# LAB 3: Classification Methods

Due date : November 16, 2023, 11:59 PM

To access assignments, go to file>make a copy.

The lab will be in your own google drive, and you can modify it.

Please answer all questions within the space provided. Please type the assignment and use R.

Your last name: Suri

Your first name: Jatin

Your student ID: 261152263

Your last name: Agarwal

Your first name: Lakshya

Your student ID: 261149449

# Part I: Political Analytics

In 2016, the U.S. had a general presidential election between two main candidates: (i) Donald Trump (from the Republican party); and (ii) Hillary Clinton (from the Democratic party). Hillary Clinton is a moderate candidate. Donald Trump, in contrast, is a right-wing candidate with more drastic conservative views. Ultimately, Trump won by a small margin, attracting voters with specific characteristics. I have collected data on all U.S. county districts, with the following information:

## Response variable

- winner: Candidate who obtained more votes in a given county {Clinton, Trump}

## Predictors

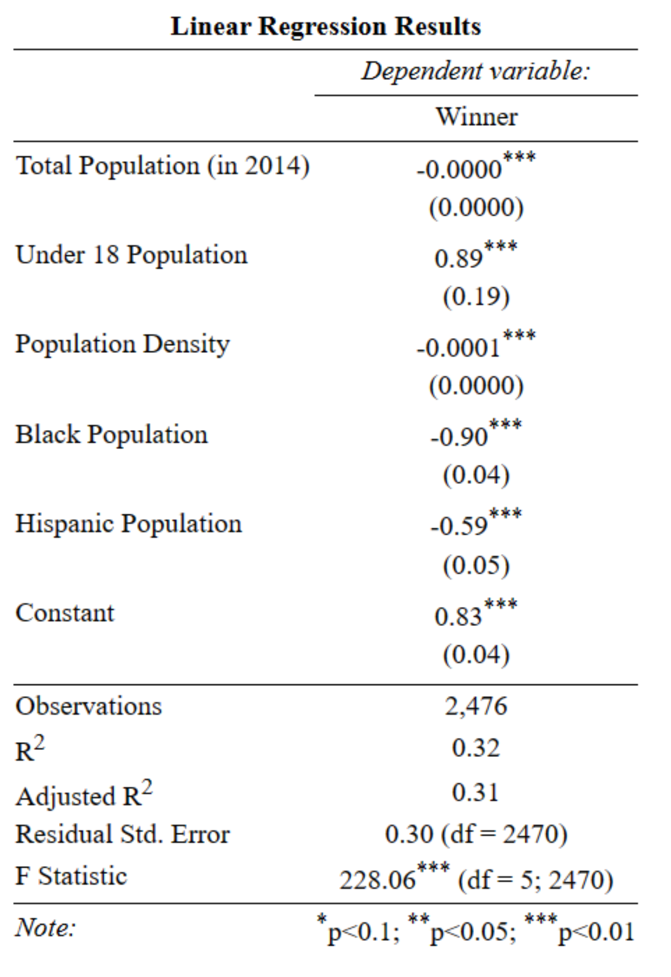
- state: 2-letter state abbreviation
- county: The county's name
- fips: The county's Federal Information Processing System (FIPS) code
- pop2010: The county's population in the 2010 census
- pop2014: The county's population in the 2014 census
- pop_change: population change in percentage terms between 2010 and 2014
- under5: % of county population under 5 years of age.
- under18: % of county population under 18 years of age.
- over65: % of county population over 65 years of age.
- female: % of county population that is female.
- black: % of county population that are Black.
- hispanic: % of county population that are Hispanic.
- undergrad: % of population (aged over 25) holding a bachelor degree.
- density: population density (habitants per square mile)

# 1. The linear probability model (5 points)

We will explore the following relationship:

winner=f(pop2014, under18, density, black, hispanic)

A) (1 point) Run a linear probability model using the variables above. Paste your regression results below:



**Linear Regression Results**

| | *Dependent variable:* |
|---|---|
| | Winner |
| Total Population (in 2014) | -0.0000*** |
| | (0.0000) |
| Under 18 Population | 0.89*** |
| | (0.19) |
| Population Density | -0.0001*** |
| | (0.0000) |
| Black Population | -0.90*** |
| | (0.04) |
| Hispanic Population | -0.59*** |
| | (0.05) |
| Constant | 0.83*** |
| | (0.04) |
| Observations | 2,476 |
| $R^2$ | 0.32 |
| Adjusted $R^2$ | 0.31 |
| Residual Std. Error | 0.30 (df = 2470) |
| F Statistic | 228.06*** (df = 5; 2470) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

B) (1 point) Interpret the value of the above regression coefficients.

- **population:** The coefficient of the population is very small and negative which suggests that for each additional person in the population, the probability of Trump winning decreases slightly, with everything else being constant. The coefficient is statistically significant at the 1% level.

- **under18:** The coefficient of the under18 suggests that for each unit increase in the population of under 18 years of age, the probability of Trump winning increases by approximately 0.89 percentage points, assuming other factors are kept constant. This coefficient is statistically significant at 1% level.

- **density:** The coefficient of density is -0.0001 which is also statistically significant at 1% level. It suggests that for each additional person per square mile, the probability of Trump winning decreases by 0.01 percentage points, holding all other variables constant.

- **black:** The coefficient of black indicates that for each one-percentage-point increase in the black population, the probability of Trump being the winner decreases by approximately 0.897 percentage points, with other variables held constant. This is a statistically significant effect at the 1% level.

