15.8 Let P(y = 1) denote the probability that a randomly selected respondent supports current laws legalizing abortion, estimated using sex of respondent (s= 0, male; s = 1, female), religious affliation (r_1 = 1, Protestant, 0 otherwise; r_2 = 1, Catholic, 0 otherwise; $r_1 = r_2 = 0$, Jewish), and political party affiliation (P_1 = 1, Democrat, 0 otherwise; P_2 = 1, Republican, 0 otherwise, $p_1 = p_2 = 0$, Independent). The logistic model with mam effects has prediction equation

$$\text{logit}\,[P(y=1)] = 0.11 + 0.16s - 0.57\,r_1 - 0.66r_2 + 0.47p_1 - 1.67p_2$$

(a) Give the effect of sex on the odds of supporting legalized abortion; that is, if the odds of support for females equal θ times the odds of support for males, report θ.

- **The logistic model**

$$\text{logit}\,[P(y=1)] = 0.11 + 0.16s - 0.57\,r_1 - 0.66r_2 + 0.47p_1 - 1.67p_2$$
$$s = 1 \text{ female} \quad s = 0 \text{ male}$$

β_1 represents the effect of gender, controlling other variables. Since s = 1 for female, the positive coefficient (0.16) of s means that the estimated odds of supporting legalized abortion are higher for female than male.

- **The effect of sex on the odds**

The antilog of β_1:

$$e^{\beta 1} = e^{0.16} = 1.17 = \theta$$

1.17 is the estimated odds ratio between gender and supporting legalized abortion, controlling other variables.
The estimated odds of the supporting legalized abortion for female equal 1.17 times the estimated odds for male.

- **More**

$$\text{logit}\,[P(y=1)] = 0.11 + 0.16s - 0.57\,r_1 - 0.66r_2 + 0.47p_1 - 1.67p_2$$

The corresponding prediction equation for odds is:

$$Odds = e^{0.11 + 0.16s - 0.57\,r1 - 0.66r2 + 0.47p1 - 1.67p2}$$
$$= e^{0.11}e^{0.16s}e^{-0.57\,r1}\,e^{-0.66r2}\,e^{0.47p1}\,e^{-1.67p2}$$

For female, the estimated odds equal:

$$s = 1$$
$$Odds_1 = e^{0.11}e^{0.16}e^{-0.57\,r1}\,e^{-0.66r2}\,e^{0.47p1}\,e^{-1.67p2}$$

For male, the estimated odds equal:

$$s = 0$$
$$Odds_2 = e^{0.11} e^{-0.57\,r1}\, e^{-0.66r2}\, e^{0.47p1}\, e^{-1.67p2}$$

The estimated odds for female divided by the estimated odds for male equal:

$$Odds_1 / Odds_2 = e^{0.16} = 1.17 = \theta$$

This shows why the antilog of the coefficient for s in the prediction equation is the estimated odds ratio between gender and supporting legalized abortion, for female and male.

(b) Give the effect of being Democrat instead of Independent on the estimated odds of support for legalized abortion.

- **The effect of being Democrat**

For Democrat, the estimated odds equal:

$$p_1 = 1 \quad p_2 = 0$$
$$Odds_1 = e^{0.11} e^{0.16s} e^{-0.57\,r1}\, e^{-0.66r2}\, e^{0.47}$$

- **The effect of being Independent**

For Independent, the estimated odds equal:

$$p_1 = 0 \quad p_2 = 0$$
$$Odds_1 = e^{0.11} e^{0.16s} e^{-0.57\,r1}\, e^{-0.66r2}$$

- **Effects on Odds of being Democrat instead of Independent**

The estimated odds for Democrat divided by the estimated odds for Independent equal:

$$Odds_1 / Odds_2 = e^{0.47`} = 1.59$$

The positive coefficient (0.47) means that the estimated odds of supporting legalized abortion are higher for Democrat than Independent.
The estimated odds of the supporting legalized abortion for Democrat instead of Independent equal 1.59.

(c) Give the effect of being Democrat instead of Republican on the estimated odds of support for legalized abortion.

