ISF 110, Lab 9 – Logistic and curvilinear regression analysis exercise

Introduction
Which type of regression analysis should we use – linear, curvilinear, or logistic – to solve a classification problem? In this lab, we will compare results obtained from linear, curvilinear, and logistic regression to choose the best model. Follow the instructions exactly and answer all the questions in the lab.

(1) Linear regression or logistic regression?

We will use the following dataset on 1200 high schools' academic performance in California. Start with a "do" file, open the dataset, and look at the DVs and IVs by running the "**tabulation**" command.

Create a Do file for Lab 9
clear all
set more off
use https://stats.idre.ucla.edu/stat/stata/webbooks/logistic/apilog, clear

The Academic Performance Index (api00) is a measurement of academic performance and progress of individual schools in California. A numeric **api00** score ranges from a low of 200 to a high of 1000.

Our dependent variable is called **hiqual**. This variable was created using a cut-off point of 745 from the **api00** score. Hence, **api00** values of 744 and below were coded as 0 (with a label of "not_high_qual") and values of 745 and above were coded as 1 (with a label of "high_qual").

Our hypothesis is that **"school quality depends on students' socioeconomic status."**

To measure students' socioeconomic status, we will use **avg_ed**, which is a continuous measure of the average education on a scale 1-5, where 1 means low and 5 means high average education of the parents of the students in the participating high schools. High average education corresponds to high socioeconomic status of the parents as well as the students.

To test the hypothesis, first run a linear regression, obtain the fitted values, and graph them against the observed values of the variables.

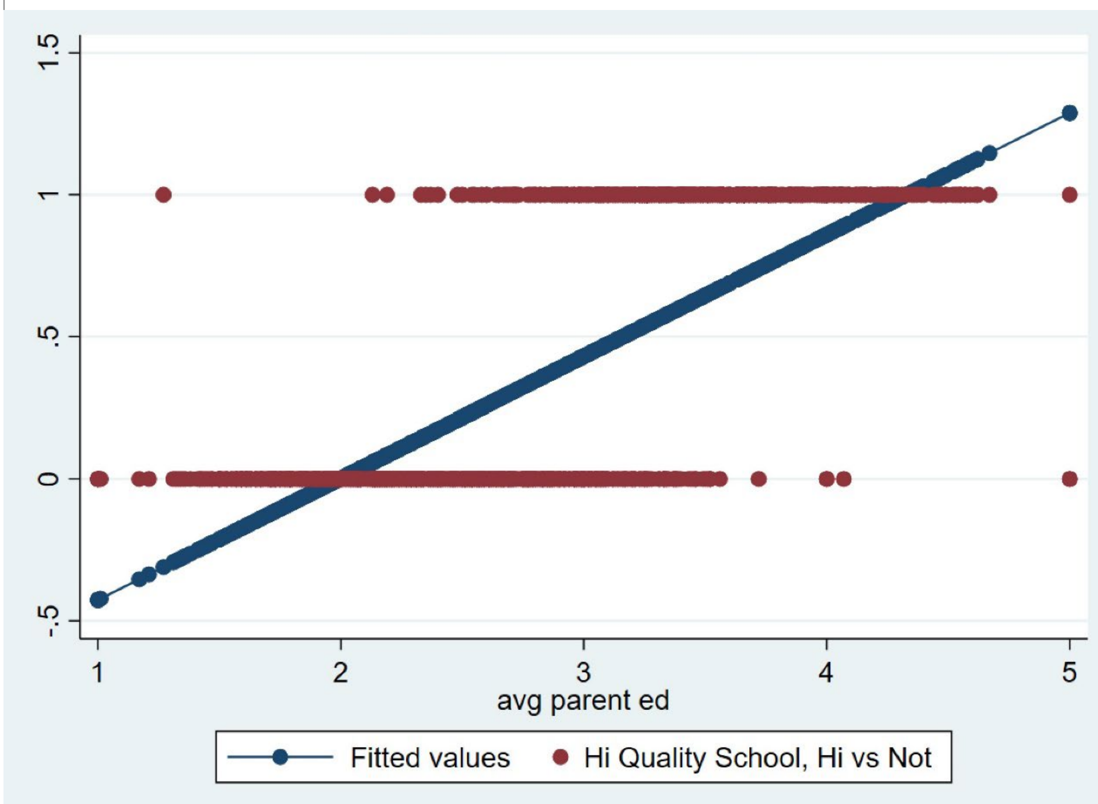
reg hiqual avg_ed
predict y
twoway scatter y hiqual avg_ed, connect(1 .)

