Step 1: Introduction
> library("Rcmdr")

> Dataset <-
readXL("//apporto.com/dfs/UALR/Users/saoyedotun_ualr/Desktop/MidusCollege2022.xls", rownames=FALSE, header=TRUE, na="", sheet="Sheet1", stringsAsFactors=TRUE)

summary(Dataset)

Step 2 - 1: Linear Probability Models
> lin_prob_model_1 <- lm(col~momedu+race0+sex, data=Dataset)

> summary(lin_prob_model_1)

```
Call:
lm(formula = col ~ momedu + race0 + sex, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9845 -0.4986  0.1592  0.3873  0.6725 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.327526   0.020345  16.098   <2e-16 ***
momedu      0.057032   0.003533  16.142   <2e-16 ***
race0       0.100415   0.057144   1.757    0.079 .  
sex         0.029624   0.020227   1.465    0.143    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4621 on 2145 degrees of freedom
Multiple R-squared:  0.1118,	Adjusted R-squared:  0.1105 
F-statistic: 89.96 on 3 and 2145 DF,  p-value: < 2.2e-16
```

Q: Are the predictor variables significant? Interpret the effect of Maternal Education and Race2?
A: Maternal education (momedu) is statistically significant predictor as p-value is less than 0.05. For each additional year of schooling, the probability of college completion increases by 5.7 *percentage points*.
Race0 is marginally significant, as P<0.10, so for people of Race0, college completion increases by 10 *percentage points.*

Step 2 - 2: Linear Probability Models
> lin_prob_model_2 <- lm(col~momedu+paedu+race0+race3+sex, data=Dataset)

> summary(lin_prob_model_2)

```
Call:
lm(formula = col ~ momedu + paedu + race0 + race3 + sex, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9188 -0.4511  0.1484  0.3834  0.9921 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.224948   0.024397   9.220  < 2e-16 ***
momedu       0.043291   0.003933  11.008  < 2e-16 ***
paedu        0.017517   0.002335   7.503 9.06e-14 ***
race0        0.132267   0.056547   2.339   0.0194 *  
race3       -0.424426   0.186404  -2.277   0.0229 *  
sex          0.023937   0.019971   1.199   0.2308    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4559 on 2143 degrees of freedom
Multiple R-squared:  0.1363,    Adjusted R-squared:  0.1343 
F-statistic: 67.66 on 5 and 2143 DF,  p-value: < 2.2e-16
```

Q: Are the predictor variables significant? Interpret the effect of the predictor variables?
A: Maternal education (momedu), paternal education, race0, race3 are all statistically significant predictor as p-value is less than 0.05. For each additional year of schooling, the probability of college completion for maternal education increases by 4.3 *percentage points*. For each additional year of schooling, the probability of college completion for paternal education increases by 1.7 *percentage points*.
For people of Race0, college completion increases by 13.2 *percentage points.*
For people of Race3, college completion decreases by -42.4 *percentage points.*

Q: Compare model 2 to model 1. Which one is better and why?
A: Model 2 is better, reason being that Model 2 has greater explanator power than model 1, with an adjusted R squared of 0.13 versus 0.11.

 # Step 2 - 3: Linear Probability Models
Q: Generate a forecast for completed education. Summarize the forecast.

> forcast <- predict(lin_prob_model_2, Dataset, type="response")
> forcast
> summary(forcast)

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-0.08612  0.46534  0.61658  0.60028  0.71381  1.05791 
```

Step 3: Logit

> GLM.1 <- glm(col ~ momedu + race0 + sex, family=binomial(logit), data=Dataset)

> summary(GLM.1)

```
Call:
glm(formula = col ~ momedu + race0 + sex, family = binomial(logit),
    data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1733  -1.1720   0.6093   0.9655   1.5368

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.81404    0.09690  -8.401   <2e-16 ***
momedu       0.26709    0.01841  14.507   <2e-16 ***
race0        0.45111    0.27221   1.657   0.0975 .
sex          0.13868    0.09501   1.460   0.1444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2892.1  on 2148  degrees of freedom
Residual deviance: 2637.5  on 2145  degrees of freedom
AIC: 2645.5

Number of Fisher Scoring iterations: 4
```

```
> exp(coef(GLM.1))  # Exponentiated coefficients ("odds ratios")

(Intercept)      momedu       race0         sex
  0.4430645   1.3061589   1.5700521   1.1487537
```

> logitmfx(formula = col ~ momedu + race0 + sex, data=Dataset, atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL, start = NULL, control = list())