- **hispanic:** The coefficient is -0.589, which is significant at the 1% level, meaning that for each one-percentage-point increase in the Hispanic population, the probability of Trump being the winner decreases by approximately 0.589 percentage points, everything else kept constant.

C) (0.5 points) According to this model, what is the probability that Trump will win a county with a 250k population in 2014, where 10% are under 18, with a population density of 175 inhabitants per square mile, and where 45% are black and 10% are hispanic?

**Your prediction:** 0.396

D) (0.5 points) According to this model, what is the probability that Trump will win a county with a 1 million population in 2014, where 20% are under 18, with a population density of 1000 inhabitants per square mile, and where 85% are black and 5% are hispanic?

**Your prediction:** -0.062

E) (2 points) What are the two problems of the linear probability model? Identify these problems, using examples from our election model, above.

**Problem 1:**
**Boundaries:** Linear probability models do not constrain the predictions to be within the [0, 1] range, which can lead to nonsensical predictions like negative probabilities or probabilities greater than 1. This can be seen from Part (D) where we got a negative probability that Trump will win within that county.

**Problem 2:**
**Non-linearity:** Linear probability models assume that all the predictors in the model have a linear relationship with the probability of the outcome, which cannot be true because of the bounded nature of probabilities. Hence the relationship is inherently non-linear.

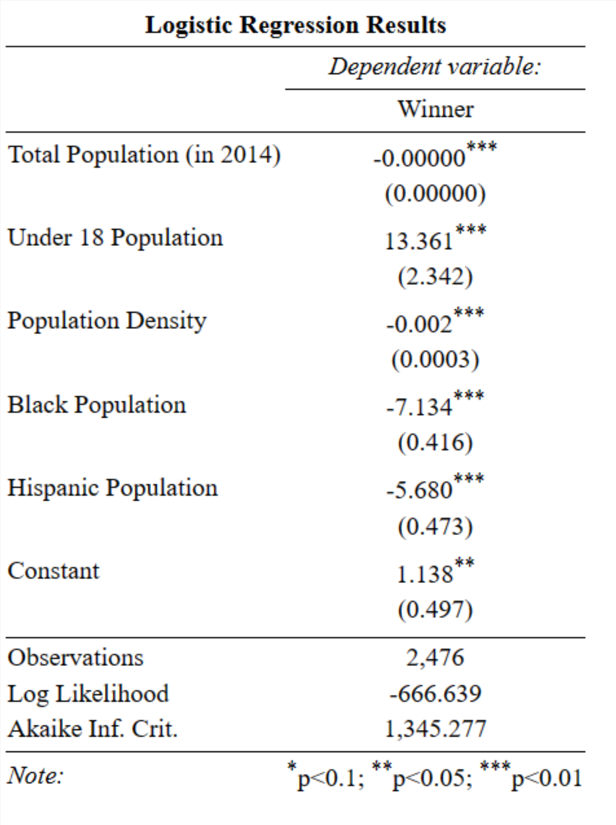
The model may suggest a positive impact on Trump's probability of winning when the under-18 population increases, but it predicts a negative probability in extreme cases like an 85% black population (as in Part D) because it doesn't account for diminishing effects as predicted probabilities approach 0 or 1.

# 2. Logistic Regression (10 points)

A) (1 point) We wish to test the following relationship

winner=f(pop2014, under18, density, black, hispanic)

This time, however, we wish to use a logistic regression. Run a logistic regression for the model above, and paste the output below:



**Logistic Regression Results**

| | *Dependent variable:* |
|---|---|
| | Winner |
| Total Population (in 2014) | -0.00000*** |
| | (0.00000) |
| Under 18 Population | 13.361*** |
| | (2.342) |
| Population Density | -0.002*** |
| | (0.0003) |
| Black Population | -7.134*** |
| | (0.416) |
| Hispanic Population | -5.680*** |
| | (0.473) |
| Constant | 1.138** |
| | (0.497) |
| Observations | 2,476 |
| Log Likelihood | -666.639 |
| Akaike Inf. Crit. | 1,345.277 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

B) (1 point) How many iterations did the model take to find the coefficients that best fit the data?

**Your answer:** 6

C) (1 point) According to this model, what is the probability that Trump will win a county with a 1 million population in 2014, where 20% are under 18, with a population density of 1000 inhabitants per square mile, and where 85% are black and 5% are hispanic?

**Your R code:**

```
prob_1 <- data.frame(pop2014 = 1000000, under18 = 20/100, density = 1000, black = 85/100,
hispanic = 5/100)
predict(reg2, prob_1, type="response")
```

**Your answer:** 0.001551897

D) (1 point) What is the R-squared value of this logistic regression?

**Your answer:**
R2 = 0.477

E) (2 points) Suppose a county has a 250,000 population in 2014, where 25% are under 18, with a density of 100 inhabitants per square mile and where 10% are hispanic. Using the regression output, complete the sentence below:

If the number of black people in the country is below 44.9%, the predicted winner is Trump.

Otherwise, the predicted winner is Clinton.

**Your work:**

The logistic regression equation is given as below -

$$log(\frac{p}{1-p}) = b0 + b1 * pop\_2014 + b2 * under\_18 + b3 * density + b4 * black + b5 * hispanic$$

where p = probability of Trump winning and b0 to b5 are the coefficients from the logistic regression output. To find the threshold value of the black population where the predicted winner changes, we will put p = ½ and then solve the equation for the variable 'black'. This calculation will give us the percentage of the black population that changes the predicted winner.