Upon inspecting the graph, what limitations do you find in the linear regression for a binary outcome (our DV)?

```
. reg hiqual avg_ed

      Source |       SS           df       MS      Number of obs   =     1,158
-------------+----------------------------------   F(1, 1156)      =   1136.02
       Model |  126.023363         1  126.023363   Prob > F        =    0.0000
    Residual |  128.240023     1,156  .110934276   R-squared       =    0.4956
-------------+----------------------------------   Adj R-squared   =    0.4952
       Total |  254.263385     1,157  .219760921   Root MSE        =    .33307

------------------------------------------------------------------------------
      hiqual |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avg_ed |   .4286426   .0127175    33.70   0.000     .4036906    .4535946
       _cons |  -.8549049   .0363655   -23.51   0.000    -.9262547   -.7835551
------------------------------------------------------------------------------
```



It doesn't fit! It's missing a lot of center values, and extends above and below the bounds of the classification scale.

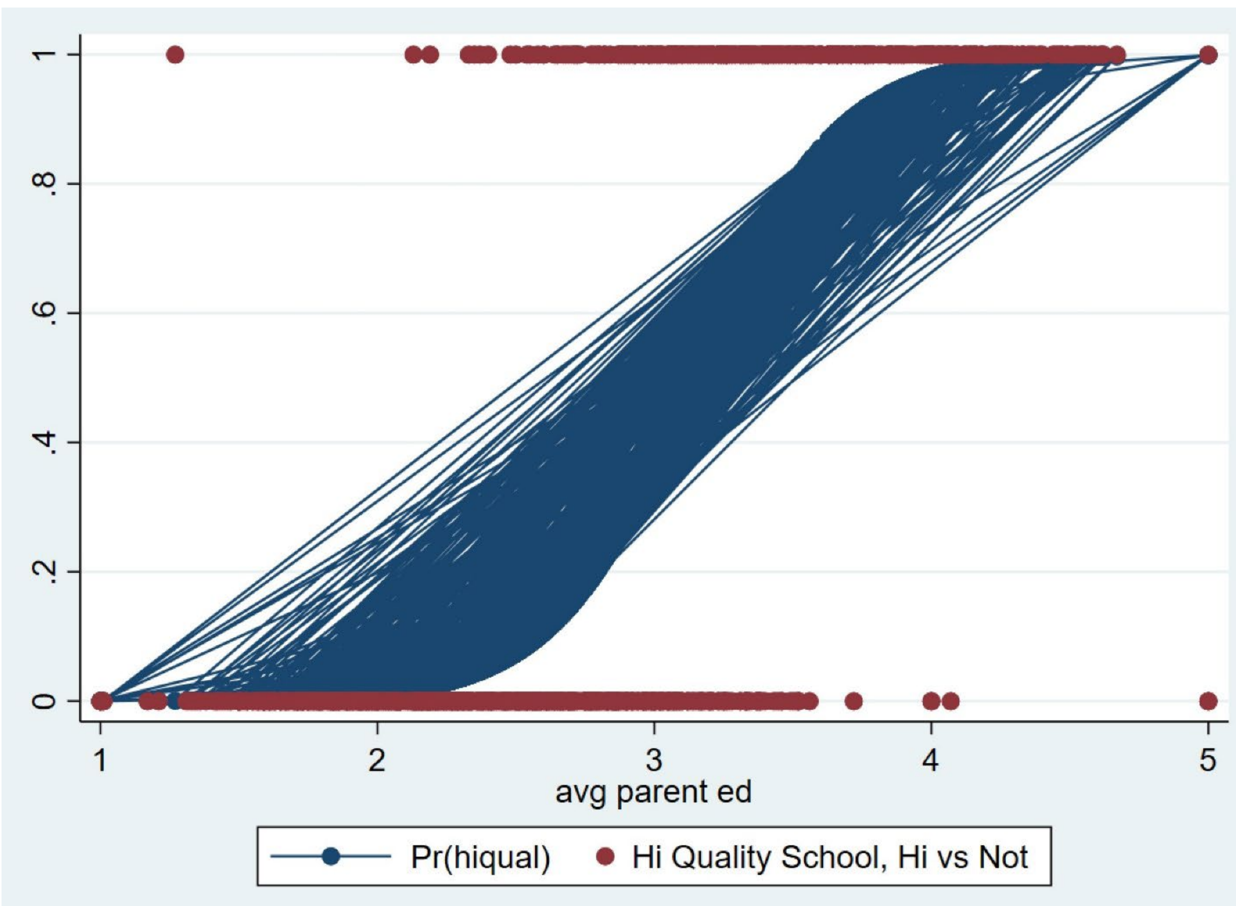
Now, run a logistic regression, obtain the fitted values, and graph them against observed values.

