e+04 0.000 1.000
LocationBrown 9.695e+01 9.734e+04 0.001 0.999
LocationBurlinson 9.700e+01 9.827e+04 0.001 0.999
LocationBurnet 9.697e+01 9.697e+04 0.001 0.999
LocationCaldwell 9.702e+01 9.790e+04 0.001 0.999
LocationCallahan 4.870e+01 1.112e+05 0.000 1.000
LocationCallahan 4.871e+01 1.113e+05 0.000 1.000
LocationCameron 9.712e+01 9.964e+04 0.001 0.999
LocationCamp 4.876e+01 1.113e+05 0.000 1.000
LocationCarson 4.877e+01 1.114e+05 0.000 1.000
LocationCass 4.878e+01 1.114e+05 0.000 1.000
LocationCastro 4.880e+01 1.114e+05 0.000 1.000
LocationChandler 9.717e+01 9.872e+04 0.001 0.999
LocationChapman 6.047e-01 9.328e+04 0.000 1.000
LocationCherokee 4.887e+01 1.116e+05 0.000 1.000
LocationChildress 3.971e-01 1.003e+05 0.000 1.000
LocationClay 4.890e+01 1.116e+05 0.000 1.000
LocationCochran 4.822e+01 1.133e+05 0.000 1.000
LocationColeman 4.894e+01 1.117e+05 0.000 1.000
LocationCollin 9.713e+01 9.426e+04 0.001 0.999
LocationCollinsworth 8.186e-01 9.828e+04 0.000 1.000
```

```
Source
Console Terminal Background Jobs
R - R 4.5.2 - ~/bhi/
LocationWashington 4.886e+01 1.088e+05 0.000 1.000
LocationWebb 4.887e+01 1.091e+05 0.000 1.000
LocationWharton 4.887e+01 1.094e+05 0.000 1.000
LocationWheeler 5.795e-01 9.649e+04 0.000 1.000
LocationWichita 9.718e+01 9.853e+04 0.001 0.999
LocationWilson 4.890e+01 1.104e+05 0.000 1.000
LocationWinkler 5.994e-01 9.590e+04 0.000 1.000
LocationWood 4.891e+01 1.111e+05 0.000 1.000
LocationYoakum 4.892e+01 1.114e+05 0.000 1.000
LocationYoung 6.189e-01 9.534e+04 0.000 1.000
LocationZapata 9.724e+01 9.750e+04 0.001 0.999
LocationZavala 9.725e+01 9.734e+04 0.001 0.999

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.3474e+03 on 971 degrees of freedom
Residual deviance: 2.7997e-08 on 727 degrees of freedom
AIC: 490

Number of Fisher Scoring iterations: 25

>
> # 7. Predictions
> logit_values <- predict(model, type = "link")
> prob_values <- predict(model, type = "response")
>
> # 8. Plot Logistic Regression Curve
> plot(
+ logit_values,
+ prob_values,
+ pch = 19,
+ col = "blue",
+ main = "Logistic Regression: High vs Low House Prices",
+ xlab = "Logit (Linear Predictor)",
+ ylab = "Predicted Probability"
+ )
>
> # 9. Add sigmoid curve
> logit_seq <- seq(min(logit_values), max(logit_values), length.out = 200)
> sigmoid <- 1 / (1 + exp(-logit_seq))
> lines(logit_seq, sigmoid, col = "red", lwd = 2)
>
```

SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