- **The effect of being Democrat**

For Democrat, the estimated odds equal:

$$Odds_1 = e^{0.11} e^{0.16s} e^{-0.57\,r1}\, e^{-0.66r2}\, e^{0.47}$$

- **The effect of being Republication**

For Independent, the estimated odds equal:

$$p_1 = 0 \quad p_2 = 1$$
$$Odds_1 = e^{0.11} e^{0.16s} e^{-0.57\,r1}\ e^{-0.66r2}\ e^{-1.67}$$

- **Effects on Odds of being Democrat instead of Republication**

The estimated odds for Democrat divided by the estimated odds for Republication equal:

$$Odds_1 / Odds_2 = e^{-1.67} = 0.19$$

The negative coefficient (-1.67) means that the estimated odds of supporting legalized abortion are lower for Democrat than Republication.
Therefore, the estimated odds of the supporting legalized abortion for Democrat instead of Republication equal 0.19.

(d) Find the estimated probability of supporting legalized abortion, for (i) female Jewish Democrats, (ii) male Catholic Republicans.

- **Formula for the estimated probability of supporting legalized abortion**

$$logit\ [P(y=1)] = 0.11 + 0.16s - 0.57\ r_1 - 0.66r_2 + 0.47p_1 - 1.67p_2$$

$$\hat{P}(y=1) = \frac{e^{0.11 + 0.16s - 0.57\,r1 - 0.66r2 + 0.47p1 - 1.67p2}}{1 + e^{0.11 + 0.16s - 0.57\,r1 - 0.66r2 + 0.47p1 - 1.67p2}}$$

- **Female Jewish Democrats**

$$s = 1; \quad r_1 = r_2 = 0; \quad p_1 = 1 \quad p_2 = 0$$

$$\hat{P}(y=1) = \frac{e^{0.11 + 0.16 + 0.47}}{1 + e^{0.11 + 0.16 + 0.47}}$$

$$P = odds / (1+odds) = 0.67$$

The estimated probability of supporting legalized abortion for female Jewish Democrats is 0.67.

- **Male Catholic Republicans**

$$s = 0; \quad r_1 = 0 \quad r_2 = 1; \quad p_1 = 0 \quad p_2 = 1$$

$$\hat{P}(y=1) = \frac{e^{0.11 - 0.66 - 1.67}}{1 + e^{0.11 - 0.66 - 1.67}}$$

$$P = odds / (1+odds) = 0.098$$

The estimated probability of supporting legalized abortion for Male Catholic Republicans is 0.098.

Using the data at http://teaching.sociology.ul.ie/so5032/hten.dta, construct a binary variable indicating home ownership (i.e., "own outright" and "own with mortgage" versus the rest). Explore the data, determining what are the main characteristics that predict home ownership. Come up with a logistic regression model that includes all the variables you think are relevant, and write a (very) short report. Include the minimal Stata code that you used.

- **Binary Variable (Tenure)**

```
. tab tenure

               housing tenure |      Freq.     Percent        Cum.
------------------------------+-----------------------------------
               owned outright |      4,701       30.33       30.33
          owned with mortgage |      7,147       46.11       76.44
       local authority rented |      1,695       10.94       87.37
         housing assoc. rented |        653        4.21       91.59
          rented from employer |         99        0.64       92.23
rented private unfurnished    |        697        4.50       96.72
rented private furnished      |        454        2.93       99.65
               other rented   |         54        0.35      100.00
------------------------------+-----------------------------------
                        Total |     15,500      100.00

. tab tenure, nol

    housing |
     tenure |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      4,701       30.33       30.33
          2 |      7,147       46.11       76.44
          3 |      1,695       10.94       87.37
          4 |        653        4.21       91.59
          5 |         99        0.64       92.23
          6 |        697        4.50       96.72
          7 |        454        2.93       99.65
          8 |         54        0.35      100.00
------------+-----------------------------------
      Total |     15,500      100.00
```

Table 1: There are eight categories of Tenure.

Recoding Tenure as a binary variable for logistic regression:

```
. label list otenure
otenure:
          -9 missing
          -7 telephone int. only
           1 owned outright
           2 owned with mortgage
           3 local authority rented
           4 housing assoc. rented
           5 rented from employer
           6 rented private unfurnished
           7 rented private furnished
           8 other rented

. gen owner = inlist(tenure, 1,2)
. tab owner

      owner |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,779       24.18       24.18
          1 |     11,848       75.82      100.00
------------+-----------------------------------
      Total |     15,627      100.00
```

Table 2: Divide Tenure into two categories.