logit hiqual avg_ed
predict y1
twoway scatter y1 hiqual avg_ed, connect(1 .)

Compare the two regression results and graphs. Which type of regression should we use to fit the data – linear or logistic? Why? Is our hypothesis supported?

```
Logistic regression                             Number of obs     =      1,158
                                                LR chi2(1)        =     753.54
                                                Prob > chi2       =     0.0000
Log likelihood = -353.91719                     Pseudo R2         =     0.5156

------------------------------------------------------------------------------
      hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avg_ed |   3.909635   .2383161    16.41   0.000     3.442544    4.376726
       _cons |  -12.30054    .731489   -16.82   0.000    -13.73423   -10.86684
------------------------------------------------------------------------------
```



Logistic regression certainly hits more of the actual points than the linear regression. Although I'd argue that right now, our hypothesis doesn't appear to be correct. At best, we're still not reaching many of the points.

(2) Logistic regression or curvilinear regression?

We need to test the hypothesis that the probability of diabetes increases with age but decreases after a certain stage of life. We will also test whether blacks and males have a higher chance of developing diabetes with age than non-blacks and females.

To test the hypotheses, choose your DV and IVs from the following dataset:

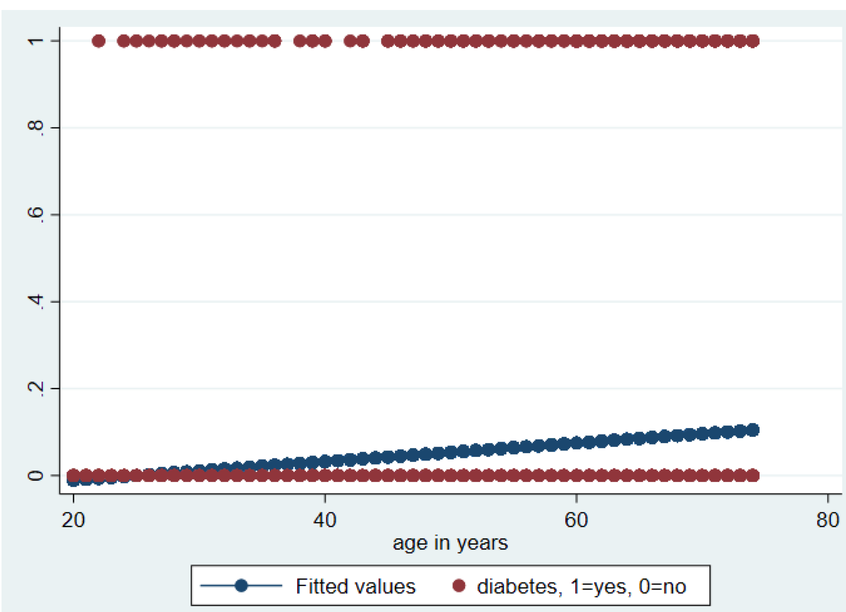
webuse nhanes2f, clear

I'm going to choose age as our IV, and diabetes as our DV.

We're then going to make black males ('black' and 'race') interact for our second test, and do the same thing when we test for non-blacks and females.

Run a curvilinear regression with the DV and IVs (using the **reg** command), obtain the fitted values, and graph them against observed values. Does the line/model fit the data? Why?

reg
predict y1
twoway scatter y1 hiqual avg_ed, connect(1 .)



It does not fit very well at all, at least for people with diabetes.

