PM592: Regression Analysis for Data Science **HW9** *Prediction Models*	Name: Flemmin g Wu

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- For the purpose of this assignment, statistical evidence refers to a test statistic and associated p-value.
- If a question requires a conclusion, it must be phrased professionally and coherently.
- There are 2 questions and 30 points possible.

Question 1 [20 points]

J. Rumbagai at the Los Angeles County Wellness Department is trying to recruit residents for a focus group on attitudes toward marijuana use. Her department's primary goal is to develop a program that will reduce the amount of smoking/vaping that occurs at public parks, and the focus group is necessary in order to determine the content and delivery of the program. Residents are generally hesitant to participate since the focus group occurs during work hours, so they are offered $50 for participation. The department has access to data on residents from last year. Of the 673 who were invited to participate, 152 actually attended the focus group.

Your main goal will be to develop a model that can be used to predict whether a resident participated in the focus group last year. Because employees go door-to-door to invite residents, knowing who is likely to participate this year will help the department manage resources more efficiently.

The data is located in **focusgroup.dta** and the data dictionary follows:

Variable	**Description**
Participated	1=participated, 0=no participation
Income	Family income level, rounded to the nearest $1,000
Profession	3=unemployed, 2=retired, 1=professional, 0=otherwise
Isfemale	1=female, 0=otherwise
White	1=white race, 0=otherwise
residencelength	Length of residency in Los Angeles (years)
English	1=English is primary language in house, 0=otherwise

You started to create this model from the beginning, but J. Rumbagai then told you her assistant already assessed the variables in the model. The final model she found was:

- Income quartile (categorical)
- Profession (categorical)
- Sex
- Ethnicity

1a. [2 points] Split the model into a training and testing data set and run the final model (with the four listed variables) on the training data set. Report the model coefficients.

```
> fg$income.q <- cut(fg$income, quantile(fg$income), include.lowest = T)
> fg$profession.f <- factor(fg$profession, levels=c(1, 0, 2, 3))
>
> #skimr::skim(fg)
>
> set.seed(11)
> fg$train <- sample(c(F,T), nrow(fg), prob = c(.2, .8), replace = T)
>
> fg_train <- fg[fg$train, ]
> fg_test <- fg[!fg$train, ]
>
> m <- glm(participated ~ income.q + profession.f + isfemale + white, family = binomial, data = fg_train)
> summary(m)

Call:
glm(formula = participated ~ income.q + profession.f + isfemale +
    white, family = binomial, data = fg_train)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -5.9582     0.7502  -7.942 1.99e-15 ***
income.q(1.4e+04,3.2e+04]  0.8570     0.6609   1.297 0.194714
income.q(3.2e+04,5.2e+04]  1.6228     0.6421   2.527 0.011497 *
income.q(5.2e+04,7.5e+04]  4.9407     0.6323   7.814 5.54e-15 ***
profession.f0              1.2578     0.3629   3.466 0.000529 ***
profession.f2              6.0835     0.7145   8.515  < 2e-16 ***
profession.f3              6.9077     0.8886   7.773 7.64e-15 ***
isfemale                   0.5228     0.3094   1.690 0.091070 .
white                      0.6283     0.3635   1.729 0.083873 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 602.87  on 543  degrees of freedom
Residual deviance: 298.89  on 535  degrees of freedom
AIC: 316.89

Number of Fisher Scoring iterations: 6
```

The model coefficients are as follows: **0.86** the second income quartile (compared to lowest), **1.62** for the third income quartile (compared to lowest), **4.94** for the highest income quartile (compared to lowest), **1.26** for non-retired, non-professional, non-unemployed group (compared to professional), **6.08** for retired group (compared to professional), **6.91** for unemployed group (compared to professional), **0.522** for gender, and **0.63** for ethnicity.

The equation for the best fit line is:

$$\hat{Y} = -6 + 0.86X_{income.q(1.4e+04,3.2e+04]} + 1.62X_{income.q(3.2e+04,5.2e+04]} + 4.94X_{income.q(5.2e+04,7.5e+04]} + 1.26X_{otherwise} + 6X_{retired} + 6.9X_{unemployed} + 0.522X_{gender} + 0.63X_{ethnicity}$$

1b. [3 points] For this model, report the following:

- What is the goodness of fit of the model?

