ECN702 - Assignment 2

2023-02-19

ECN702 - Econometrics II
Professor Hyunhak Kim
Name: Ha Doan
Student ID: 501026801

Question 1:

a.

Create the variable deny by checking whether denial_reason_1 is empty. If the entry is empty, the mortgage is approved. Create the variable LTI by dividing loan_amount_000s by applicant_income_000s. This represents the applicant's loan to income.

```
HMDA2017$deny <- as.numeric(is.na(HMDA2017$denial_reason_1))
HMDA2017$lti = HMDA2017$loan_amount_000s/HMDA2017$applicant_income_000s
```

b.

Regress deny on LTI using a linear probability model, probit and logit regression.

1. **Linear Probability Model**

```
denymod1 <- lm(deny ~ lti, data = HMDA2017)
coeftest(denymod1, vcov. = vcovHC, type = "HC1")
```

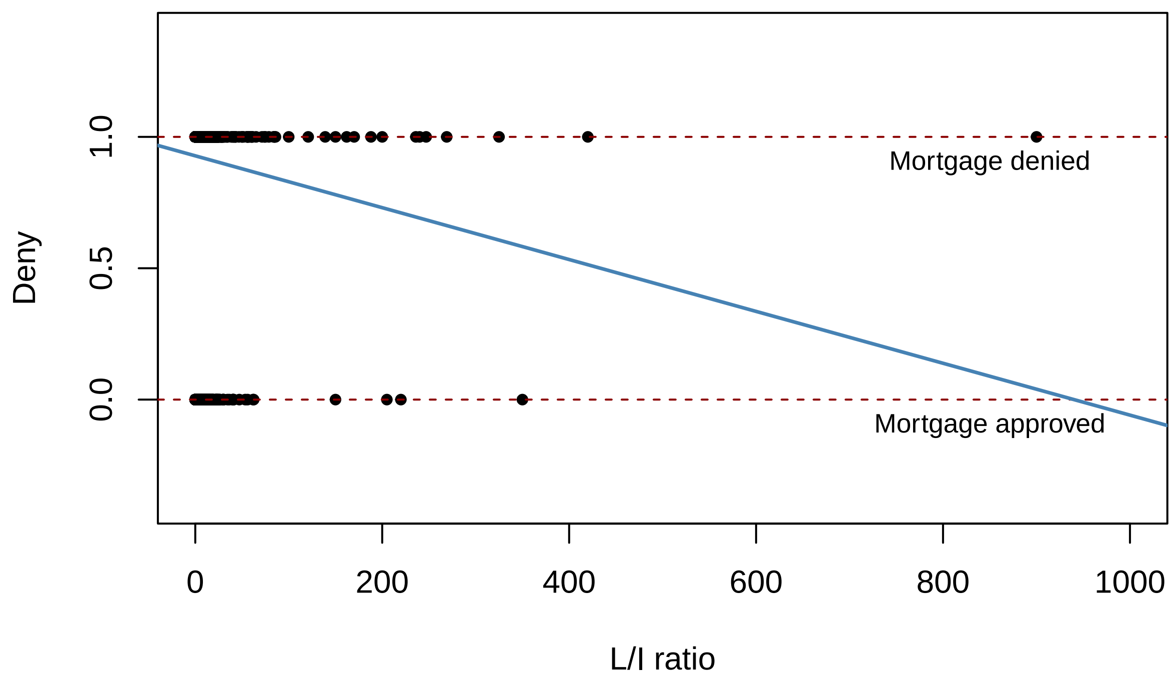
```
##
## t test of coefficients:
##
##                 Estimate  Std. Error  t value Pr(>|t|)
## (Intercept)  0.92791874  0.00173887 533.6335  < 2e-16 ***
## lti         -0.00098699  0.00047146  -2.0935  0.03631 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated regression line for the **Linear Probability Model** is
$\widehat{deny} = \underset{(0.002)}{0.928} - \underset{(0.0004)}{0.001}\ L/IRatio$

The coefficient for $L/Iratio$ is significant at the 0.01 level with a negative relationship with the dependent variable, $\widehat{deny}$. The model implies that an unit increase in $L/Iratio$ leads to a $0.00098699 \approx 0.1$ decrease in the probability of a loan denial.

The plot for the **Linear Probability Model** is shown as the figure below

Figure 1.1 Scatterplot Mortgage Application Denial and the Loan−to−Income Ratio



2. Probit Model