Dealing with missing values:

```
. replace owner = . if missing(tenure)
(127 real changes made, 127 to missing)
```

Table 3: Delete missing values.

Generate a binary variable of new Tenure:

```
. tab own2

  RECODE of |
     tenure |
   (housing |
   tenure   ) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,652       23.56       23.56
          1 |     11,848       76.44      100.00
------------+-----------------------------------
      Total |     15,500      100.00

. label define own2 0 "rest" 1 "owned outright & owned with mortgage"

. label values own2 own2

. tab own2

 RECODE of tenure (housing tenure  ) |      Freq.     Percent        Cum.
-------------------------------------+-----------------------------------
                                rest |      3,652       23.56       23.56
owned outright & owned with mortgage |     11,848       76.44      100.00
-------------------------------------+-----------------------------------
                               Total |     15,500      100.00
```

Table 4: Create binary variable Own2 as a binary variable of Tenure.

- **Logistic Regression with single explanatory variable**
- **Backward Elimination**

Placing all of the predictors under consideration in this model. It seems all variables make significant partial contributions to predicting y, as p-value of all variables are lower than 0.01.

```
. stepwise, pr(.01): logistic own2 sex jbstat mastat age nchild qfedhi spjb educ class
                      begin with full model
p < 0.0100            for all terms in model

Logistic regression                               Number of obs   =      10891
                                                  LR chi2(9)      =    1298.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -5249.5106                       Pseudo R2       =     0.1100

------------------------------------------------------------------------------
        own2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .8477247   .0436104    -3.21   0.001     .7664178    .9376572
      jbstat |    .893303   .0112874    -8.93   0.000     .8714517    .9157021
      mastat |   .7267797   .0164553   -14.10   0.000     .6952331    .7597578
         age |   1.008972   .0016746     5.38   0.000     1.005695     1.01226
      nchild |   .8347827   .0256499    -5.88   0.000     .7859936    .8866003
      qfedhi |   .8602635   .0132587    -9.77   0.000     .8346655    .8866465
        spjb |    1.29579   .0716389     4.69   0.000      1.16272     1.44409
        educ |   1.202879   .0642435     3.46   0.001      1.08333     1.33562
       class |    .822211   .0141739   -11.36   0.000     .7948948     .850466
------------------------------------------------------------------------------
```

Table 5: step regression for all variables.

Looking at details, educational situation can be presented by both ‘educ’ and ‘qfedhi’, and both ‘jbstat’ and ‘class’ can analyze occupational situation. Thus, considering z-value, this project will make analysis about age, mastat, qfedhi and class varibales. In other words, it is assumed that the probability of having home ownership can be impacted by those four variables.

```
. des
Contains data from /Users/wang/Downloads/hten-4.dta
  obs:        15,627
 vars:            13                          22 Mar 2011 15:24
 size:       453,183 (99.9% of memory free)
-----------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------------
hid             long   %12.0g                 household identification number
sex             byte   %8.0g       osex       sex
jbstat          byte   %8.0g       ojbstat    current economic activity
mastat          byte   %8.0g       omastat    marital status
age             byte   %8.0g       oage       age at date of interview
nchild          byte   %8.0g       onchild    number of own children in household
qfedhi          byte   %8.0g       oqfedhi    highest educational qualification
spjb            byte   %23.0g      ospjb      whether spouse/partner employed now
tenure          byte   %8.0g       otenure    housing tenure
educ            float  %9.0g       ed3        Highest Educational Level
class           float  %26.0g      class      Social Class (Goldthorpe scheme)
owner           float  %9.0g
own2            byte   %36.0g      own2       RECODE of tenure (housing tenure )
-----------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

Table 6: Looking at details of all variables.

➢ Quadratic Regression Models for Age

```
. logit own2 age

Iteration 0:   log likelihood = -8462.2668
Iteration 1:   log likelihood = -8396.1105
Iteration 2:   log likelihood = -8395.8593
Iteration 3:   log likelihood = -8395.8593

Logistic regression                               Number of obs   =      15499
                                                  LR chi2(1)      =     132.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -8395.8593                       Pseudo R2       =     0.0078