Substituting the values of coefficients from the logistic regressions summary results and calculating the value of black threshold -

$$log(1) = 1.14 - 0.0000021 * 250000 + 13.4 * 0.25 - 0.00183 * 100 - 7.13 * black \ - 5.68 * 0.1$$

Solving for black threshold we get,

$$black \ = 0.4489935$$

Implying that the probability of Trump winning from not winning changes (keeping everything else constant) as the black population percentage goes below 44.89%

F) (1.5 points) Suppose I wish to test the following relationship using logistic regression:

winner=f(hispanic, undergrad)

And that I have four observations in my dataset:

- **Observation 1:** winner=Trump, hispanic=0.15, undergrad=0.20
- **Observation 2:** winner=Clinton, hispanic=0.25, undergrad=0.55
- **Observation 3:** winner=Trump, hispanic=0.05, undergrad=0.05
- **Observation 4:** winner=Clinton, hispanic=0.75, undergrad=0.10

What is the value of the likelihood function *$L(b_o,b_1,b_2)$* if we let $b_o$=-0.75, $b_1$=0.03,$b_2$=0.01? Estimate the likelihood manually, and show your work below:

**Your work:**
The likelihood function is given as below -

$$L(b_0, b_1, b_2) = \prod_{i:y=1}^{4} p(x_i) \prod_{i:y=0}^{4} (1 - p(x_i))$$

where,

$$p(x_i) = \frac{e^{(b_0 + b_1 .hispanic_i + b_2 .undergrad_i)}}{1 + e^{(b_0 + b_1 .hispanic_i + b_2 .undergrad_i)}},$$

and $y_i = 1$ if Trump wins and $y_i = 0$ if Clinton wins

Given that $b_0 = -0.75, b_1 = 0.03 \; and \; b_2 = 0.01$, we will calculate $p(x_i)$ for all the observations given -

Observation 1: winner = Trump, hispanic = 0.15, undergrad = 0.20

$$p(x_1) = \frac{e^{(-.075+0.03*0.15+0.01*0.20)}}{1+e^{(-.075+0.03*0.15+0.01*0.20)}} \approx 0.3222393$$

Observation 2: winner=Clinton, hispanic=0.25, undergrad=0.55

$$p(x_2) = \frac{e^{(-.075+0.03*0.25+0.01*0.20)}}{1+e^{(-.075+0.03*0.25+0.01*0.20)}} \approx 0.3236605$$

Observation 3: winner=Trump, hispanic=0.05, undergrad=0.05

$$p(x_3) = \frac{e^{(-.075+0.03*0.05+0.01*0.05)}}{1+e^{(-.075+0.03*0.05+0.01*0.05)}} \approx 0.3212572$$

Observation 4: winner=Clinton, hispanic=0.75, undergrad=0.10

$$p(x_4) = \frac{e^{(-.075+0.03*0.75+0.01*0.1)}}{1+e^{(-.075+0.03*0.75+0.01*0.1)}} \approx 0.3259632$$

Solving for the likelihood function now -

$$L(b_0, b_1, b_2) = p(x_1) * (1 - p(x_2)) * p(x_3) * (1 - p(x_4))$$
$$L(b_0, b_1, b_2) = 0.322 * (1 - 0.324) * 0.321 * (1 - 0.326) = 0.0472$$

**Your answer:**
Observation 1: 0.0.3222393
Observation 2: 0.3236605
Observation 3: 0.3212572
Observation 4: 0.3259632

Likelihood: 0.04719323

G) (1.5 points) What is the value of the likelihood function *L(b₀,b₁,b₂)* if we let $b_o$=-1, $b_1$=2, $b_2$=4? Estimate the likelihood manually, and show your work below:

**Your work:**
As above, given that $b_0 = -1, b_1 = 2 \text{ and } b_2 = 4$, we will calculate $p(x_i)$ for all the observations given -

Observation 1: winner = Trump, hispanic = 0.15, undergrad = 0.20

$$p(x_1) \approx 0.5249792$$

Observation 2: winner=Clinton, hispanic=0.25, undergrad=0.55

$$p(x_2) \approx 0.8455347$$

Observation 3: winner=Trump, hispanic=0.05, undergrad=0.05

$$p(x_3) \approx 0.3318122$$

Observation 4: winner=Clinton, hispanic=0.75, undergrad=0.10

$$p(x_4) \approx 0.7109495$$

Solving for the likelihood function now -

$$L(b_0, b_1, b_2) = p(x_1) * (1 - p(x_2)) * p(x_3) * (1 - p(x_4))$$

$$L(b_0, b_1, b_2) = 0.525 * (1 - 0.846) * 0.332 * (1 - 0.711) = 0.0077$$

**Your answer:**
Observation 1: 0.5249792
Observation 2: 0.8455347
Observation 3: 0.3318122
Observation 4: 0.7109495

Likelihood: 0.007777482

H) (1 point) Which one of the two cases above more closely fits the data? Explain your reasoning:

**Your answer:**
The values in Part F (bo=-0.75, b1=0.03, b2=0.01) more closely fit the data.

**Your explanation:**
Since logistic regression tries to maximize the likelihood function (through maximum likelihood estimation), the coefficient values that give a higher value of $L$ will more closely fit the given data. Therefore, in the given scenario with 4 observations, the coefficients in Part F give a higher likelihood of ~0.0472, as compared to the coefficients in Part G, which give a likelihood value of ~0.0078.

