Final - Homework 220

Phan Anh Le

2024-05-03

```
library(dplyr)
library(ggplot2)
library(rms)
library(ResourceSelection)
Data <- read.csv("C:/Users/admin/Downloads/heart.csv")
View(Data)
```

Introduction

In my final project, I will apply logistic regression to a dataset that provided the data of heart attack. This dataset contains variables as age, sex, chest pain type, resting blood pressure rate, cholestoral rate, fasting blood sugar, maximum heart rate achieved, vv. By applying the logistic regression, I aim to predict if the heart attack happens base on the cholestoral rate and the maximum heart rate achieved.

Validate variables

```
summary(Data)
```

```
##       age             sex              cp            trtbps     
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       chol            fbs             restecg          thalachh    
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##       exng           oldpeak          slp              caa        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
##      thall           output     
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000
```

```
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

Clean data

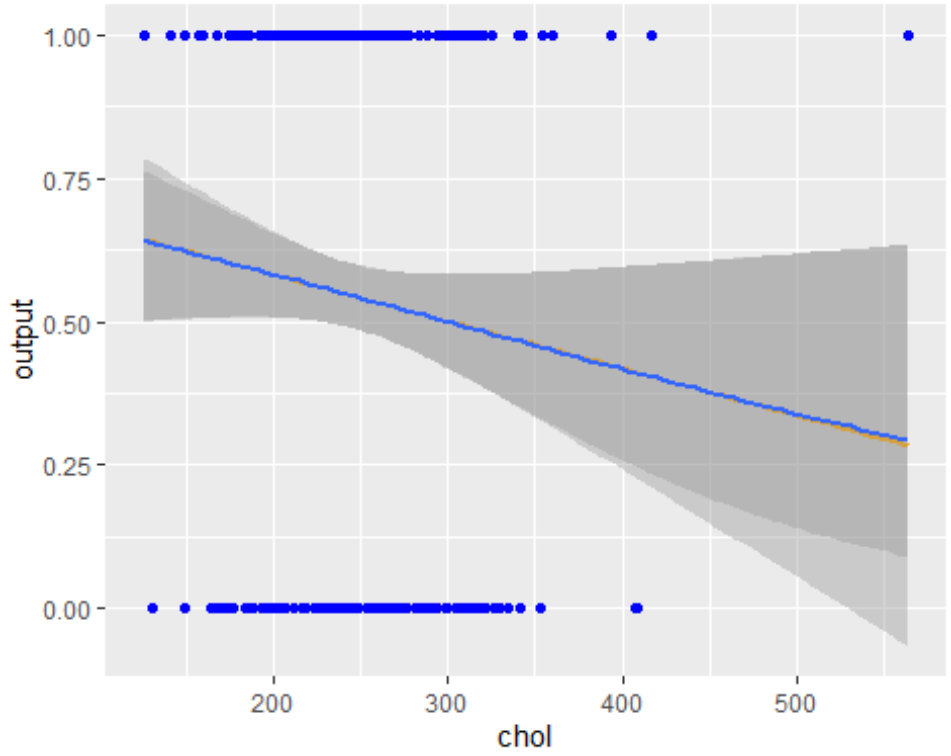
```
Data_new <- subset(Data, select = c(age, chol, thalachh, output))
Data_new <- na.omit(Data_new)
summary(Data_new)
```

