```
---
title: "Multiple Logistic Regression Practice"
author: [Your name(s) here]
output: html_document
---
```

A. Research Question

Students at a small liberal arts college took a placement exam prior to entry in order to provide them guidance when selecting their first math course. The dataset **MathPlacement** (in Stat2Data) contains the placement scores for 2696 students, along with whether they took the recommended course, and several admissions variables (GPA, SAT score, etc.). We want to use this data to decide how well the math placement process is working. If they take the recommended course, do they succeed (where "success" is defined as a grade of "B" or above)?

B. Exploratory Data Analysis -- Simple Logistic Regression
---
1. Load the `mosaic` and `Stat2Data` packages. Load the data set and look in the manual (or Help menu) to see what variables are contained within it.

2. EDA of Response Variable.
Investigate the response variable, `CourseSuccess`. What percentage of students got a "B" or better? How many students are missing this variable?

3. EDA of First Explanatory Variable.
First we will try to predict course success using `RecTaken`. What type of variable is `RecTaken` (binary, categorical, numerical)? What percentage of students took the recommended course? How many students are missing this variable?

C. Analysis of the relationship -- Simple Logistic Regression
---
1. Make a two-way table of `CourseSuccess` and `RecTaken`. Of those who did take the recommendation, what percentage were successful? Of those who didn't take the recommendation, what percentage were successful?

2. Fit the logistic regression to predict `CourseSuccess` from `RecTaken`, and call this `model1`.

3. Write out the fitted model.

4. Interpret the slope coefficient (in terms of an odds ratio), in the context of this situation.

5. Use the fitted logistic model to predict the *probability* of success for a student who took the recommended course, and the *probability* of success for a student who didn't take the recommended course. What do you notice about these values?

D. Inference -- Simple Logistic Regression

For significance tests, be sure to state the hypotheses, give the values of the test statistic and the p-value, and state your conclusion in context.

1. Checking conditions

Are the conditions of linearity, randomness, and independence met in this situation? Make

sure you discuss each condition.

2. Use the code below to compute a 95% confidence interval for your slope and use it to find a confidence interval for the odds ratio. Does your interval include the value 1? Why does that matter?

```{r}
confint(model1) #CI for the slope
```

3. Test the claim that the slope is 0.

4. Use the G-test to test the overall effectiveness of the model.

E. Exploratory Data Analysis -- Other Potential Predictors

Other variables that may be useful in predicting course success are gender, ACT math score, and GPA. (You may feel there are other possibilities, but let's focus on these three.)

1. Calculate summary statistics for the 3 potential predictor variables.

2. Use *appropriate* graphs and/or tables to investigate the relationship between each potential predictor and the response variable. Write a sentence for each predictor, summarizing what you see.

3. Checking conditions - Linearity

We need to check that the relationship between the log odds (logits) and each predictor is approximately linear. (We've already discussed randomness and independence above.)

a. ACT math score

We could use the "grouping" technique discussed in Chapter 9, and used in the "Logistic Regression Practice" activity. Recall that in this technique, we create similarly-sized groups of ACT scores and plot the mean score of each group against the log odds of success in each group. A more "quick-and-dirty" method is to use code similar to that used at the end of Part B in the "Logistic Regression Practice" activity to calculate log odds for each ACT score, and plot this against ACT. (Make sure you understand what *every line* of code below is doing!)

```{r fig.height=4, fig.width=4}
tab.ACT <- xtabs(~ACTM+CourseSuccess,data=MathPlacement)
prop.ACT <- tab.ACT[,2]/(tab.ACT[,2] + tab.ACT[,1])
plot(log(prop.ACT/(1-prop.ACT))~sort(unique(MathPlacement$ACTM)),xlab="ACT
math",ylab="log(odds) of success")
```

Based on the plot above, do you think that linearity of the logits is a reasonable assumption for this variable?

b. adjusted GPA

Follow the method in part (a) to plot the logits against adjusted GPA. Based on that plot, do you think that linearity of the logits is a reasonable assumption for this variable?

Notice that there is one group of data points that needs to be deleted from the data set because they don't make any sense. **Delete those points now.**

c. Gender

Is linearity a reasonable assumption for the gender variable?

F. Multiple Logistic Regression -- 2 variables

Model 2: For the math professors, GPA is the easiest information to get, so let's start with a model that adds GPA to the existing model with `RecTaken`.

1. Fit the logistic regression to predict course success from `RecTaken` and GPA and call this `model2`. (Make sure you use the version of GPA that you created in #E3b, which has had the incorrect values removed.)

2. Write out the fitted model.

3. Comment on the effectiveness of each predictor in the model as well as the overall fit. Be sure to indicate what value(s) from the output lead to your conclusions.

4. Find and interpret the slope coefficient (in terms of an odds ratio) of GPA, in the context of this situation.

5. Find the confidence interval for slope coefficient of `RecTaken`. Interpret this CI in terms of odds ratios, in the context of this situation.

6. Use the fitted logistic model to predict the *probability* of success for a student who took the recommended course and had a GPA of 3.0, and the *probability* of success for a student who didn't take the recommended course and had a GPA of 3.0. Then use the model to predict the *probability* of success for a student who took the recommended course and had a GPA of 3.9, and the *probability* of success for a student who didn't take the recommended course and had a GPA of 3.9. Comment on what you see.

G. Multiple Logistic Regression -- 3 variables

Model 3:

1. Fit the logistic regression to predict course success from `RecTaken`, GPA, and gender and call this `model3`.

2. Write out the fitted model.

3. Comment on the effectiveness of each predictor in the model as well as the overall fit. Be sure to indicate what value(s) from the output lead to your conclusions.

H. Multiple Logistic Regression -- 4 variables

Model 4:

1. Fit the logistic regression to predict course success from `RecTaken`, GPA, gender, and ACT math score and call this `model4`.

2. Write out the fitted model.

3. Comment on the effectiveness of each predictor in the model as well as the overall

fit. Be sure to indicate what value(s) from the output lead to your conclusions.

I. Multiple Logistic Regression -- with Interaction

Model 5:

1. Add the following interactions to `model4`: ACTxGender and GPAxGender. Call this `model5`.

2. Comment on the effectiveness of the interaction terms based on their Wald test results.

3. Conduct a nested drop-in-deviance test to make a conclusion about whether the interaction terms are useful.

J. Comparison of models
---
1. Compare and contrast models 1 - 5.

2. Are there any additional changes or different models you'd like to investigate? For example: deleting an existing term, trying a different interaction, or a squared term? If so, fit that model below.

3. Which model (of all the ones you've fit) do you prefer? Explain why.

K. Prediction/Misclassification Table

Using *your* preferred model from J3...

1. The predicted probabilities for all of the data cases can be accessed through `fitted(model)`. Classify each data point as being a predicted "success" (1) if the predicted probability is greater than 0.5, and a predicted "failure" (0) if the predicted probability is less than 0.5.

2. Look at the classifications for each of the data points and create a 2 x 2 table showing proportions of how the data are classified (predicted success or predicted failure) versus their *actual* response values.

3. Comment on the accuracy of the classifications for your model. What percentage of cases were "misclassified" (i.e., predicted to be success when actually a failure or vice versa)?

4. Now pick one of the other, simpler models from this activity, and create its misclassification table. Compare it to the misclassification table for your preferred model that you created in #2 above. Does your model have a substantial improvement in misclassification rate, compared to the simpler model? Do you think the additional predictor(s) in your preferred model are worth it?