Now run a logistic model with the same DV and IVs but obtain the fitted values and graph using the following command: **marginscontplot** or **mcp** in short. To use this command, you need to first install it using the following command:

//net install gr0056.pkg

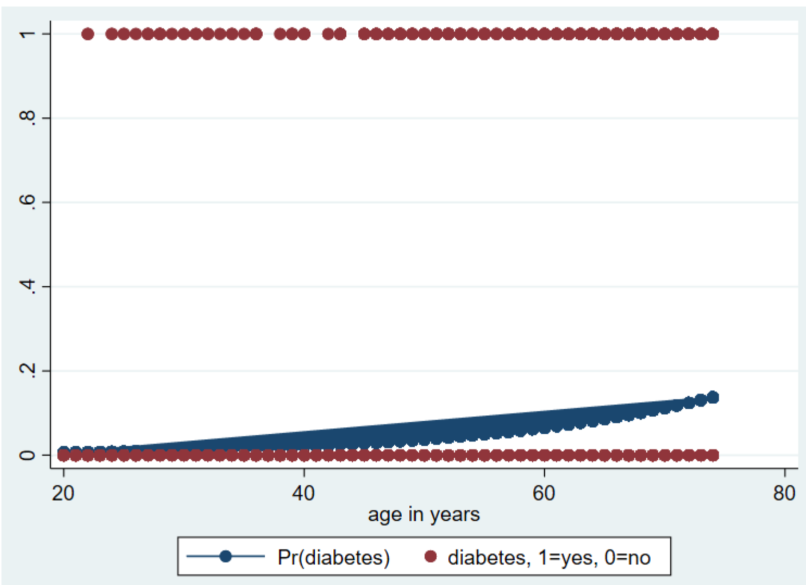
sysdir set PLUS "\\Client\C$\ISF_110"
ssc install mcp

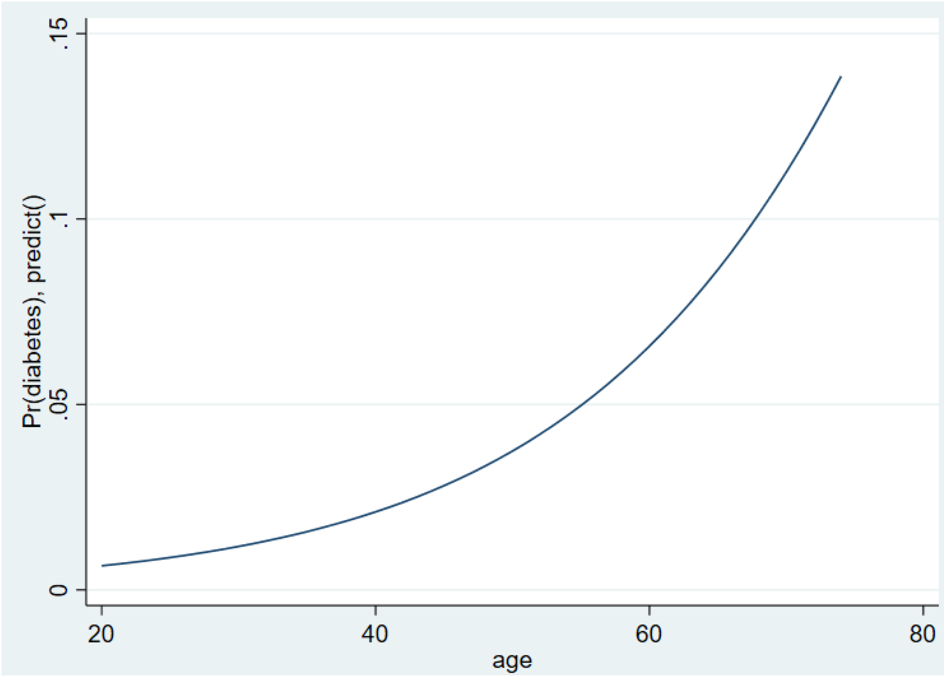
```
package installation

package name:  marginscontplot2.pkg
        from:  http://fmwww.bc.edu/RePEc/bocode/m/

checking marginscontplot2 consistency and verifying not already installed...
installing into \\Client\C$\ISF_110\...
installation complete.

(click here to return to the previous screen)
```





The logistic is a slightly better fit, but still doesn't fit well.