# Part II: Using Data to Detect Fake wines

Fake wines are becoming a pervasive problem. The high demand for high-quality wine has driven prices up across some vineyards. Counterfeiters have taken advantage of this occurrence and began forging fine-wine bottles.[1]

Detecting fake wine bottles is hard, especially because counterfeiters can replicate wine bottles labels with incredible accuracy. It is, however, nearly impossible to replicate the chemical properties of a fine wine in a lab. Nowadays, chemical engineers and data scientists are using classification methods to detect fake wines.

In particular, wines have thirteen primary chemical concentrations, which serve as the fingerprints for each winery/cultivar. By carefully studying these properties, we can identify the origin of the grapes and, hence, trace back the origin of the wine.

I have gathered real data from 178 wine bottles belonging to three cultivars of wine in Italy: Cultivar 1, Cultivar 2, and Cultivar 3. Each cultivar produces wines with similar characteristics. It would be very hard to distinguish the flavor of these three wines, even by expert tasters. However, we can use discriminant analysis to achieve this goal in a more scientific way.

## Response variable

- Cultivar*:* The origin of the wine. In our dataset, the bottles of the three cultivars from a region in South Italy.

## Predictors

I have gathered thirteen properties for each wine bottle:

- Alcohol (in %)
- Malic acid (in g/L)
- Ash (in g/L)
- Alkalinity of ash
- Magnesium
- Total phenols
- Flavonoids
- Nonflavonoid phenols
- Proanthocyanidins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

[1] I recommend you to watch the documentary "Sour Grapes," an entertaining story of the world's biggest wine forger.

# 3. Linear discriminant analysis with two predictors (10 points)

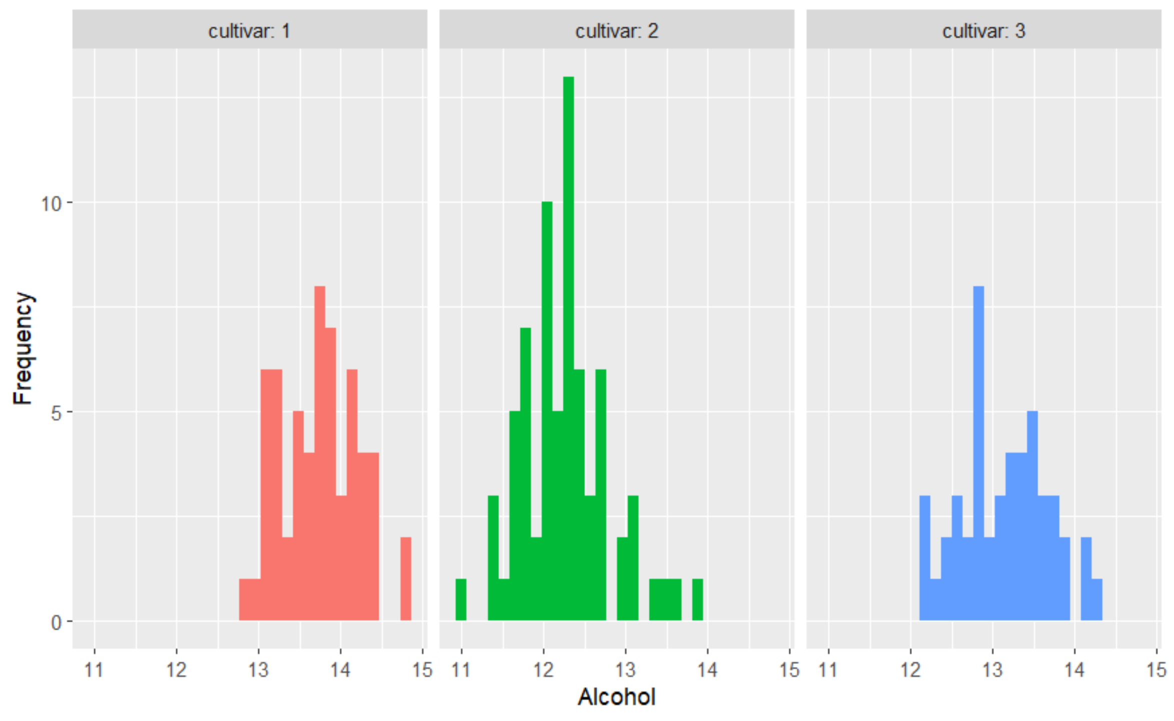
We want to understand the probability that wine bottle *i* belongs to cultivar Y={1,2,3}, given the alcohol and acid properties of the bottle:

- Pr(Y=k | alcohol, acid)