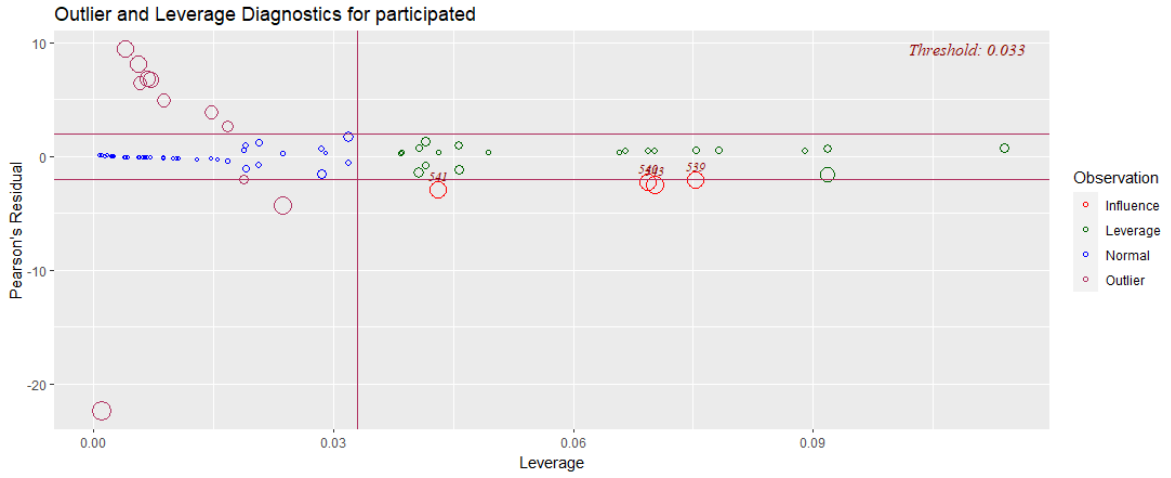
- What is the pseudo-R^2 for the model?
- Did any observations appear to be influential?

```
> hoslem.test(m$y, fitted(m))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  m$y, fitted(m)
X-squared = 21.793, df = 8, p-value = 0.005315
```

The Hosmer Lemeshow test returns a significant p-value ($\chi^2_8 = 21.8, p = 0.005$), indicating statistically significant departure from goodness of fit.

```
> DescTools::PseudoR2(m)
 McFadden
0.5042289
```

The pseudo R-squared value is 0.504

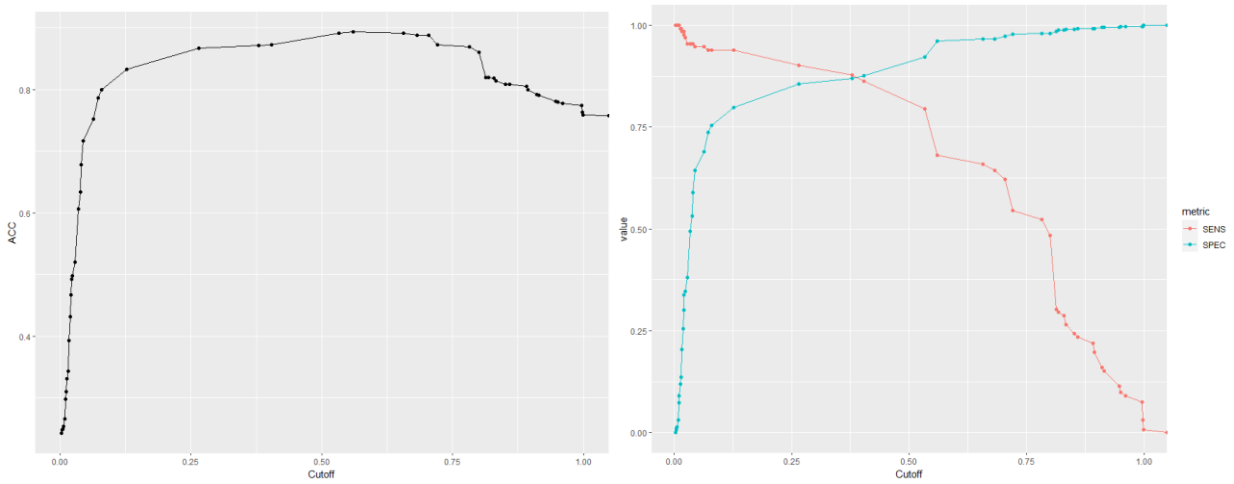


It appears that observations 541, 540, 543, and 539 are influential, however they are not very strong outliers. Overall, I would say that the model is pretty well-fit.

