**Instructions: Name:** Craig Shaffer

You may work with a partner on this project which covers material from chapters 6, 7, and 8. You will build both a descriptive and a predictive logistic regression model. You will need to install the *ISLR* package in RStudio. You are going to use a dataset from this package called *Hitters*. Answer the questions below. You will submit your answers to Bb, along with ALL R code and ALL output.

Write in complete sentences and with correct grammar. Pretend you are explaining what you are doing, step by step, to someone who is unfamiliar with statistical analysis. Therefore, you will need make sure that you are very clear about what you are doing and why.

**Data Set:**

The data set ***Hitters*** from ISLR package contains Major League Baseball Data from the 1986 and 1987 seasons. *Hitters* is a data frame with 322 observations of major league players for 20 different variables. See the help file for more information.

***GOAL: We wish to predict whether a baseball player made an above average salary in 1987 (in thousands of dollars) on the basis of various statistics associated with performance.***

First, we need to get the data set ready. The *Salary* column has many missing (NA) values that need to be removed, and then we need to create a new column indicating whether the player earned an above average salary (more than $536,000). Use the following R code to do this. Note that the code creates a new data set called ***Hitters2***. Also note that the code will display the first few rows of the data set, produce a histogram displaying the distribution of salaries, and compute summary statistics for salary. This is for your reference only.

library(ISLR)

head(Hitters)

Hitters2<-Hitters[!is.na(Hitters$Salary),]

hist(Hitters2$Salary)

summary(Hitters2$Salary)

Hitters2$AboveAvgSalary="Yes"

for(i in 1:nrow(Hitters2)){

if(Hitters2$Salary[i]<=536){

Hitters2$AboveAvgSalary[i]="No"

}

}

Hitters2$AboveAvgSalary<-as.factor(Hitters2$AboveAvgSalary)

head(Hitters2)

**Part A: Descriptive Model**

You’ll begin by creating a descriptive model that predicts whether a baseball player made an above average Salary in 1987 (in thousands of dollars) on the basis of various statistics associated with performance *in the previous year (1986) and their current league*. Specifically, we will consider the following predictors:

* AtBat - Number of times at bat in 1986
* Hits - Number of hits in 1986
* HmRun - Number of home runs in 1986
* Runs - Number of runs in 1986
* RBI - Number of runs batted in in 1986
* Walks - Number of walks in 1986
* Assists - Number of assists in 1986
* Errors - Number of errors in 1986
* NewLeague - A factor with levels A and N indicating player's league at the beginning of 1987

Answer the following 5 questions.

1. Using *forward selection*, find the “best” logistic regression model to predict whether a player’s salary is above average. Use all quantitative variables for now, i.e. do not include NewLeague. Include all two-way interaction terms. You may choose the variable that you wish to start with in the preliminary model.

Text

Description automatically generated

A picture containing text, receipt

Description automatically generated A close-up of a document

Description automatically generated with medium confidence Text

Description automatically generated Text

Description automatically generated with medium confidence

1. What is your model estimating/predicting exactly?

The model is predicting/estimating the probability of AboveAvgSalary being Yes.

You can figure this out by running visualize\_model on the preliminary model and looking at the plot. Or since No is first alphabetically, it is 0 while Yes is 1. We predict probabilities for 1.

 Chart

Description automatically generated

1. What was your preliminary model? Why did you choose this model?

My preliminary model used Hits to predict AboveAvgSalary. I chose this model because being able to hit the ball seems like an important skill for professional baseball players. Also looking at the data (by using view(Hitters2)), many players with high numbers of hits seemed to have high salaries. Ultimately, I thought it would be a good predictor for having an AboveAvgSalary.

1. What predictors/terms are in the final model?

The predictors in the final model were Hits, Walks, Runs, Assists, and Hits:Walks

1. Explain how R chose this model. In other words, walk me through the forward selection process for this data.

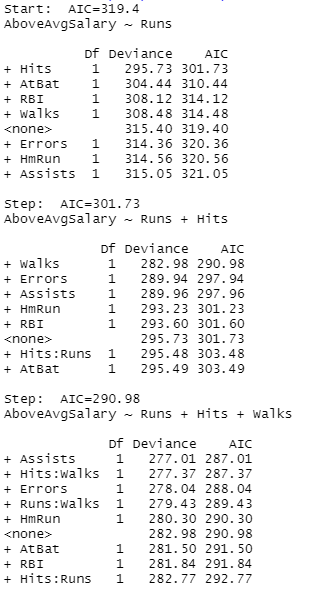
Refer to the output of forward selection:

First the model fits the preliminary model with just Hits and got an AIC of 301.54. It then checks if adding the other variables will lower the AIC. Walks lowered it by the largest amount (AIC down to 297.77), so it gets added to the model. Then it proceeds to try to add another variable to the model, the variable that lowered the AIC by the most was Runs (AIC down to 290.98), so it gets added to the model. For the next run through, it sees that Assists drops the AIC to 287.01 so it adds it to the model. The next iteration finds that the interaction Hits:Walks lowers the AIC to 283.55 so it gets added to the model. The next iteration finds that no variables lower the AIC, so it stops trying to add variables and returns the best model into result.