------------------------------------------------------------------------------
        own2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0119102   .0010456    11.39   0.000     .0098608    .0139596
       _cons |   .6421907   .0495884    12.95   0.000     .5449992    .7393822
------------------------------------------------------------------------------
```

Table 7: Logistic regression of age and own2, this model cannot make a well explanation about the relation between age and home ownership.

Making a quadratic regression model of age and comparing to the logistic regression. In table 8, Pseudo $R^2 = 0.023 > r^2 = 0.078$. Also, the chi-square value is much higher in table 8 than table 7 (393 > 132). The quadratic regression model is more effective for age variable:

```
. logit own2 c.age##c.age

Iteration 0:   log likelihood = -8462.2668
Iteration 1:   log likelihood = -8268.4968
Iteration 2:   log likelihood = -8265.6118
Iteration 3:   log likelihood = -8265.6116

Logistic regression                               Number of obs   =      15499
                                                  LR chi2(2)      =     393.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -8265.6116                       Pseudo R2       =     0.0232

------------------------------------------------------------------------------
        own2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0902015   .0049133    18.36   0.000     .0805715    .0998315
             |
 c.age#c.age |  -.0008169   .0000498   -16.42   0.000    -.0009145   -.0007194
             |
       _cons |  -.9270015   .1072065    -8.65   0.000    -1.137122   -.7168807
------------------------------------------------------------------------------
```

Table 8: Graph of Two Second-Degree Polynomials.

In table 8, Mound-shaped function have $\beta_2 < 0$ (-0.00082). Also, since the coefficient 0.9 of x is positive, the curve is increasing as it crosses the y-axis. A mound-shaped quadratic equation takes its minimum at $x = -\beta_1/(2\beta_2) = -0.09/(2*(-0.00082) = 55$. The predicted probability of having home owner increases as age increases under 55 years old, but decrease after 55 years old. As showing below:

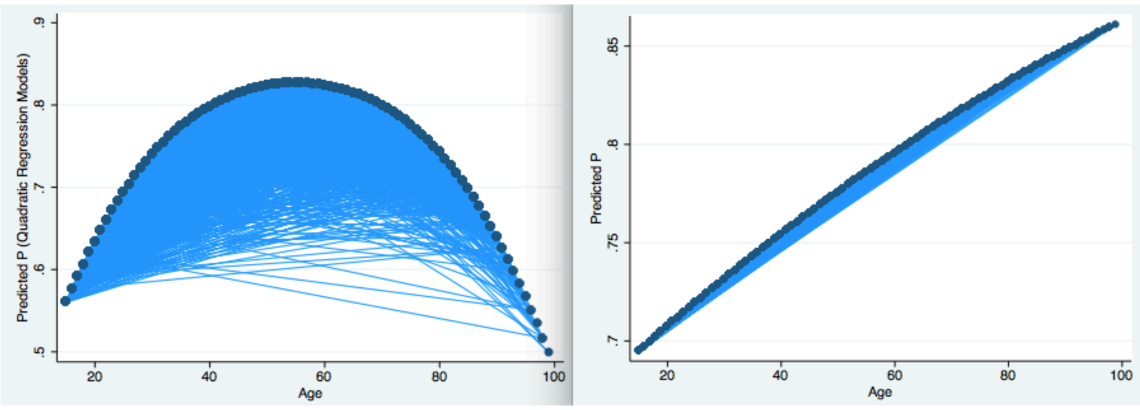

Figure 1: The Quadratic Regression Models makes a better explanation for age variable.

➢ **Logit and Recode Mastat/Qfedhi/Class variables**

Take Mastat as an example:

```
tab mastat

     marital status  |      Freq.     Percent        Cum.
---------------------+-----------------------------------
      child under 16 |         50        0.32        0.32
             married |      8,106       51.89       52.21
living as couple     |      1,838       11.77       63.97
             widowed |      1,152        7.37       71.35
      divorced       |        858        5.49       76.84
      separated      |        278        1.78       78.62
      never married  |      3,340       21.38      100.00
---------------------+-----------------------------------
               Total |     15,622      100.00