```
denyprobit1 <- glm(deny ~ lti,
                   family = binomial(link = "probit"),
                   data = HMDA2017)
coeftest(denyprobit1, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  1.4507663  0.0120924 119.9739   <2e-16 ***
## lti         -0.0038459  0.0031761  -1.2109   0.2259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated model obtained from the Z-test for the **Probit Model** is
$P(deny|\widehat{L/Iratio}) = \Phi(\underset{(0.012)}{1.451} - \underset{(0.003)}{0.004} L/Iratio)$

The **Probit Model** still shows a negative relationship between $L/Iratio$ and the probability of a loan denial (As $\beta_1 < 0$). We use `predict()` to compute the predicted change in the denial probability when $L/Iratio$ changed from 0.3 to 0.4. It shows a decreasing effect, but the effect is nearly 0.

```r
predictions <- predict(denyprobit1,
                       newdata = data.frame("lti" = c(0.3,0.4)),
                       type = "response")
diff(predictions)
```

```
##             2
## -5.366841e-05
```

3. **Logit Model**

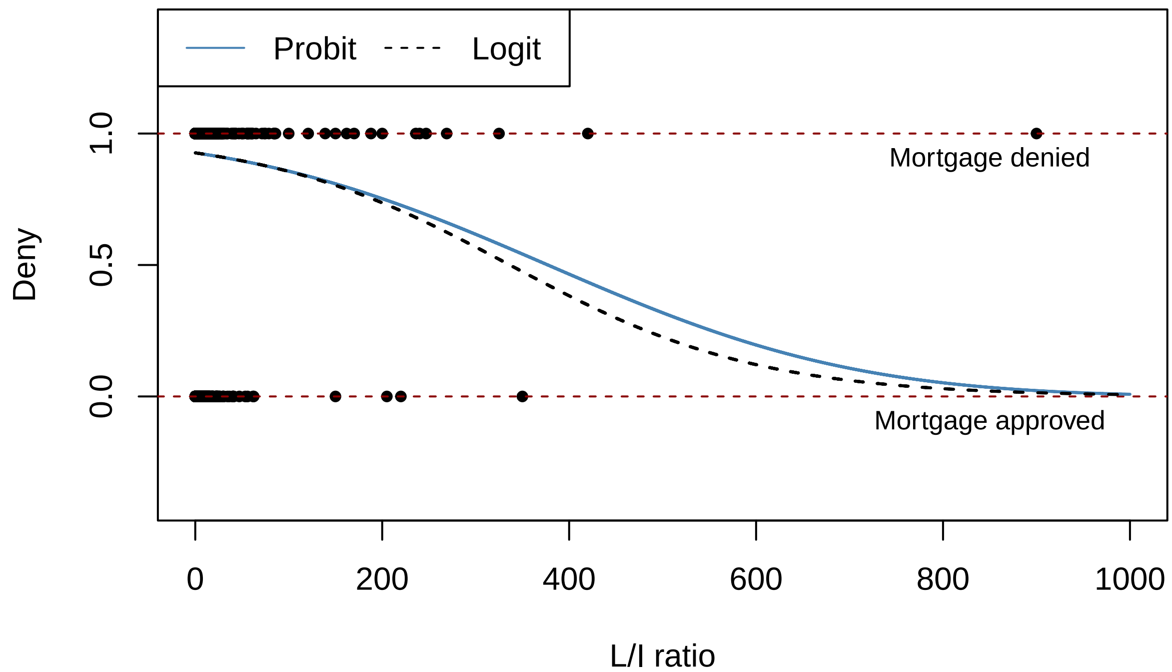
```r
denylogit1 <- glm(deny ~ lti,
                  family = binomial(link = "logit"),
                  data = HMDA2017)
coeftest(denylogit1, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  2.5340766  0.0193846 130.7261  < 2e-16 ***
## lti         -0.0075287  0.0042194  -1.7843  0.07437 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated model obtained from the Z-test for the **Logit Model** is
$P(deny\widehat{|L/I}ratio) = F(\underset{(0.019)}{2.534} - \underset{(0.004)}{0.008}L/Iratio)$

The plot for the **Probit Model** and **Logit Model** is shown as the figure below:

Figure 1.2 Probit and Logit Models of the Probability of Denial, Given L/I Ratio



The figure 1.2 shows that both **Probit** and **Logit** models produce very similar estimates of the probability that a mortgage application will be denied depending on the applicants loan-to-income ratio.

Comparison Models

The summary result for *Question 1b* is shown as below:

```
b <- list(denymod1, denyprobit1, denylogit1)
stargazer(b, type = "text", title = "Table 1: Summary Statistics", no.space = TRUE)
```