```
##       age             chol          thalachh         output
##  Min.   :29.00   Min.   :126.0   Min.   : 71.0   Min.   :0.0000
##  1st Qu.:47.50   1st Qu.:211.0   1st Qu.:133.5   1st Qu.:0.0000
##  Median :55.00   Median :240.0   Median :153.0   Median :1.0000
##  Mean   :54.37   Mean   :246.3   Mean   :149.6   Mean   :0.5446
##  3rd Qu.:61.00   3rd Qu.:274.5   3rd Qu.:166.0   3rd Qu.:1.0000
##  Max.   :77.00   Max.   :564.0   Max.   :202.0   Max.   :1.0000
```

Check for correlation

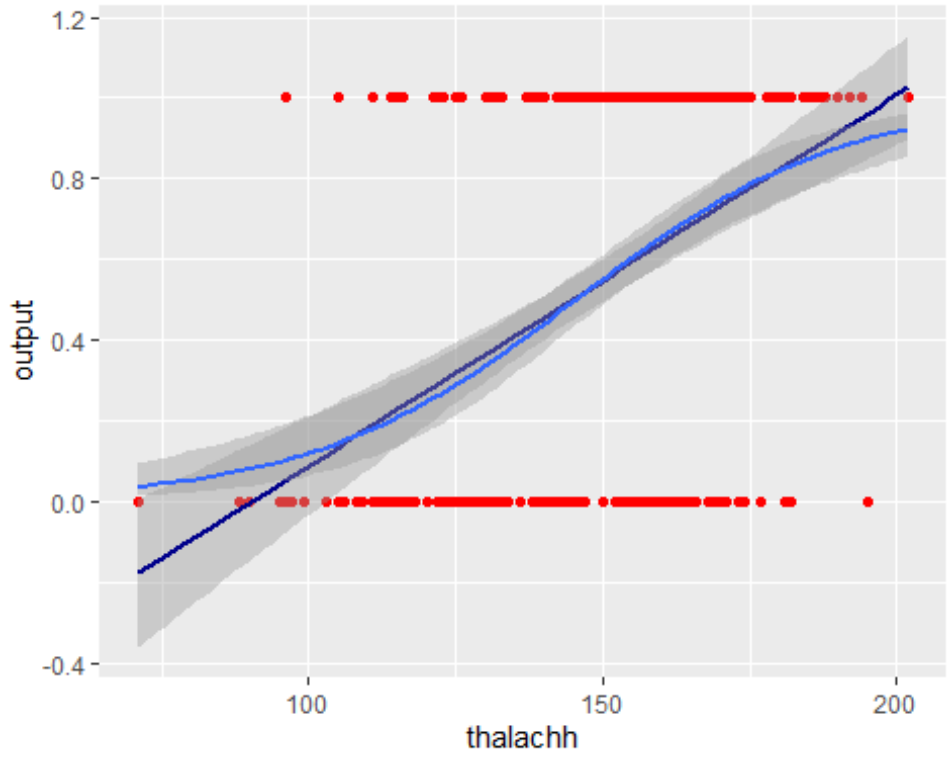
```
Data_new %>% ggplot(aes(chol, output))+
  geom_point(color = "blue")+
  geom_smooth(method = "lm",color = "orange")+
  stat_smooth(method = "glm",method.args = list(family= "binomial"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



```
Data_new %>% ggplot(aes(thalachh, output))+
  geom_point(color = "red")+
  geom_smooth(method = "lm",color = "darkblue")+
  stat_smooth(method = "glm",method.args = list(family= "binomial"))

## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



The first graph shows a negative correlation between chol and output. It means that if the cholestoral increase, there will be a lower chance of heart attack.

The second graph indicates a positive relationship between thalachh and output. It can be observable that the maximum heart rate achieved, the higher chance of heart attack happened.

Generated Linear Model

```
# Model 1
glm1 <- glm(output~ chol + thalachh, Data_new, family = binomial)
summary(glm1)