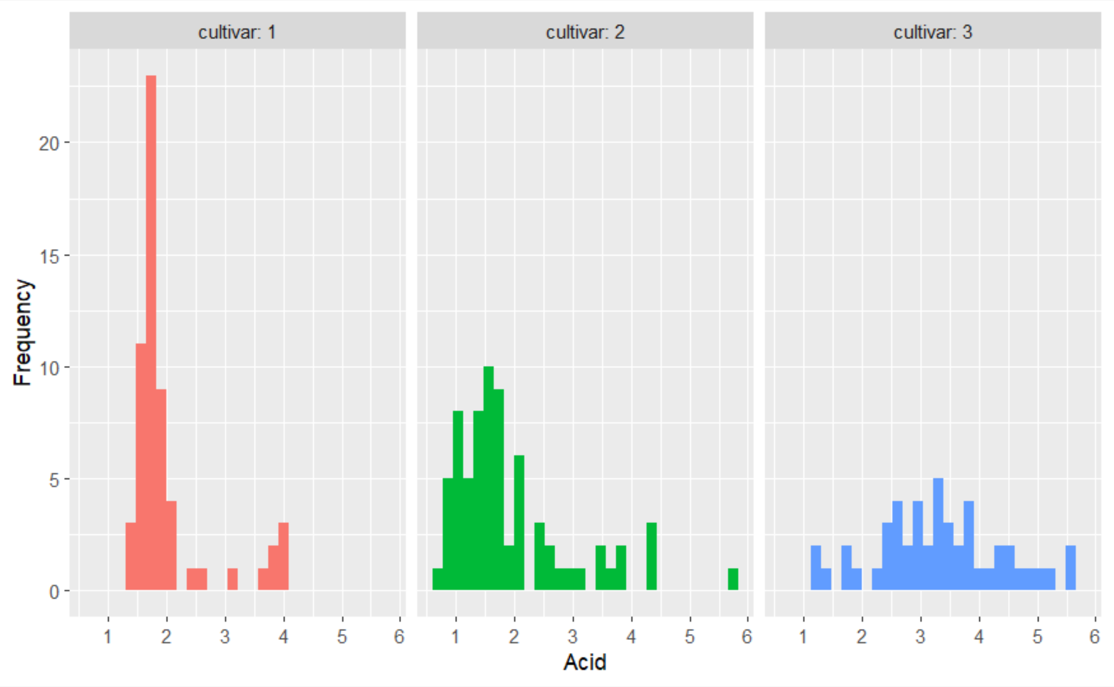
A) (1 point) What are the prior probabilities of each class?

- $\pi_1 = 59/178 = 0.331$
- $\pi_2 = 71/178 = 0.399$
- $\pi_3 = 48/178 = 0.270$

B) (1 point) We want to find the probability density functions $f_k$(alcohol) for k=1,2,3. Plot the histograms of alcohol level for each class. Make sure all histograms have the same range in the x axis (please ensure all axis information is legible):



C) (1 point) We want to find the functions $f_k$(acid). Plot the histograms of acid level for each class. Make sure all histograms have the same range in the x axis:



D) (1 point) Is it reasonable to assume that all $f_k$() functions are normally distributed?

**Choose one:**

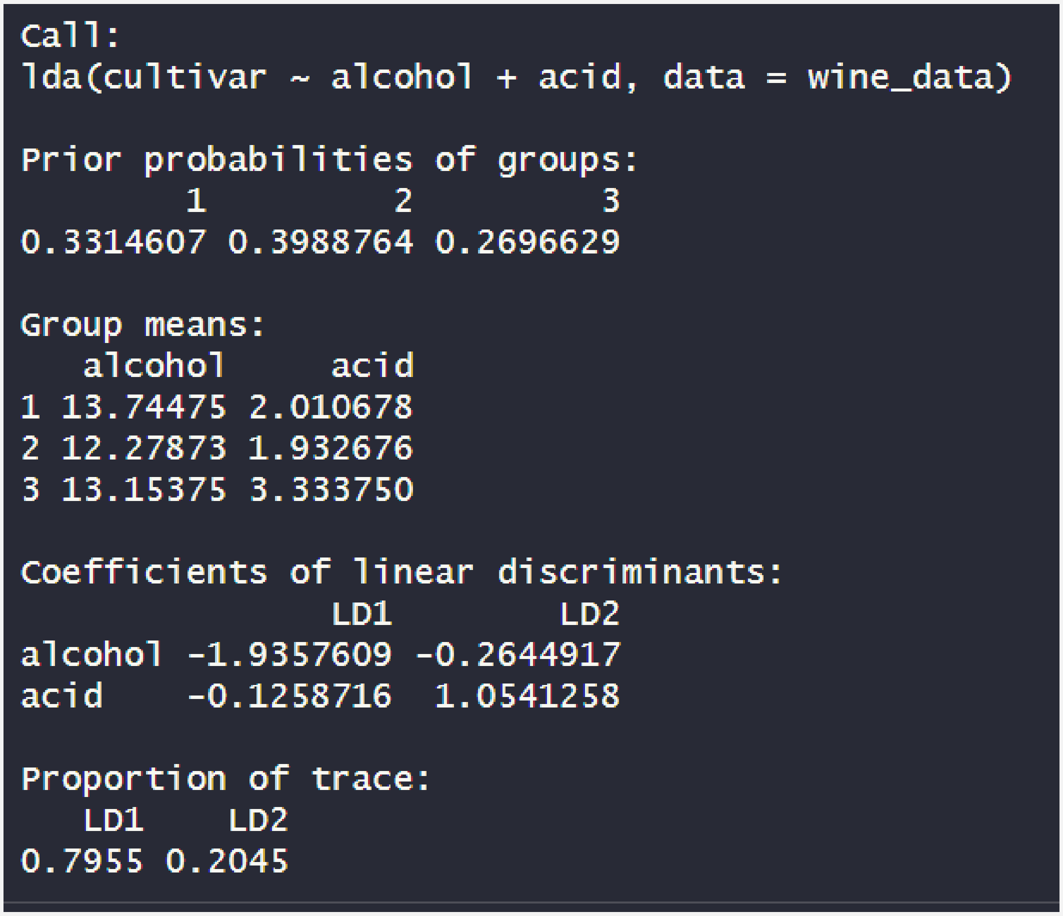
- **Yes**
- **No** ✓

**Your explanation:** No, it is reasonable to assume that $f_k$() functions are normally distributed since the data appears to be skewed (at least for the `acid` variable).