. tab mastat, nol

    marital |
     status |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         50        0.32        0.32
          1 |      8,106       51.89       52.21
          2 |      1,838       11.77       63.97
          3 |      1,152        7.37       71.35
          4 |        858        5.49       76.84
          5 |        278        1.78       78.62
          6 |      3,340       21.38      100.00
------------+-----------------------------------
      Total |     15,622      100.00
```

Table 9: tab mastat.

In logistic Regression, recoding mastat is necessary, and it must to be divided into two categories:

```
. recode mastat 0=6
(mastat: 50 changes made)

recode mastat 1/2=1 3/6=0, gen (nmastat)
(7516 differences between mastat and nmastat)
```

Table 10: Recode mastat.

In nmastat variable, the first category involves married and living as couple (=1), and the second includes widowed, divorced, separated and never married (=0), naming otherwise. The first category (1) represents a relatively complete family:

```
. tab nmastat, nol

  RECODE of |
     mastat |
   (marital |
   status  ) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,678       36.35       36.35
          1 |      9,944       63.65      100.00
------------+-----------------------------------
      Total |     15,622      100.00

. tab nmastat

 RECODE of mastat (marital |
                  status  ) |      Freq.     Percent        Cum.
---------------------------+-----------------------------------
                 otherwise |      5,678       36.35       36.35
married & living as couple |      9,944       63.65      100.00
---------------------------+-----------------------------------
                     Total |     15,622      100.00
```

Table 11: nmastat.

Making a logistic regression for own2 and nmastat. As shown in table 12:

- Firstly, the formula equal:

$$logit\ [p(y=1)] = a + \beta x$$
$$logit\ [p(y=1)] = 0.61 + 0.97(nmastat)$$

- Then, the antilog of β_1:

$$e^{\beta 1} = e^{0.96} = 2.63$$

- The estimated odds of having home ownership for people with complete family equal 1.17 times for people with incomplete family.
- The probability with a greater $x^2 = 638$, with 2 degree freedom, is low enough = 0.0000 to reject null hypothesis, indicating nmastat indeed has an effect for predicted p.
- Also, with one predictor variable, that predictor's z statistic and overall x^2 statistic test equivalent hypotheses:

```
. logit own2 nmastat

Iteration 0:   log likelihood = -8460.2839
Iteration 1:   log likelihood = -8146.4054
Iteration 2:   log likelihood = -8140.9333
Iteration 3:   log likelihood = -8140.9325

Logistic regression                               Number of obs   =      15496
                                                  LR chi2(1)      =     638.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -8140.9325                       Pseudo R2       =     0.0377

------------------------------------------------------------------------------
        own2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nmastat |   .9759509   .0387291    25.20   0.000     .9000432    1.051859
       _cons |   .6102845   .0280026    21.79   0.000     .5554005    .6651686
------------------------------------------------------------------------------

. dis exp(_b[nmastat])
2.6536895
```

Table 12: Logistic regression for own2 and nmastat.

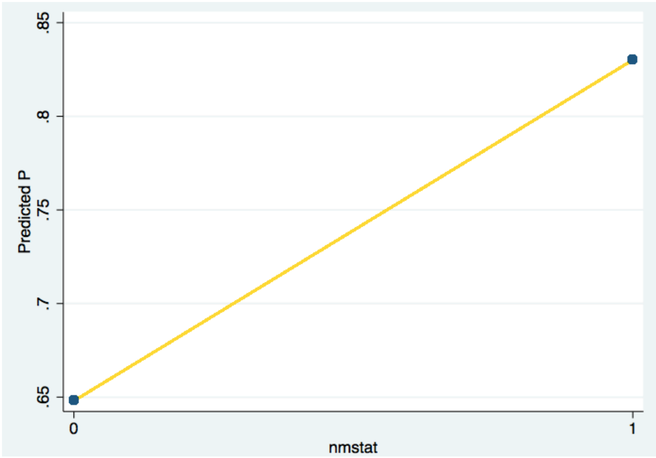

Figure 2: Logistic regression for own2 and nmastat. As can be seen, The probability for nmastat = 1 is higher than nmstat = 0, and queal 2.63 times.

In this sense, recoding class, qfedhi as well:

. tab nclass

RECODE of class (Social Class (Goldthorpe scheme))	Freq.	Percent	Cum.
otherwise	4,062	34.12	34.12
managerial & routine non manual & self-	7,843	65.88	100.00
Total	11,905	100.00	

. tab nclass

RECODE of class (Social Class (Goldthorpe scheme))	Freq.	Percent	Cum.
0	4,062	34.12	34.12
1	7,843	65.88	100.00
Total	11,905	100.00	

Table 13: Dividing class into two categories, that is, category of higher social class and others.

. tab nqfedhi

RECODE of qfedhi (highest educational qualification)	Freq.	Percent	Cum.
otherwise	6,554	45.76	45.76
higher educational qualification	7,768	54.24	100.00
Total	14,322	100.00	

. tab nqfedhi

RECODE of qfedhi (highest educational qualificati on)	Freq.	Percent	Cum.
0	6,554	45.76	45.76
1	7,768	54.24	100.00
Total	14,322	100.00	

Table 13: Dividing qfedhi into two categories, that is, category of higher educational qualification and others.

➢ **Logistic Regression**

Now, making a logistic regression for those four variables:

```
. logit own2 nclass nqfedhi nmastat c.age##c.age

Iteration 0:   log likelihood = -5899.3944
Iteration 1:   log likelihood = -5410.5733
Iteration 2:   log likelihood = -5394.9472
Iteration 3:   log likelihood = -5394.9296
Iteration 4:   log likelihood = -5394.9296

Logistic regression                               Number of obs   =      10894
                                                  LR chi2(5)      =    1008.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -5394.9296                       Pseudo R2       =     0.0855

------------------------------------------------------------------------------
        own2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nclass |   .6561952   .0499964    13.12   0.000     .5582041    .7541864
     nqfedhi |   .6360862   .0523193    12.16   0.000     .5335423    .7386301
     nmastat |   .7851561   .0529116    14.84   0.000     .6814513     .888861
         age |   .0468735   .0069804     6.72   0.000     .0331923    .0605548
             |
 c.age#c.age |  -.0003305   .0000701    -4.72   0.000    -.0004679   -.0001931
             |
       _cons |  -1.352842   .1482633    -9.12   0.000    -1.643433   -1.062251
------------------------------------------------------------------------------
```

Table 14: Logistic Regression.

The logistic regression model equal:

$$logit\,[P(y=1)] = -1.4 + 0.66nc + 0.64nq + 0.79nm + 0.47age - 0.0003age^2$$

$$\hat{P}(y=1) = \frac{e^{-1.4 + 0.66nc + 0.64nq + 0.79nm + 0.47age - 0.0003age^2}}{1 + e^{-1.4 + 0.66nc + 0.64nq + 0.79nm + 0.47age - 0.0003age^2}}$$

The corresponding prediction equation for odds is:

$$Odds = e^{-1.4 + 0.66nc\ 0.64nq\ 0.79nm\ 0.47age - 0.0003age2}$$
$$= e^{0.11} e^{0.66nc} e^{0.64nq} e^{0.79nm} e^{0.47age} e^{-0.0003age2}$$

As $e^{\beta 1} = e^{0.66} = 1.93$. The estimated odds of having home ownership for higher class people (nclass =1) equal 1.93 times for lower class people (nclass =0).

As $e^{\beta 2} = e^{0.64} = 1.89$. The estimated odds of having home ownership for high-education level people (nqfedhi =1) equal 1.89 times for low-education level class people (nqfedhi =0).

For example, for people with lower class, lower educational qualification, incomplete family, the probability of having home ownership is:

$$nclass = 0 \quad nqfedhi = 0 \quad nclass = 0$$
$$Odds_1 = e^{0.11}\, e^{0.47age}\, e^{-0.0003age2}$$

and for people with higher class, higher educational qualification, complete family, the probability of having home ownership is:

$$nclass = 1 \quad nqfedhi = 1 \quad nclass = 1$$

$$Odds_2 = e^{0.11} e^{0.66} e^{0.64}\ e^{0.79}\ e^{0.47age}\ e^{-0.0003age2}$$

$$Odds_1 / Odds_2 = e^{0.66} e^{0.64}\ e^{0.79} = 1.40$$

The estimated odds of the having home ownership for people in higher qualification (education, family and job) equal 1.4 times with people in lower qualification (education, family and job).

This, in fact, corresponds to the previous assumption. The higher educational qualification, the higher probability of having home ownership, which are same to family and class situation. But the age variable should consider carefully, as it is a Quadratic Regression Model.