##
## Call:
## glm(formula = output ~ chol + thalachh, family = binomial, data =
Data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0894  -1.0199   0.5823   0.9523   2.0313
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.506770   1.127885  -4.882 1.05e-06 ***
## chol        -0.003805   0.002425  -1.569    0.117
## thalachh     0.044342   0.006588   6.731 1.69e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 356.76  on 300  degrees of freedom
## AIC: 362.76
##
## Number of Fisher Scoring iterations: 3
```

Coefficients

Intercept (-5.506770): The intercept in this model represents the log-odds of the outcome being positive when both cholesterol and maximum heart rate are zero. The value is highly significant (p-value = 1.05e-06), suggesting a strong baseline effect against the positive outcome under theoretical conditions where chol and thalach are at their minimum.

Cholesterol (Chol) (-0.003805): The coefficient for cholesterol indicates that an increase in cholesterol levels decreases the log-odds of the positive outcome by -0.003805, although this effect is not statistically significant (p-value = 0.117). This suggests that cholesterol may not have a strong independent impact on the outcome in the presence of heart rate as another predictor.

Maximum Heart Rate (Thalach) (0.044342): Each unit increase in maximum heart rate significantly increases the log-odds of the positive outcome by 0.044342, with a very high level of statistical significance (p-value = 1.69e-11). This indicates that thalach is a powerful predictor of the outcome, providing clear evidence that higher heart rate is associated with an increased likelihood of the positive outcome.

Residual Deviance (356.76 on 300 degrees of freedom): The substantial reduction in deviance suggests that the model, including both cholesterol and heart rate, explains a significant portion of the variability in the outcome. This improvement in fit compared to the null model, which includes no predictors, is notable.

AIC (362.76): The Akaike Information Criterion score of 362.76 balances the model's accuracy against its complexity. While this AIC is lower than some previous single-predictor models, indicating a more effective model in terms of explanatory power and simplicity.

```r
# Model 2
glm2 <- glm(output ~ chol, Data_new, family = binomial)
summary(glm2)
```

```
##
## Call:
## glm(formula = output ~ chol, family = binomial, data = Data_new)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.425  -1.241   1.015   1.093   1.567
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.001617   0.571467   1.753   0.0797 .
## chol        -0.003338   0.002269  -1.471   0.1412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 415.43  on 301  degrees of freedom
## AIC: 419.43
##
## Number of Fisher Scoring iterations: 4
```

Coefficient

Intercept (1.001617): The intercept value indicates the log-odds of the outcome being positive when cholesterol level is at zero. This coefficient is not statistically significant (p-value = 0.097), suggesting that the baseline effect in the model, without considering cholesterol, does not significantly influence the outcome.

Cholesterol (Chol) (-0.003338): Each unit increase in cholesterol level decreases the log-odds of the outcome by -0.003338. Although the effect suggests a negative association between cholesterol levels and the outcome, this coefficient is not statistically significant (p-value = 0.1412), indicating that cholesterol levels might not have a strong impact on the outcome in this model setting.

Residual Deviance (415.43 on 301 degrees of freedom): The model reduces the residual deviance slightly from the null deviance of 417.64 on 302 degrees of freedom, indicating a minimal improvement in model fit upon including cholesterol as a predictor.

AIC (419.43): The Akaike Information Criterion score of 419.43 reflects the model's balance between accuracy and complexity. The slight decrease compared to the null model suggests a minimal contribution of cholesterol to improving the model's explanatory power.

```r
# Model 3
glm3 <- glm(output~thalachh, Data_new, family = binomial)
summary(glm3)
```

```
##
## Call:
## glm(formula = output ~ thalachh, family = binomial, data = Data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1383  -1.0780   0.6043   0.9200   2.1354
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.391452   0.987133  -6.475 9.50e-11 ***
## thalachh     0.043951   0.006531   6.729 1.71e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 359.26  on 301  degrees of freedom
## AIC: 363.26
##
## Number of Fisher Scoring iterations: 4
```

Coefficient

Intercept (-6.391452): The intercept value indicates the log-odds of the outcome being positive when the maximum heart rate (thalach) is zero. This coefficient is highly significant (p-value = 9.50e-11), suggesting a strong baseline effect in the model. The negative value indicates a low likelihood of the outcome being positive at this baseline level of heart rate, which is theoretically very low (near zero).

Thalach (0.043951): Each unit increase in maximum heart rate significantly increases the log-odds of the outcome by 0.043951. This effect is highly significant (p-value = 1.71e-11), highlighting a strong positive association between heart rate and the likelihood of the outcome. The odds ratio, exp(0.043951), suggests that each unit increase in heart rate significantly increases the probability of the positive outcome.

Residual Deviance (359.26 on 301 degrees of freedom): The decrease in residual deviance from the null deviance of 417.64 on 302 degrees of freedom to 359.26 on 301 degrees of freedom indicates that the model fits the data significantly better than the null model, which only includes the intercept.

AIC (363.26): The Akaike Information Criterion score of 363.26 reflects the model's balance between accuracy and complexity. This score demonstrates the model's effectiveness in

using maximum heart rate to describe the outcome while considering the complexity of the dataset, which might influence comparisons with other models.

Do the logistic regression model