1. Select a different preliminary model and run forward selection again. Did you get the same result as in #1? If not, what’s different? Which model has the lower AIC?

Text

Description automatically generated

 Table

Description automatically generated

Table

Description automatically generated with low confidence

Yes, I got the same result as #1. The model it decided on has the predictors Runs, Hits, Walks, Assists, and Hits:Walks. It has the same AIC with 283.55.

1. Looking at the models from #1 and #2, select the one with the lowest AIC. Add the categorical predictor *NewLeague* to it.



Table

Description automatically generated

1. In order to add this predictor to the model, what dummy variable would you create?

You would create the dummy variable NewLeagueN because NewLeague ‘A’ would come first alphabetically and become the reference level. Therefore, NewLeagueN would become the dummy variable.

1. Provide an interpretation for its coefficient.

[Just comment on what the sign of the coefficient tells you.]

The coefficient for NewLeagueN is positive, therefore people in the N (National) league will be more likely to have AboveAvgSalary = ‘Yes’ than players in NewLeague ‘A’.

1. Now compute and interpret the odds ratio for two identical players, except for the fact that they play in different leagues.

=

For two identical players (outside of the fact they play in different leagues), the player that plays in the N league is 1.81 times more likely to have an above average salary than a player that plays in the A league. (Odds of AboveAvgSalary = ‘Yes’ are 81% higher for a player in the N league).

1. Did adding this variable further lower the AIC?

Yes, it lowered it from 283.55 to 282.01. However, it’s not that much of a difference.

1. Use the model from #3.
2. Write out the fitted logistic regression model.

Y = -2.0867896 -0.0554058 Runs + 0.0295681 Hits -0.0266012 Walks -0.0027333 Assists

+ 0.5912650 NewLeagueN + 0.0005232 Hits:Walks

1. Are there any interaction terms? If so, then describe the interaction(s).

Yes, the interaction Hits:Walks is included in the model. The slope changes (increases) by 0.0005232 for each one increase in Walks relationship between AboveAvgSalary and Hits.

1. Then provide interpretations for two additional coefficients (your choice).

[Just comment on what the sign of the coefficient tells you.]

Coefficient

Hits 0.0295681

Walks -0.0266012

Since the coefficient for hits is positive, people with larger values for hits will have a higher probability of AboveAvgSalary = ‘Yes’. Since the coefficient for walks is negative, people with larger values for hits will have a lower probability of AboveAvgSalary = ‘Yes’.

1. Compute and interpret at least 2 odds ratios (your choice).

[For example, you could compare two players, call them A and B, who are identical except for their number of homeruns. Let’s say that player A had 10 more homeruns than player B.

Player 1 has 50 more hits than player 2 (everything else identical):

=

For two identical players (Except for player 1 has 50 more hits than player 2), Player 1 is 4.39 times more likely to have an above average salary than a Player 2. (Odds of AboveAvgSalary = ‘Yes’ are 339% higher for Player 1).

Player 1 has 50 more walks than player 2 (everything else identical):

= 

= 

For two identical players (Except for player 1 has 50 more walks than player 2), Player 2 is 3.78 times more likely to have an above average salary than a Player 1. (Odds of AboveAvgSalary = ‘Yes’ are 278% higher for Player 2).

1. Are all predictors significant? Explain.

No, Walks is not significant (p-value of 0.30507 < 0.05). However, we cannot remove it from the model because the interaction Hits:Walks is included in the model and significant (p-value of 0.01727).

1. Comment on model utility. Start by creating a confusion matrix for the model that you found in #3.



Text

Description automatically generated with medium confidence

1. What is the misclassification rate? Is this rate low enough to be useful? Explain. Also talk about the number of false negatives and false positives.

= 

The misclassification rate is 24.3%. I would say that this is low enough to be useful because it is right more than 3/4ths of the time. That means the model correctly classifies 75.7% of cases. I think that is pretty good considering how contracts may vary based on team situations and young players who haven’t proved their abilities enough to get a major contract.

The model had 40 false negatives which means that the model predicted 40 people to not have an above average salary when they actually did.

The model had 24 false positives which means that the model predicted 24 people to have an above average salary who did not.

Niether misclassification is particularly harmful in this scenario because we are just predicting whether a player would have an above average salary. Had we been predicting diseases, then it would be a bigger deal to try and have a low false negative rate because we would rather be over cautious than being under cautious and leave a patient untreated.

1. How does the misclassification rate compare to that of the naïve model?

= 

The misclassification rate of the model is much better than the naïve. The logistic regression model is good because it decreased the misclassification rate from 39.9% down to 24.3%.