1c. [6 points] Assess the predictive ability of your model. Include:

- What cutoff did you use for classification and how did you arrive at that value?
- What is the predictive accuracy of your model?
- What are the values of sensitivity and specificity for the model?
- What is the value of the AUC for the model?

```
> m.p <-
+   tibble(
+     pred_p = m$fitted.values,
+     y = m$y
+   )
>
> roc <- ROCit::measureit(score = m$fitted.values,
+                         class = m$y,
+                         measure = c("ACC", "SENS", "SPEC"))
> # plot accuracy at different cutoff values
> tibble(
+   Cutoff = roc$Cutoff,
+   ACC = roc$ACC
+ ) %>%
+   ggplot(aes(x = Cutoff, y = ACC)) +
+   geom_point() +
+   geom_line() # optimal cutoff seems to be about .5
>
> # plot sensitvity and specificity tradeoff
> tibble(
+   Cutoff = roc$Cutoff,
+   SENS = roc$SENS,
+   SPEC = roc$SPEC
+ ) %>%
+   pivot_longer(., cols = c("SENS", "SPEC"), values_to = "value", names_to = "metric") %>%
+   ggplot(aes(x = Cutoff, y = value, color = metric)) +
+   geom_point() +
+   geom_line()
>
```



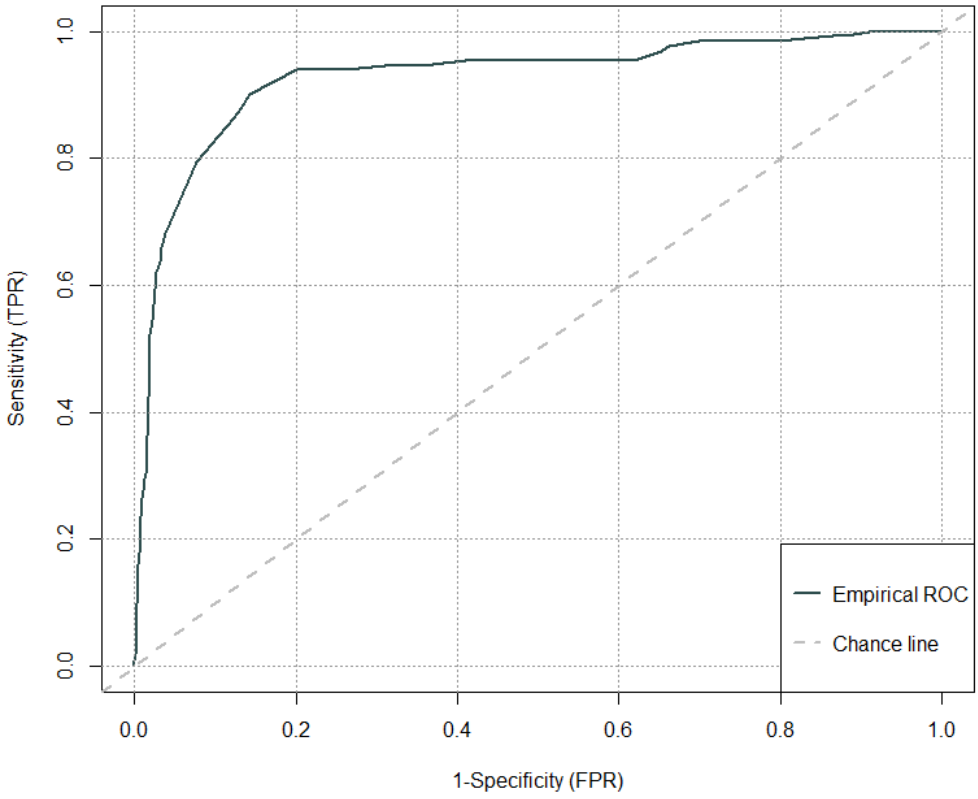
I plotted the accuracy at different cutoffs, and it seems to suggest that the highest accuracy is achieved when the cutoff value is about 0.55. I also plotted the tradeoff between sensitivity and specificity and that plot seems to suggest the optimal tradeoff is by using a cutoff value of about 0.39. I think that it would be okay to lower the cutoff value (i.e. slightly increase sensitivity and decrease specificity), this way more participants will be detected at the cost of inviting more residents that will not participate. For this reason, I will use 0.38 as the cutoff value.

```
> DescTools::Conf(m, cutoff = 0.38, pos=1)

Confusion Matrix and Statistics

          Reference
Prediction   1   0
         1 114  51
         0  18 361

               Total n : 544
              Accuracy : 0.8732
                95% CI : (0.8426, 0.8985)
    No Information Rate : 0.7574
    P-Value [Acc > NIR] : 1.13e-11

                 Kappa : 0.6819
 Mcnemar's Test P-Value : 1.17e-04

            Sensitivity : 0.8636
            Specificity : 0.8762
         Pos Pred Value : 0.6909
         Neg Pred Value : 0.9525
             Prevalence : 0.2426
         Detection Rate : 0.3033
   Detection Prevalence : 0.2096
      Balanced Accuracy : 0.8699
         F-val Accuracy : 0.7677
     Matthews Cor.-Coef : 0.6900

       'Positive' Class : 1
```