```
lrm(output~ chol + thalachh, Data_new)
```

```
## Logistic Regression Model
##
## lrm(formula = output ~ chol + thalachh, data = Data_new)
##
##                       Model Likelihood      Discrimination    Rank
Discrim.
##                             Ratio Test             Indexes
Indexes
## Obs           303    LR chi2      60.88      R2       0.243    C
0.750
##  0            138    d.f.             2      R2(2,303)0.177    Dxy
0.500
##  1            165    Pr(> chi2) <0.0001    R2(2,225.4)0.230    gamma
0.500
## max |deriv| 2e-06                           Brier    0.202    tau-a
0.249
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept -5.5068 1.1279 -4.88  <0.0001
## chol      -0.0038 0.0024 -1.57  0.1166
## thalachh   0.0443 0.0066  6.73  <0.0001
```

Remove maximum heart rate

```
lrm(output ~ chol, Data_new)
```

```
## Logistic Regression Model
##
## lrm(formula = output ~ chol, data = Data_new)
##
##                       Model Likelihood      Discrimination    Rank
Discrim.
##                             Ratio Test             Indexes
Indexes
## Obs           303    LR chi2      2.21      R2       0.010    C
0.570
##  0            138    d.f.            1      R2(1,303)0.004    Dxy
0.140
##  1            165    Pr(> chi2) 0.1373    R2(1,225.4)0.005    gamma
0.141
## max |deriv| 2e-10                           Brier    0.246    tau-a
0.070
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
```

```
## Intercept  1.0016 0.5715  1.75  0.0797
## chol      -0.0033 0.0023 -1.47  0.1412
```

Remove cholestoral

```
lrm(output ~ thalachh, Data_new)