After running the logistic model, run

mcp, ci

Does the line/model fit the data better? Note that the lower confidence interval shows a curvilinear pattern. But our **age** variable has only this range (20-74). Check it using the **sum** command. We can "project" the trend line beyond age 74, say to 100, to see if later ages show a curvilinear patter for diabetes diagnosis. To see this, run the following commands:

logit diabetes black female age c.age#c.age //the last variable is an interaction term for age (with later age).
mcp age, at1(20(1)100) //at1 is an option to project the age range to 20-100 with a 20-year interval.

```
Logistic regression                             Number of obs     =     10,335
                                                LR chi2(4)        =     381.03
                                                Prob > chi2       =     0.0000
Log likelihood = -1808.5522                     Pseudo R2         =     0.0953
```

diabetes	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
black	.7207406	.1266509	5.69	0.000	.4725093	.9689718
female	.1566863	.0942032	1.66	0.096	-.0279486	.3413212
age	.1324622	.0291223	4.55	0.000	.0753836	.1895408
c.age#c.age	-.0007031	.0002753	-2.55	0.011	-.0012428	-.0001635
_cons	-8.14958	.7455986	-10.93	0.000	-9.610926	-6.688233

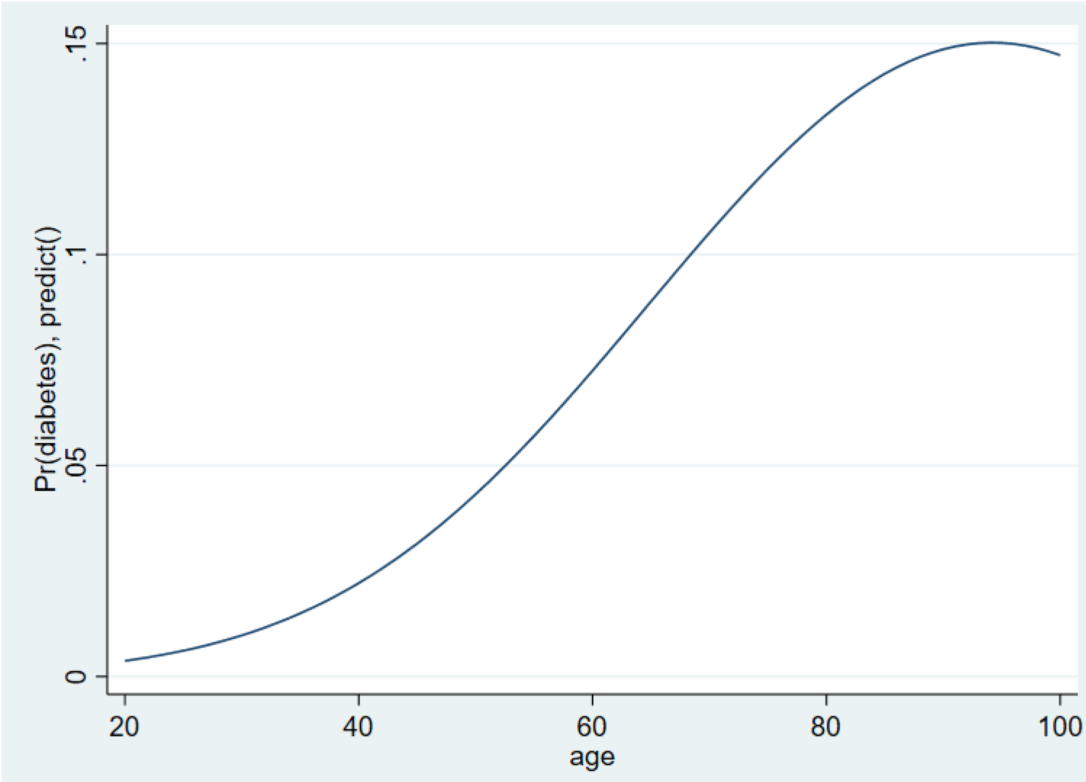
logit diabetes black female age c.age#c.age, or

this is odds in odds col

```
Logistic regression                             Number of obs     =     10,335
                                                LR chi2(4)        =     381.03
                                                Prob > chi2       =     0.0000
Log likelihood = -1808.5522                     Pseudo R2         =     0.0953
```

diabetes	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
black	2.055955	.2603886	5.69	0.000	1.604014	2.635234
female	1.169629	.1101828	1.66	0.096	.9724383	1.406805
age	1.141636	.033247	4.55	0.000	1.078298	1.208694
c.age#c.age	.9992971	.0002751	-2.55	0.011	.998758	.9998365
_cons	.0002889	.0002154	-10.93	0.000	.000067	.0012455

Note: _cons estimates baseline odds.