The accuracy for using a 0.38 cutoff is 0.8732 (95%CI = (0.8426, 0.8985)). Sensitivity is 0.8636 and specificity is 0.8762.

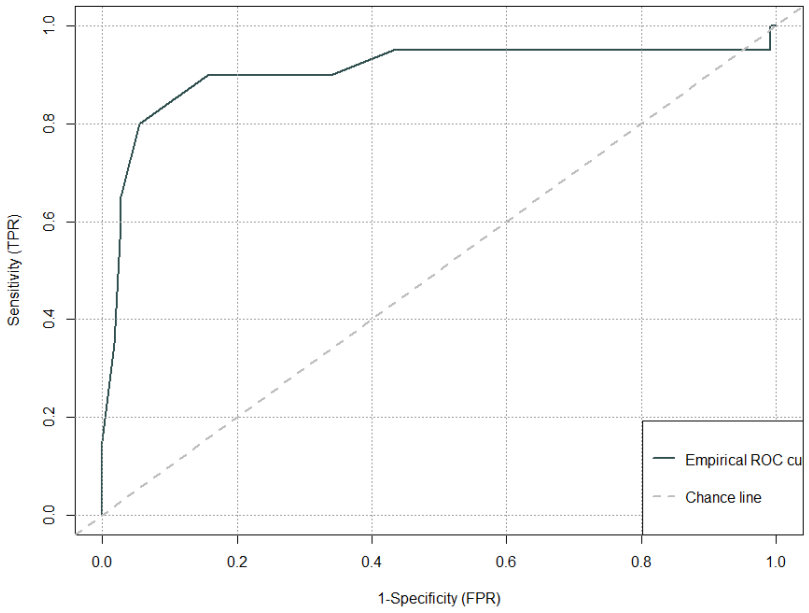
```
> roc_empirical <- rocit(score = m$fitted.values, class = m$y)
> plot(roc_empirical, YIndex = F)
> summary(roc_empirical)

 Method used: empirical
 Number of positive(s): 132
 Number of negative(s): 412
 Area under curve: 0.9271
> ciAUC(roc_empirical)

   estimated AUC : 0.927147690497205
   AUC estimation method : empirical

   CI of AUC
   confidence level = 95%
   lower = 0.895673788302731     upper = 0.958621592691679
```

The model's AUC is 0.93 (95% CI = (0.896, 0.959))



1d. [5 points] Apply the model's predictions to the testing data set. Report the AUC applied to the testing data set.

```
> m.test.p <-
+   tibble(
+     pred_p = predict(m, newdata = fg_test, type = "response"),
+     y = fg_test$participated
+   )
>
> test.roc <-
+   ROCit::measureit(score = predict(m, newdata = fg_test, type = "response"),
+                    class = fg_test$participated,
+                    measure = c("ACC", "SENS", "SPEC"))
> test_roc_empirical <-
+   rocit(score = predict(m, newdata = fg_test, type = "response"),
+         class = fg_test$participated)
> plot(test_roc_empirical, YIndex = F)
> summary(test_roc_empirical)

 Method used: empirical
 Number of positive(s): 20
 Number of negative(s): 109
 Area under curve: 0.9055
> ciAUC(test_roc_empirical)

   estimated AUC : 0.905504587155963
   AUC estimation method : empirical

   CI of AUC
   confidence level = 95%
   lower = 0.815302503054299     upper = 0.995706671257628
>
```



The model achieved an AUC of 0.906 on the testing set (95% CI = (0.815, 0.996))

1e. [4 points] Provide a professionally worded conclusion paragraph. This conclusion should state what your final model found, touch on how each of the variables is related to the likelihood of participation, and explain the equation for predicting participation this year.