1. Is there a simpler model that you would choose to use instead? Explain why or why not. If there is a simpler model, make sure to state the model here.

I would consider using the model from Question 1 and 2 (the same model as this without *NewLeague* in it).

The model would predict AboveAvgSalary from the predictors Hits, Walks, Runs, Assists, and Hits:Walks.

I would consider this model because it had one less variable and the AIC was within 2 (283.55 compared to 282.01).

**Part B: Predictive Model**

The data set also contains the following variables:

* CAtBat - Number of times at bat during career
* CHits - Number of hits during career
* CHmRun - Number of home runs during career
* CRuns - Number of runs during career
* CRBI - Number of runs batted during career
* CWalks - Number of walks during career

Now you’ll create a predictive model that predicts whether a baseball player made an above average Salary in 1987 (in thousands of dollars) using the statistics associated with performance in the previous year (1986), their current league, plus these new statistics on the player’s performance over the course of their career. Use the following code to create a training data set and a holdout data set (containing 50 players chosen at random).

set.seed(12345)

ii<-sample(1:263, 50)

TRAIN<-Hitters2[-ii,]

HOLDOUT<-Hitters2[ii,]

Answer the following 2 questions.

1. Before you begin to create a predictive logistic regression model, explain the following:
2. What is the difference between a descriptive and a predictive model?

The goal of a descriptive model is to be able to reflect reality as closely as possible. It seeks to determine the “right” variables to predict some y and to estimate the coefficients as precisely as possible. For predictive models, we don’t care about how well our model fits our data because we already have the answers for these individuals. Instead, we care about how well it predicts on new data (or data it hasn’t seen yet). We measure the success of these models on their ability to generalize to new data. Instead of a low AIC, we care about having a small generalization error.

1. How does K-fold cross-validation work, in general, and what is its purpose in predictive modeling?

K-fold cross-validation works by splitting the data into K groups/subsets. In each of K different rounds, one group is reserved as the validation set while the model is fit on the rest. The value of the error (for example, RMSEvalidation) is then calculated for the validation set. Each of the K subsets gets to serve as the validation set once. The estimated generalization error for the model is the average of these validation errors.

The purpose of k-fold c cross-validation is to minimize the sensitivity of the model to one set of the data. It alleviates some of the bias in the training data.

1. How are the generalization errors for multiple regression and logistic regression different?

The error for multiple regression is quite simple as it’s just how far off you typically are for predictions (RMSE, etc.). In comparison, logistic regression involves predicting whether something is or isn’t (1 or 0). So, your generalization error will be rather small since it’s the error of a probability for 1. This gives you a multitude of decisions to make. You could consider misclassification rates, false positive rates, false negative rates, etc. for how you will judge your models. For example, if you were to predict fatal diseases, you would want a model with a low rate of false negatives.

1. What does it mean for a model to be “overfit”? Why is this a danger in predictive modeling?

A model is overfit when it is very biased toward the training data and has memorized quirks in the data that might not exist in a general setting. Essentially, it is very good at predicting for the training data, but it is inaccurate on new data. This is a danger in predictive modeling because the whole point of predictive modeling is to be good at predicting on new data/observations rather than the data it was built on.

1. Now, using all 15 predictors specified in this document, build a predictive model. Do NOT include interaction terms. You may use the default all possible models search procedure or the backwards selection method. Then answer the following questions:



Chart

Description automatically generated

Text

Description automatically generated

A picture containing text, receipt

Description automatically generated

1. Which model had the lowest generalization error? Which model was chosen as “best” based on the one-standard deviation rule?

The model with the lowest generalization error had seven predictors: Hits, Runs, Walks, Assists, NewLeagueN, CRuns, and CWalks. Its estimated generalization error is 0.1437038.

The “best” model based on the one-standard deviation rule had two predictors: Hits and CHits. Its estimated generalization error is 0.1502927.

1. Which model would you choose? Explain. Then provide the fitted logistic regression model for your choice.

I’d choose the model with two predictors that performs within 1 standard deviation of the seven-predictor model that had the lowest generalization error. I’d choose this model because performs similar to the best model and is a lot less complex (2 predictors vs 7 predictors).



A screenshot of a computer

Description automatically generated with medium confidence

Y= 0.0245552 Hits + 0.0017448 CHits

1. Using the model from b), what is the generalization error on the holdout data set? Use this to explain whether your model “generalizes well.”



A screenshot of a computer

Description automatically generated with medium confidence

= 

= -> 1-1.19766=.19766=19.8% increase

The generalization error on the holdout is the misclassification rate of 18% (.18). The estimated generalization error was ~15% misclassification rate (0.1502927). If you take the estimated generalized error times 1.2, you’ll get the value of 0.1803512. This is greater than the misclassification rate of the model on the holdout data. Therefore, the model has less than a 20% increase compared to the estimated generalization data. Since this is true, then our model generalizes well because the increase was less than 20% (19.8% increase).