## Logistic Regression Model
##
## lrm(formula = output ~ thalachh, data = Data_new)
##
##                       Model Likelihood      Discrimination    Rank
Discrim.
##                             Ratio Test             Indexes
Indexes
## Obs           303    LR chi2      58.38      R2       0.234    C
0.748
##  0            138    d.f.             1      R2(1,303)0.173    Dxy
0.497
##  1            165    Pr(> chi2) <0.0001    R2(1,225.4)0.225    gamma
0.502
## max |deriv| 5e-08                           Brier    0.203    tau-a
0.247
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept -6.3915 0.9871 -6.47  <0.0001
## thalachh   0.0440 0.0065  6.73  <0.0001
```

We can observe the performance differences among the three models analyzed. Each model showcases unique strengths, with their C and Dxy values indicating various levels of predictive accuracy.

Firstly, Model 1, which includes predictors cholesterol and thalach, exhibits the best overall fit among them. Its C-stat of 0.75 and Dxy of 0.50 demonstrate solid discriminative power, establishing it as the most effective model in differentiating between the positive and negative outcomes of the dataset.

Secondly, Model 2, relying solely on cholesterol as a predictor, shows considerably weaker performance. The C-stat of 0.570 and a Dxy of 0.140 suggest limited ability to discriminate effectively between outcomes, highlighting the insufficiency of using cholesterol alone as a reliable predictor.

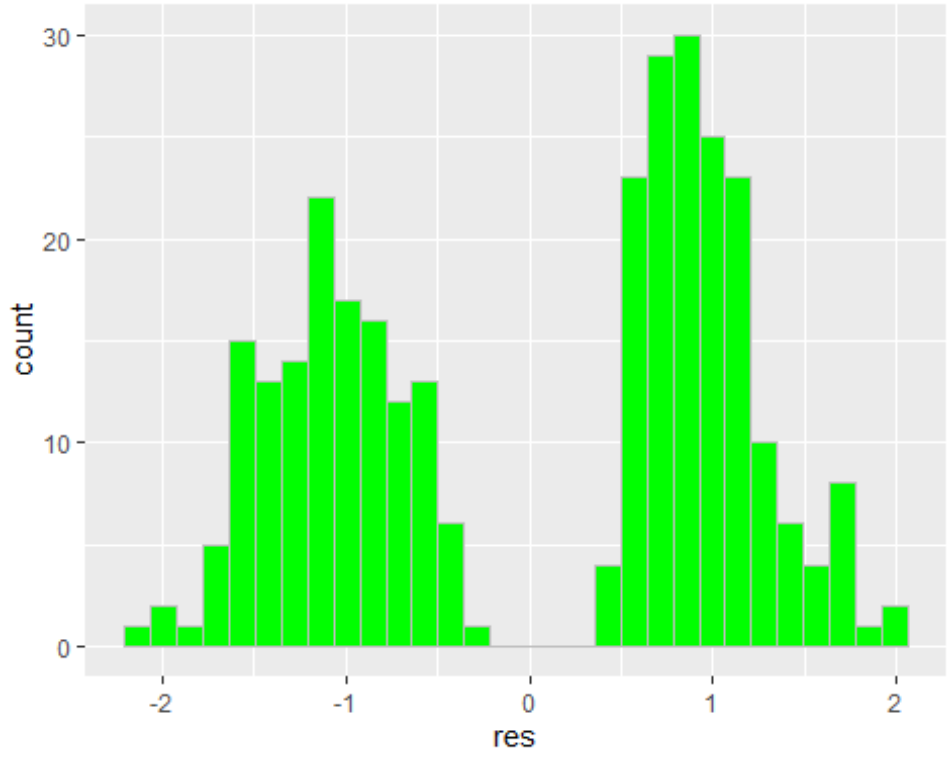
Lastly, Model 3, which uses thalach as the sole predictor, performs commendably for a model based on a single variable. Its C-stat of 0.748 and Dxy of 0.497 are indicative of good predictive accuracy and discriminative ability, making it a robust choice for simpler predictive needs.

Thus, Model 1 demonstrates remarkable discriminative power, establishing it as the most effective model.

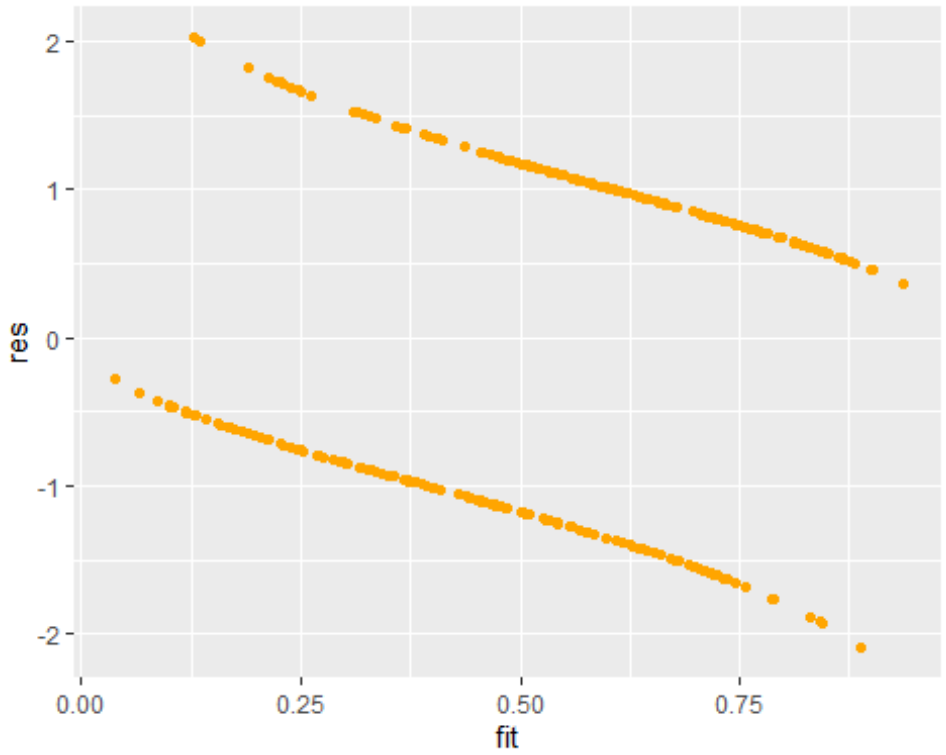
Residue's Distribution