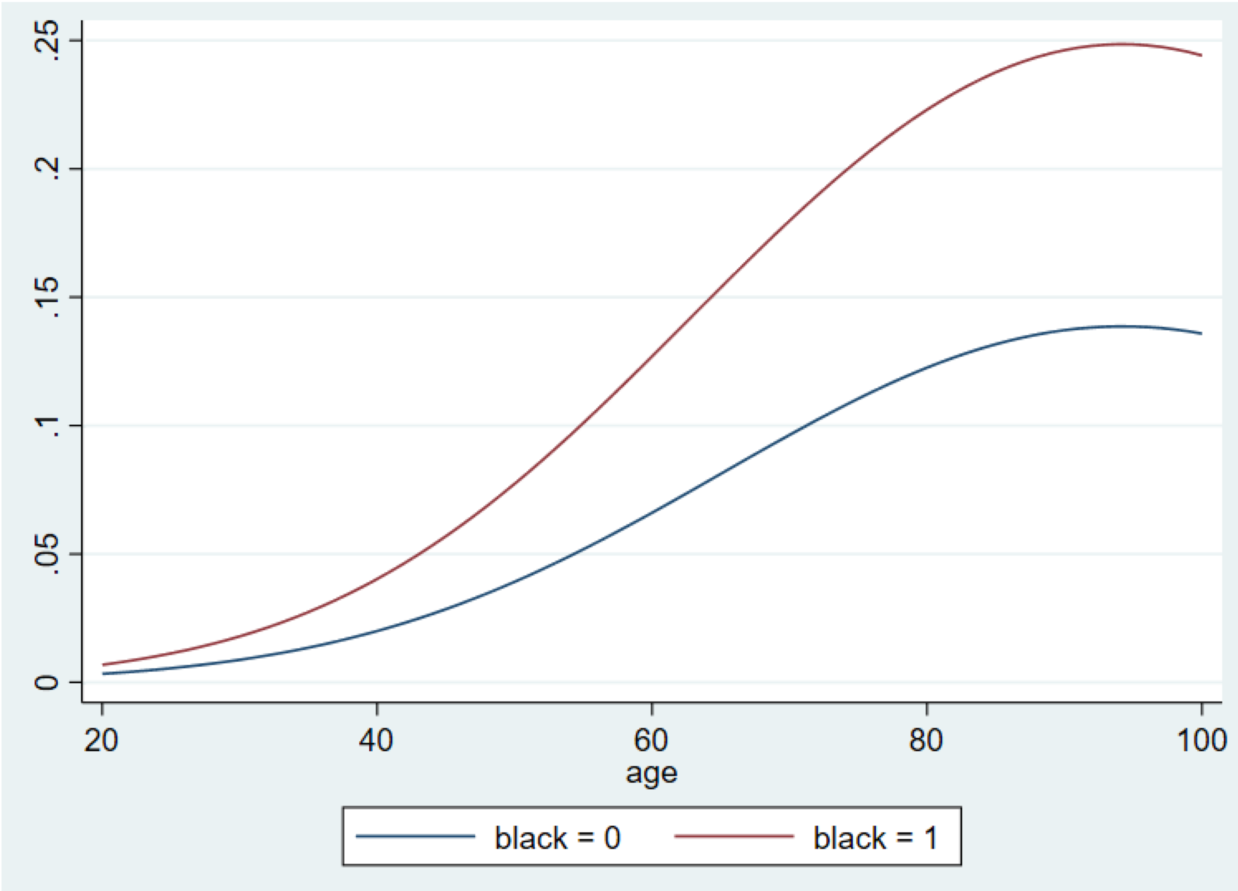
//** prof solution
//margins, dx/dy
//margins plot, at(age = 20 1 (74))

///**
//quietly margins, at(age = 20 1 (74))
//marginsplot, noci

What does the graph say? Compare this graph with the one obtained from curvilinear regression above. Which model to choose – curvilinear or logistic?
The logistic curve fits substantially better than the other curves here. The curvilinear curve is slightly steeper, and also doesn't capture the dropoff in likelihood following age 80. This is good to know. Definitely choosing logistic here.