```
Call:
logitmfx(formula = col ~ momedu + race0 + sex, data = Dataset,
    atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL,
start = NULL, control = list())

Marginal Effects:
           dF/dx Std. Err.       z   P>|z|
momedu 0.0633229 0.0043189 14.6619 < 2e-16 ***
race0  0.1005896 0.0562302  1.7889 0.07363 .
sex    0.0328023 0.0224109  1.4637 0.14328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

dF/dx is for discrete change for the following variables:

[1] "race0" "sex"
```

Q: Are the predictor variables significant? Interpret the effect of the predictor variables?
A: Maternal education (momedu) is statistically significant predictor as p-value is less than 0.05. For each additional year of schooling, the probability of college completion increases by 26.7 *percentage points*.

> GLM.2 <- glm(col ~ momedu + race0 + sex + paedu + race3, family=binomial(logit), data=Dataset)

> summary(GLM.2)

```
Call:
glm(formula = col ~ momedu + race0 + sex + paedu + race3, family =
binomial(logit), data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0721  -1.0628   0.5798   0.9499   2.3683

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.33145    0.12199 -10.914  < 2e-16 ***
momedu       0.20981    0.01981  10.590  < 2e-16 ***
race0        0.60337    0.27690   2.179   0.0293 *
sex          0.11469    0.09644   1.189   0.2343
paedu        0.08444    0.01135   7.442 9.91e-14 ***
race3       -2.41048    1.17956  -2.044   0.0410 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2892.1  on 2148  degrees of freedom
Residual deviance: 2576.1  on 2143  degrees of freedom
AIC: 2588.1

Number of Fisher Scoring iterations: 4
```

Q: Are the predictor variables significant? Interpret the effect of the predictor variables?
A: Maternal education (momedu), paternal education, race0, race3 are all statistically significant predictor as p-value is less than 0.05. For each additional year of schooling, the probability of college completion for maternal education increases by 20.9 *percentage points*. For each additional year of schooling, the probability of college completion for paternal education increases by 8.4 *percentage points*.
For people of Race0, college completion increases by 11.4 *percentage points.*
For people of Race3, college completion decreases by -24.1 *percentage points.*

> exp(coef(GLM.2)) # Exponentiated coefficients ("odds ratios")

```
(Intercept)      momedu       race0         sex       paedu       race3
  0.2640952   1.2334418   1.8282712   1.1215313   1.0881022   0.0897723
```

> logitmfx(formula = col ~ momedu + race0 + sex + paedu + race3, data=Dataset, atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL, start = NULL, control = list())

```
Call:
logitmfx(formula = col ~ momedu + race0 + sex + paedu + race3,
    data = Dataset, atmean = TRUE, robust = FALSE, clustervar1 = NULL,
    clustervar2 = NULL, start = NULL, control = list())

Marginal Effects:
            dF/dx  Std. Err.       z     P>|z|
momedu  0.0495382  0.0046462 10.6621 < 2.2e-16 ***
race0   0.1302580  0.0532068  2.4481 0.0143594 *
sex     0.0270289  0.0226780  1.1919 0.2333175
paedu   0.0199361  0.0026682  7.4716  7.92e-14 ***
race3  -0.4919448  0.1316338 -3.7372 0.0001861 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

dF/dx is for discrete change for the following variables:

[1] "race0" "sex"   "race3"
```

Q: Use the AIC to pick the best models for Logit between Equation 1 and 2
A: Model 2 has greater explanatory power. The smaller the AIC the better the model. Model 2 – 2588.1, model 1 – 2645.5

> logit <- predict(GLM.2, Dataset, type="response")
> logit
> summary(logit)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03938 0.45817 0.63690 0.60028 0.73621 0.93676
```

 # Step 3: Probit

> GLM.3 <- glm(col ~ momedu + race0 + sex, family=binomial(probit), data=Dataset)

> summary(GLM.3)

```
Call:
glm(formula = col ~ momedu + race0 + sex, family = binomial(probit), 
    data = Dataset)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2382  -1.1735   0.5971   0.9667   1.5330  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.49920    0.05857  -8.523   <2e-16 ***
momedu       0.16448    0.01082  15.197   <2e-16 ***
race0        0.26562    0.16436   1.616    0.106    
sex          0.08374    0.05778   1.449    0.147    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2892.1  on 2148  degrees of freedom
Residual deviance: 2634.9  on 2145  degrees of freedom
AIC: 2642.9

Number of Fisher Scoring iterations: 4
```

Q: Are the predictor variables significant? Interpret the effect of the predictor variables?
A: Maternal education (momedu) is statistically significant predictor as p-value is less than 0.05. For each additional year of schooling, the probability of college completion increases by 16.4 *percentage points*.