E) (1 point) Run a linear discriminant analysis to find

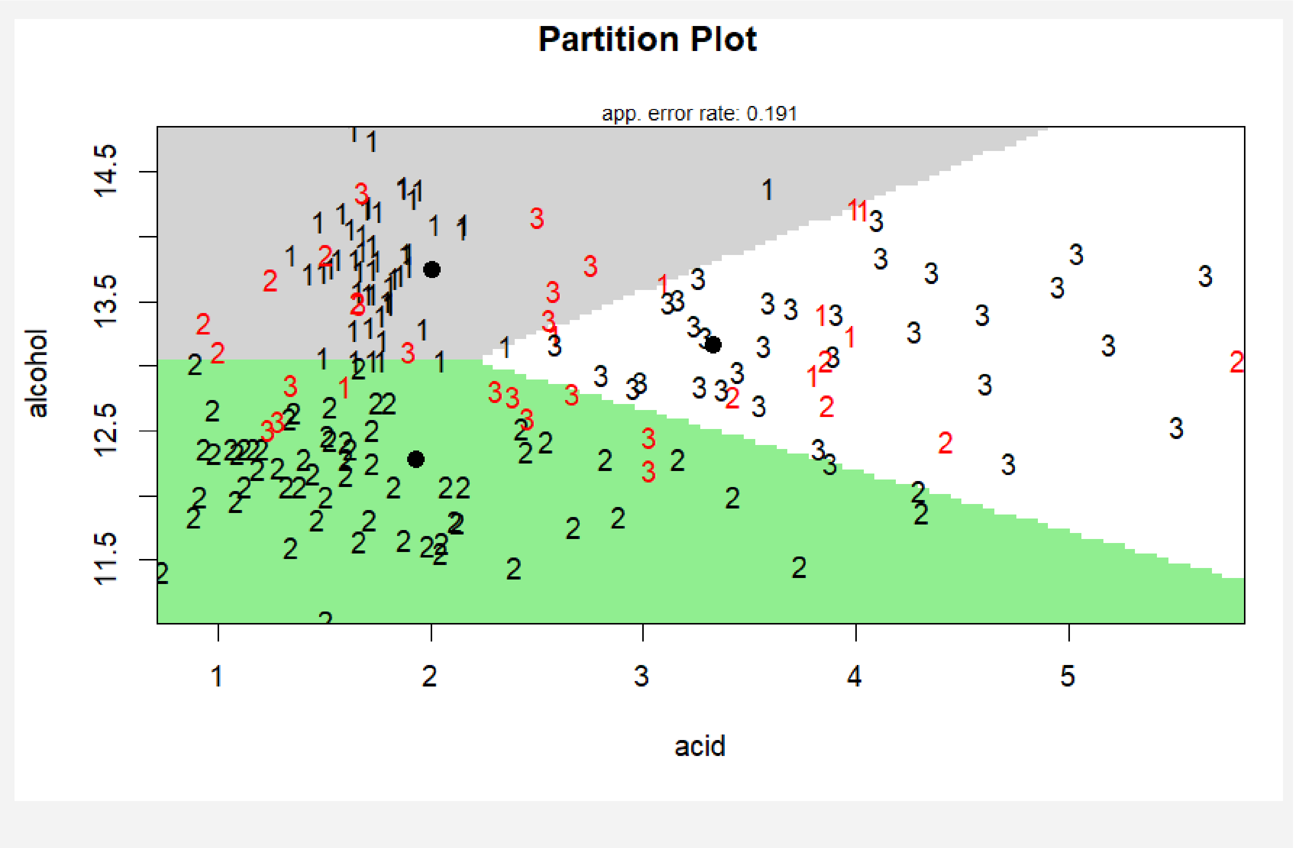
Pr(Y=K | alcohol, acid)

Paste the lda's output below:



```
Call:
lda(cultivar ~ alcohol + acid, data = wine_data)

Prior probabilities of groups:
        1         2         3 
0.3314607 0.3988764 0.2696629 

Group means:
   alcohol     acid
1 13.74475 2.010678
2 12.27873 1.932676
3 13.15375 3.333750

Coefficients of linear discriminants:
               LD1        LD2
alcohol -1.9357609 -0.2644917
acid    -0.1258716  1.0541258

Proportion of trace:
   LD1    LD2 
0.7955 0.2045 
```

F) (1 point) Plot the classification regions using the *partimat* function:



G) (1 point) Interpret the results above from the partimat() function.

**Your answer:** The plot generated by the partimat() function illustrates the decision regions that are determined through Linear Discriminant Analysis (LDA). The shaded gray area on the plot represents the classification boundary for Cultivar 1, indicating where wine bottle observations are categorized within this class. Similarly, the green shaded area demarcates the region for Cultivar 2, and the uncolored (white) region signifies the classification space for Cultivar 3. Each region corresponds to the respective cultivar classification based on the properties analyzed by the LDA. The red colored observations are incorrectly predicted by the model.