```
Data_new1 <- Data_new %>% mutate(res=resid(glm1),fit = fitted(glm1))
Data_new1 %>%
  ggplot(aes(res))+
  geom_histogram(fill = "green", color = "grey")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
Data_new1 %>%
  ggplot(aes(fit, res))+
  geom_point(color = "orange")
```



The histogram indicates that the distribution is not normal, which is expected for residuals from a logistic regression model given the binary nature of the outcome. Moreover, according to the plot, as we can see, more than 95% of the residuals are between -2 and 2. This suggests that the model provides a good fit for certain sections of the data. The residuals versus fitted values plot shows a clear funnel shape, with residuals spreading out as the fitted values approach 0.25 and 0.75. This pattern is typical in logistic regression models due to the nature of the binomial distribution and indicates heteroscedasticity.

Do the Hosmer–Lemeshow test

```
hoslem.test(Data_new1$output, Data_new1$fit, g = 10)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  Data_new1$output, Data_new1$fit
## X-squared = 9.2343, df = 8, p-value = 0.3229
```

Chi-Squared (X-squared): The test statistic is reported at 9.2343. This value represents the sum of squared deviations between the observed and expected outcomes across deciles of predicted probabilities, based on the model's estimations.

Degrees of Freedom (df): The test utilizes 8 degrees of freedom. This count generally arises from dividing the dataset into ten groups (deciles), with the degrees of freedom calculated as the number of groups minus two.

P-value: The recorded p-value is 0.3229, which is significantly above the commonly accepted significance threshold of 0.05.

The high p-value (0.3229) from the Hosmer and Lemeshow test indicates a good fit for the model, as it suggests there is no significant deviation between observed and predicted outcomes across the various deciles of predicted probabilities. This result implies that the model is effectively capturing the actual trends within the data.

Make Prediction

```
make_predict <- data.frame(
  chol = c(170,283),
  thalachh = c(180,85)
)

pre <- predict(glm1, newdata = make_predict, interval = "prediction", level = 0.95)
print(pre)

##         1         2
##  1.828049 -2.814387
```

The two predictions yield one positive, which means “closed” and one negative, which is “not closed”, result. This outcome aligns with expectations, as the first prediction involves a lower cholestoral rate and a higher maximum heart rate achieved, while the second involves a higher cholestoral rate and lower maximum heart rate achieved.

Conclusion

This project applied logistic regression to predict heart attack occurrences based on cholesterol levels and maximum heart rate and the result shows some good sights. The most comprehensive model, combining both predictors, demonstrated the strongest fit and predictive power, as evidenced by significant p-values and a robust Hosmer and Lemeshow test result, indicating a good model fit.

The analysis confirmed the model’s effectiveness in practical scenarios, such as predicting heart attack risk, which can aid in clinical decision-making and preventative healthcare strategies. Overall, this project highlights the value of logistic regression in health sciences, offering a solid foundation for further research and application in medical settings to enhance patient care and preventative measures.