> exp(coef(GLM.3)) # Exponentiated coefficients ("odds ratios")

```
(Intercept)      momedu       race0         sex 
  0.6070162   1.1787769   1.3042365   1.0873488 
```

> logitmfx(formula = col ~ momedu + race0 + sex, data=Dataset, atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL, start = NULL, control = list())

```
Call:
logitmfx(formula = col ~ momedu + race0 + sex, data = Dataset,
    atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL,
start = NULL, control = list())

Marginal Effects:
           dF/dx Std. Err.       z   P>|z|
momedu 0.0633229 0.0043189 14.6619 < 2e-16 ***
race0  0.1005896 0.0562302  1.7889 0.07363 .
sex    0.0328023 0.0224109  1.4637 0.14328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

dF/dx is for discrete change for the following variables:

[1] "race0" "sex"
```

> GLM.4 <- glm(col ~ momedu + race0 + sex + paedu + race3, family=binomial(probit), data=Dataset)

summary(GLM.4)

```
Call:
glm(formula = col ~ momedu + race0 + sex + paedu + race3, family =
binomial(probit), data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1180  -1.0687   0.5641   0.9505   2.3106

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.816960   0.072943 -11.200  < 2e-16 ***
momedu       0.128252   0.011840  10.832  < 2e-16 ***
race0        0.347001   0.166042   2.090   0.0366 *
sex          0.070050   0.058449   1.198   0.2307
paedu        0.052495   0.006845   7.669 1.73e-14 ***
race3       -1.282439   0.641047  -2.001   0.0454 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2892.1  on 2148  degrees of freedom
Residual deviance: 2572.0  on 2143  degrees of freedom
AIC: 2584

Number of Fisher Scoring iterations: 4
```

Q: Are the predictor variables significant? Interpret the effect of the predictor variables?
A: Maternal education (momedu), paternal education, race0, race3 are all statistically significant predictor as p-value is less than 0.05. For each additional year of schooling, the probability of college completion for maternal education increases by 12.8 *percentage points*. For each additional year of schooling, the probability of college completion for paternal education increases by 5.2 *percentage points*.
For people of Race0, college completion increases by 34.7 *percentage points.*
For people of Race3, college completion decreases by -128.2 *percentage points.*

> exp(coef(GLM.4)) # Exponentiated coefficients ("odds ratios")

```
(Intercept)      momedu       race0         sex       paedu       race3
  0.4417724   1.1368395   1.4148186   1.0725623   1.0538977   0.2773601
```

> probitmfx(formula = col ~ momedu + race0 + sex + paedu + race3, data=Dataset, atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL, start = NULL, control = list())

```
Call:
probitmfx(formula = col ~ momedu + race0 + sex + paedu + race3,
    data = Dataset, atmean = TRUE, robust = FALSE, clustervar1 = NULL,
    clustervar2 = NULL, start = NULL, control = list())

Marginal Effects:
            dF/dx  Std. Err.       z     P>|z|
momedu  0.0489309  0.0045026 10.8673 < 2.2e-16 ***
race0   0.1239681  0.0544569  2.2764   0.02282 *
sex     0.0266858  0.0222283  1.2005   0.22993
paedu   0.0200281  0.0026069  7.6826 1.559e-14 ***
race3  -0.4552943  0.1584487 -2.8734   0.00406 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

dF/dx is for discrete change for the following variables:

[1] "race0" "sex"   "race3"
```

Q: Use the AIC to pick the best models for Probit between Equation 1 and 2
A: Model 2 has greater explanatory power. The smaller the AIC the better the model. Model 2 AIC is 2584, while model 1 AIC is 2642.9

> probit <- predict(GLM.4, Dataset, type="response")
> probit
> summary(probit)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03911 0.45793 0.63655 0.60247 0.73826 0.95167
```

Q: Compare forecasts from OLS with Logit and Probit.
A: Our summary shows that OLS forecast are bounded between 1 and below 0 whereas Logit and Probit forecasts are bounded between 0 and 1.