H) (1 point) What is the error rate of the linear discriminant model? What does this mean?

**Your answer:** The model has an error rate of 19.1% which means that 19 observations out of 100 would be misclassified.

l) (2 points) I'm selling a wine that supposedly comes from Cultivar 1. The wine has an Alcohol level of 14% and a Malic Acid Level of 3 g/L. What is the probability that I'm lying about the wine's origin?

**Your answer:**
The probability that I'm lying about the wine's origin is ~28%.

The model predicts a ~72% probability that a wine with the given alcohol and acid level is from Cultivar 1. Therefore, the probability that I'm lying is = (1 - 0.72) = 0.28

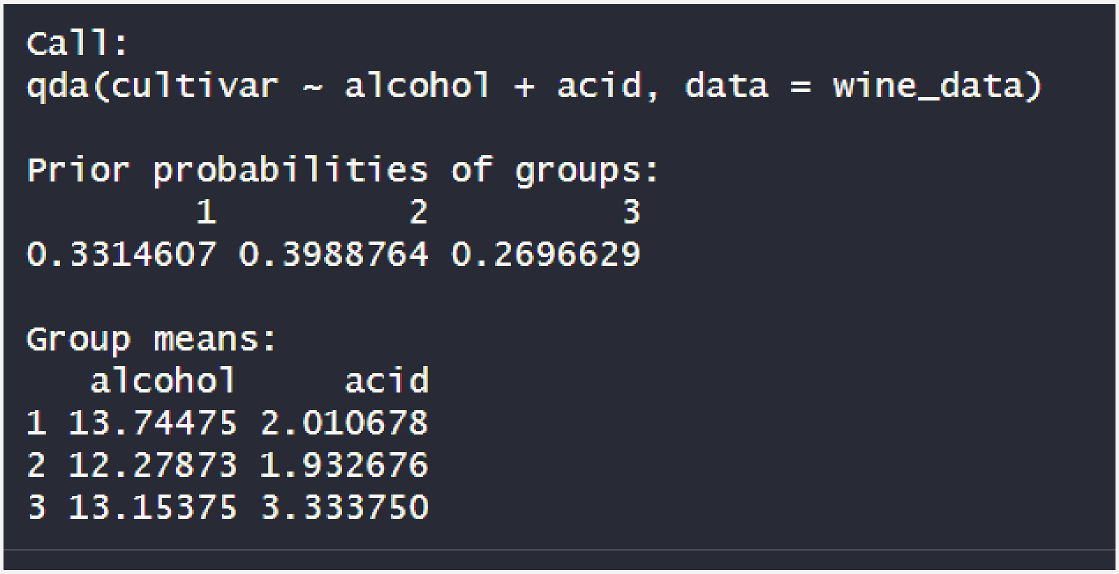
# 4. Quadratic discriminant analysis with two predictors (8 points)

A) (1.5 point) What is the difference between the linear and quadratic discriminant analysis?

**Your answer:**
The difference between linear and quadratic discriminant analysis lies in the assumption that standard deviation across classes is equal. In QDA, we relax the above assumption and find K prior probabilities, K means of each class, and K standard deviations of each class, as opposed to LDA, where we find K prior probabilities, K means of each class, and 1 common standard deviation.

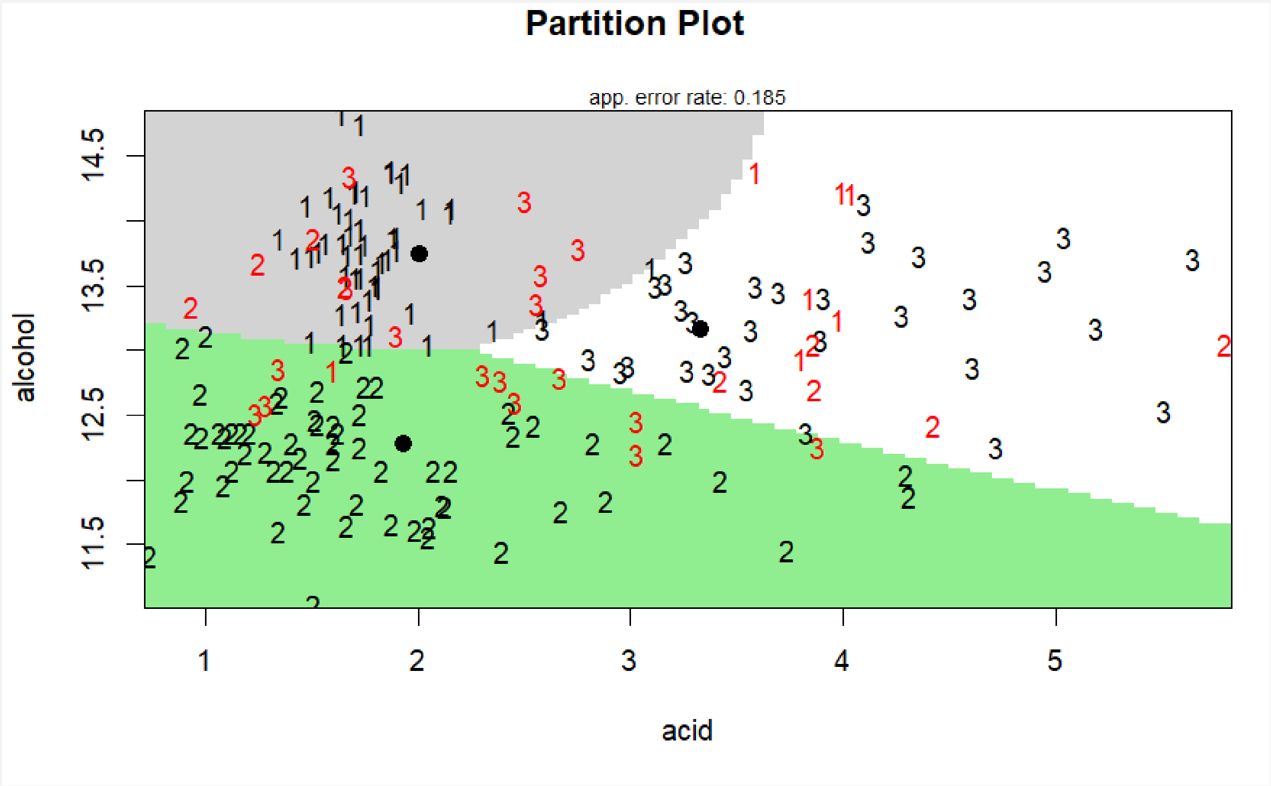
B) (1 point) Run a quadratic discriminant analysis to find