```
##
## Table 1: Summary Statistics
## ======================================================================
##                                     Dependent variable:
##                      -------------------------------------------------
##                                            deny
##                                  OLS                probit   logistic
##                                  (1)                  (2)       (3)
## ----------------------------------------------------------------------
## lti                           -0.001***          -0.004***  -0.008***
##                               (0.0002)            (0.001)    (0.002)
## Constant                      0.928***           1.451***   2.534***
##                                (0.001)            (0.008)    (0.016)
## ----------------------------------------------------------------------
## Observations                   72,734              72,734     72,734
## R2                             0.0004
```

```
## Adjusted R2                 0.0004
## Log Likelihood                                 -19,396.230 -19,397.240
## Akaike Inf. Crit.                              38,796.460  38,798.480
## Residual Std. Error    0.264 (df = 72732)
## F Statistic         29.555*** (df = 1; 72732)
## ========================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

From the table, we know that $L/Iratio$ is statistically significant in changing the probability of the mortgage denial at the 1% significance level in all models. Additionally, the *linear probability* model cannot capture the nonlinear nature of the population regression function and it may predict probabilities to lie outside the interval $[0; 1]$. *Probit* and *Logit* models are harder to interpret but capture the nonlinearities better since both models produce predictions of probabilities that lie inside the interval $[0; 1]$.

c.

Create the subsample which include only African American(3) and White(5) applicant based on applicant_race_1.

```r
# Check which entries are 3 and 5
W1 <- which(HMDA2017$applicant_race_1==3|HMDA2017$applicant_race_1==5)
# Assign the dataset
HMDA2017_AAW <- HMDA2017[W1,]
```

Then repeat (b). Does the result change? Why or why not?

1. **Linear Probability Model**

```r
denymod2 <- lm(deny ~ lti, data = HMDA2017_AAW)
coeftest(denymod2, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error  t value Pr(>|t|)
## (Intercept)  0.93205181  0.00191769 486.0277  < 2e-16 ***
## lti         -0.00091107  0.00052483  -1.7359  0.08258 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated regression line for the **Linear Probability Model** is

$$\widehat{deny} = \underset{(0.002)}{0.932} - \underset{(0.0004)}{0.001}\, L/IRatio$$

2. **Probit Model**

```r
denyprobit2 <- glm(deny ~ lti,
                   family = binomial(link = "probit"),
                   data = HMDA2017_AAW)
```

```
## Warning: glm.fit: algorithm did not converge
```

```r
coeftest(denyprobit2, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  1.4813580  0.0120704 122.7268   <2e-16 ***
## lti         -0.0035205  0.0029490  -1.1938   0.2326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated model obtained from the Z-test for the **Probit Model** is
$P(\widehat{deny|L/I}ratio) = \Phi(\underset{(0.012)}{1.481} - \underset{(0.003)}{0.004}L/Iratio)$

3. Logit Model

```r
denylogit2 <- glm(deny ~ lti,
                 family = binomial(link = "logit"),
                 data = HMDA2017_AAW)
coeftest(denylogit2, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  2.5974266  0.0249565 104.0783   <2e-16 ***
## lti         -0.0070900  0.0060221  -1.1773   0.2391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated model obtained from the Z-test for the **Logit Model** is
$P(\widehat{deny|L/I}ratio) = F(\underset{(0.018)}{2.597} - \underset{(0.006)}{0.007}L/Iratio)$

Comparison Models

The summary result for *Question 1c* is shown as below:

```r
c <- list(denymod2, denyprobit2, denylogit2)
stargazer(c, type = "text", title = "Table 2: Summary Statistics", no.space = TRUE)
```

```
##
## Table 2: Summary Statistics
## =======================================================================
##                                   Dependent variable:
##                     ---------------------------------------------------
##                                          deny
##                               OLS              probit     logistic
##                               (1)                (2)         (3)
## -----------------------------------------------------------------------
## lti                        -0.001***          -0.004***   -0.007***
##                            (0.0002)            (0.001)     (0.002)
```

```
## Constant                     0.932***          1.481***    2.597***
##                               (0.001)           (0.009)     (0.018)
## ----------------------------------------------------------------------
## Observations                 56,619            56,619      56,619
## R2                           0.0004
## Adjusted R2                  0.0004
## Log Likelihood                               -14,467.000 -14,467.990
## Akaike Inf. Crit.                            28,938.010  28,939.990
## Residual Std. Error    0.256 (df = 56617)
## F Statistic         21.500*** (df = 1; 56617)
## ======================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

Both questions b and c generate very similar result, which indicates that $L/Iratio$ has a negative effect on the probability of mortgage denial. It means that if the loan-to-income ratio increase, the probability of denying the mortgage will decrease.
The results from question b and c remains unchanged because the only difference is the sample size. The sample in question c only counts for White race and American African, exclude other races. If the result remains similar to question b, it implies that American African and White are greatly represented in both question b and c.

d.