A logistic regression model was trained to predict whether or not a resident participated in the focus group last year with the purpose of identifying people likely to participate in the current year. The variables selected as informative to making the predictions were income quartile, profession, sex, and ethnicity. Income and profession were both converted to categorical variables for model training. First, the data was split into training and testing sets with 80% used for training and the other 20% reserved for model validation. The model was fit to the training set and gave the following best fit equation for the line: $\hat{Y} = -6 + 0.86X_{income.q(1.4e+04,3.2e+04]} + 1.62X_{income.q(3.2e+04,5.2e+04]} + 4.94X_{income.q(5.2e+04,7.5e+04]} + 1.26X_{otherwise} + 6X_{retired} + 6.9X_{unemployed} + 0.522X_{gender} + 0.63X_{ethnicity}$. The coefficient for the second income quartile $(1.4\text{e}+04, 3.2\text{e}+04]$ indicates that compared to the lowest income quartile, the odds of someone in the second income quartile participating in the focus group last year increases by $e^{0.857} = 2.35$ $(p = 0.19)$ times, holding all other variables constant. The coefficient for the third income quartile $(3.2\text{e}+04, 5.2\text{e}+04]$ indicates that compared to the lowest income quartile, the odds of someone in the third income quartile participating in the focus group last year increases by $e^{1.62} = 5.06$ $(p = 0.01)$ times, holding all other variables constant. The coefficient for the highest income quartile $(5.2\text{e}+04,\ 7.5\text{e}+04]$ indicates that compared to the lowest income quartile, the odds of someone in the highest income quartile participating in the focus group last year increases by $e^{4.94} = 139$ $(p < 0.01)$ times, holding all other variables constant. The coefficient for the "otherwise" group (non-employed, non-professional, non-retired) indicates that compared to professionals, the odds of participating increase by $e^{1.26} = 3.52$ $(p < 0.01)$ times. The coefficient for the retired group indicates that compared to professionals, the odds of participating increases by $e^{6.08} = 438.6$ $(p < 0.01)$ times, holding all other variables constant. The coefficient for the non-employed group indicates that compared to professionals, the odds of participating increase by $e^{6.91} = 1000$ $(p < 0.01)$ times, holding all other variables constant. The coefficient for sex indicates that compared to non-females, the odds of females participating increases by $e^{0.52} = 1.69$ $(p = 0.09)$ times, holding all other variables constant. The coefficient for ethnicity indicates that compared to non-whites, the odds of whites participating increases by $e^{0.63} = 1.87$ $(p = 0.08)$, holding all other variables constant. The model was assessed for influential points and no points were found that were suspected to highly influence model parameters. A cutoff point for classification of 0.38 was chosen, and the model was then scored, achieving an AUC of 0.93 (95% CI = (0.896, 0.959)), accuracy of 0.87 (95%CI = (0.84, 0.90)), sensitivity of 0.86, and specificity of 0.88. Finally, the model was validated with the testing set, resulting in an AUC of 0.91, indicating that the model is not overfit to the training data.

Question 2 [10 points]

Read the article by Guo et al. (2019) available on Blackboard. The authors develop a model to predict mortality risk in patients with viral pneumonia.

2a. [2 points] How many individuals were in the final analytic sample? What percentage of these were assigned to the training vs. testing data set?

After exclusion criteria, 528 pneumonia patients with a positive viral detection were included in the analysis. 80% of the data (n = 423) was used for training and 20% (n = 105) was used for testing.

2b. [2 points] How did the authors define the outcome groups for the prediction model?

The outcome groups were divided between those who died within 90 days and those who did not (survival group).

2c. [2 points] How did the authors choose the cut-point for classifying outcome?

The authors assigned weights to the predictors in the model, giving them each a point number, and assigned a score based on these variables. They used Youden's index of ROC curve (max(specificity + sensitivity – 1)) to identify best cut-off points and used 12 in the final model.

2d. [2 points] What modeling approach did the authors use to select the variables for the final model?

The authors used a backward stepwise logistic regression analysis to select variables for the final model.

2e. [2 points] What did the authors do to convert the regression model into a tool that is more easily used by clinicians?

They developed a scoring system called the MuLBSTA score that can be calculated with easy-to-get clinical lab results for 6 variables, which simplifies the coefficients for the model. Plugging in values and obtaining the score for a patient makes it easy to determine high-risk patients from low-risk patients. An online calculator is available at: https://www.mdcalc.com/calc/10279/mulbsta-score-viral-pneumonia-mortality