Pr(Y=K | alcohol, acid)



```
Call:
qda(cultivar ~ alcohol + acid, data = wine_data)

Prior probabilities of groups:
        1         2         3 
0.3314607 0.3988764 0.2696629 

Group means:
   alcohol     acid
1 13.74475 2.010678
2 12.27873 1.932676
3 13.15375 3.333750
```

C) (1.5 point) Plot the classification regions using the partimat() function:



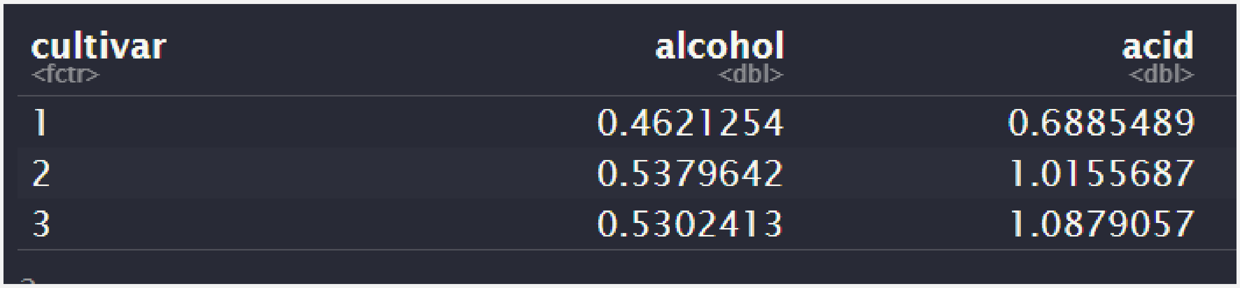
D) (2 points) What is the error rate of the quadratic discriminant model?

**Your answer:**
The quadratic discriminant model has an error rate of 18.5% which means that 18.5 observations out of 100 would be misclassified.

E) (2 points) Which model performs better in the training data: the linear or the quadratic model? How did you reach that conclusion?

**Your answer:**
The quadratic model performs better in the training data since it has a lower error rate (18.5%), as compared to the linear model (19.1%).

Furthermore, if we compute the standard deviation of our predictors (alcohol and acid) for each class, we see that they are different. Therefore, the quadratic model is theoretically a better fit to the training data as well.



| cultivar | alcohol | acid |
|---|---:|---:|
| <fctr> | <dbl> | <dbl> |
| 1 | 0.4621254 | 0.6885489 |
| 2 | 0.5379642 | 1.0155687 |
| 3 | 0.5302413 | 1.0879057 |

# For further practice (Not graded): Cross-validation for Classification methods

A) Run a validation-set test test to determine the out-of-sample performance of the linear and quadratic discriminant models from the questions above. (You will have to figure out how by combining your knowledge from Lectures 6 and 8).

Paste the code below:

**Your answer:**

```
set.seed(420)
sample = sample.split(wine_data$cultivar, SplitRatio = 0.6)
train_set = subset(wine_data, sample == TRUE)
test_set = subset(wine_data, sample == FALSE)

lda_model <- lda(cultivar ~ alcohol + acid, data = train_set)
lda_predictions <-
  predict(lda_model, newdata = test_set)$class

lda_accuracy <- sum(lda_predictions == test_set$cultivar) / length(lda_predictions)
print(glue("Accuracy of LDA: {round(lda_accuracy, 3)}"))

qda_model <- qda(cultivar ~ alcohol + acid, data = train_set)
qda_predictions <-
    predict(qda_model, newdata = test_set)$class

qda_accuracy <- sum(qda_predictions == test_set$cultivar) / length(qda_predictions)
print(glue("Accuracy of QDA: {round(qda_accuracy, 3)}"))
```

B) Paste the result of the test for the linear and quadratic models, and the average error rate for each test.

**Your answer:**
**The accuracy for the linear model is:** 0.789
**The accuracy for the quadratic model is:** 0.803

- Submission: Please save in colour as a PDF and submit via MyCourses. If you don't submit a color PDF, there will be a 2-point penalty.

- Code: Submit code in a separate file.

- Due date: November 16, 11:59 pm.