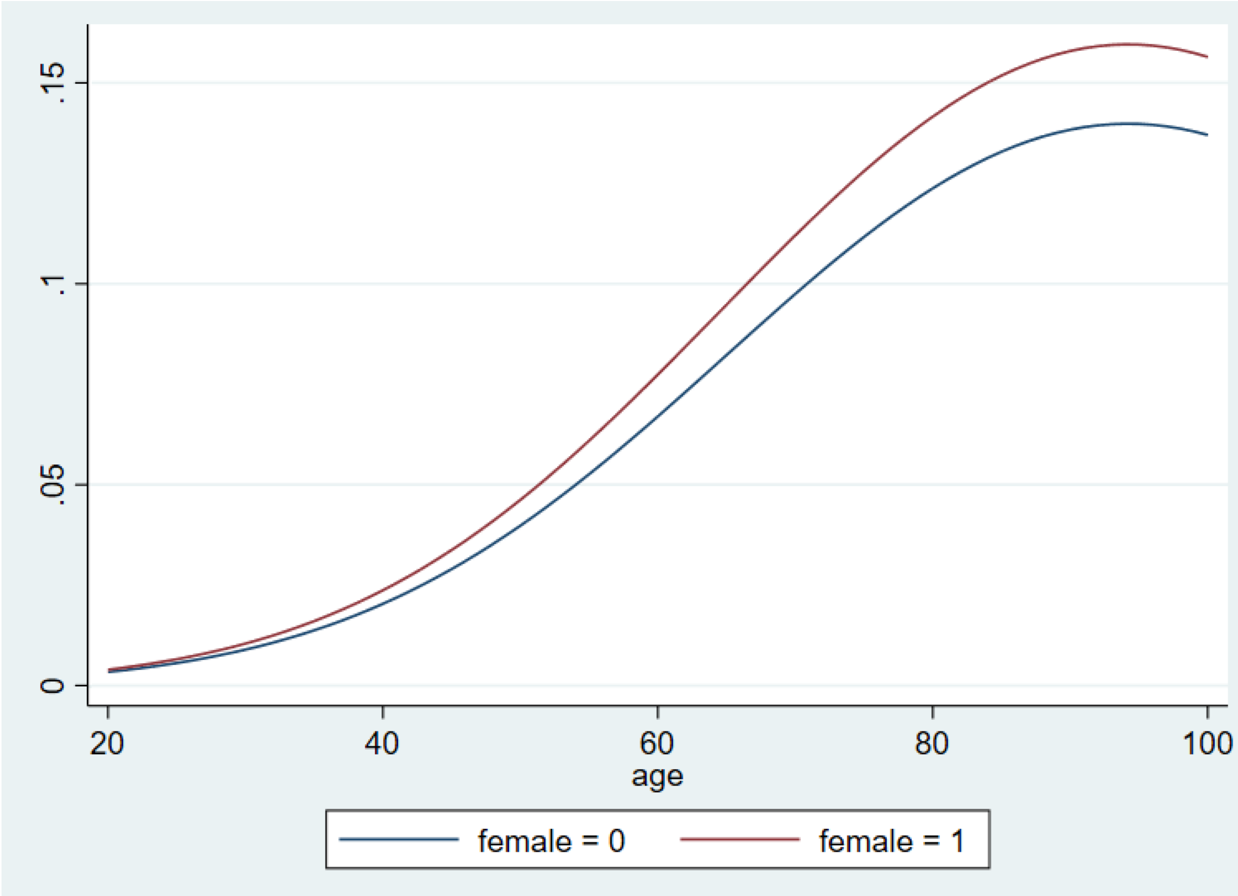
Now, obtain separate graphs for "age and black" and "age and female":

mcp age black



It appears that if you are black, you have a higher likelihood of getting diabetes, particularly as you pass age 50. We can also observe this difference in the pvalue and odds ratio for black in the tables above.

mcp age female



Also similar to the table above, it appears the gap between male female isn't significant. Not only is the p-value not significant, but the odds (displayed here) almost mirror eachother.

Explain the graphs to decide about the hypotheses.

Post your do file as a separate document and all tables, graphs, and interpretations as one word document.

End of Lab