Regress deny on LTI, applicant_race_1, loan_type, property_type, loan_purpose, owner_occupancy, sex, income using linear probability model, probit and logit regression.
Note that income is the variable you have to create by setting 0 for lower income group(<70) and 1 for middle, and 2 for the high income group(>200) based on applicant_income_000s.

```
# Create income variable
HMDA2017_AAW$income = as.numeric(case_when(HMDA2017_AAW$applicant_income_000s<70 ~ "0",
                                HMDA2017_AAW$applicant_income_000s>70 & HMDA2017_AAW$applicant_income_0
                                HMDA2017_AAW$applicant_income_000s>200 ~ "2"))
```

1. **Linear Probability Model**

```
denymod3 <- lm(deny ~ lti + applicant_race_1 + loan_type + property_type + loan_purpose + owner_occupan
coeftest(denymod3, vcov. = vcovHC, type = "HC1")
```

2. **Probit Model**

```
denyprobit3 <- glm(deny ~ lti + applicant_race_1 + loan_type + property_type + loan_purpose + owner_occu
                  family = binomial(link = "probit"),
                  data = HMDA2017_AAW)
coeftest(denyprobit3, vcov. = vcovHC, type = "HC1")
```

3. **Logit Model**

```
denylogit3 <- glm(deny ~ lti + applicant_race_1 + loan_type + property_type + loan_purpose + owner_occu
                 family = binomial(link = "logit"),
                 data = HMDA2017_AAW)
coeftest(denylogit3, vcov. = vcovHC, type = "HC1")
```

Comparison Models

The summary result for *Question 1d* is shown as below:

```
##
## Table 3: Summary Statistics
## =======================================================================
##                                     Dependent variable:
##                     ---------------------------------------------------
##                                             deny
##                              OLS                probit     logistic
##                              (1)                  (2)         (3)
## -----------------------------------------------------------------------
## lti                       -0.0004*             -0.001      -0.002
##                           (0.0002)             (0.001)     (0.002)
## applicant_race_1          0.021***            0.131***    0.254***
##                           (0.002)              (0.012)     (0.023)
## loan_type                  0.001               -0.004      -0.003
##                           (0.002)              (0.017)     (0.034)
## property_type            -0.263***            -1.085***   -1.963***
##                           (0.015)              (0.078)     (0.129)
## loan_purpose             -0.023***            -0.183***   -0.365***
##                           (0.001)              (0.009)     (0.019)
## owner_occupancy          -0.017***            -0.121***   -0.239***
##                           (0.004)              (0.030)     (0.060)
## applicant_sex            -0.006***            -0.045***   -0.093***
##                           (0.002)              (0.017)     (0.034)
## income                    0.030***            0.232***    0.485***
##                           (0.002)              (0.014)     (0.029)
## Constant                  1.134***            2.325***    4.111***
##                           (0.019)              (0.114)     (0.211)
## -----------------------------------------------------------------------
## Observations               55,864               55,864      55,864
## R2                         0.024
## Adjusted R2                0.024
## Log Likelihood                               -13,647.410 -13,653.580
## Akaike Inf. Crit.                             27,312.810  27,325.160
## Residual Std. Error     0.253 (df = 55855)
## F Statistic         171.554*** (df = 8; 55855)
## =======================================================================
## Note:                                       *p<0.1; **p<0.05; ***p<0.01
```

e.

Assume that model described in (d) is the one with every available data. Construct the table like Table 11.2 of the textbook comparing the various model. You could take out the variables from (d) or you could add more from the original data set. When you include the data, be cautious what each of variable stands for. You have to consider at least five specifications using probit or logit.

```r
denyprobit4 <- glm(deny ~ applicant_race_1 + property_type + loan_purpose + owner_occupancy + applicant_
                  family = binomial(link = "probit"),
                  data = HMDA2017_AAW)
```

```r
denylogit4 <- glm(deny ~ applicant_race_1 + property_type + loan_purpose + owner_occupancy + applicant_s
                  family = binomial(link = "logit"),
                  data = HMDA2017_AAW)
```

```
##
## Table 4: Summary Statistics
## ================================================================================================
##                                                 Dependent variable:
##                      ---------------------------------------------------------------------------
##                                                        deny
##                        probit     logistic              OLS                 probit     logistic
##                          (1)         (2)                (3)                   (4)         (5)
## ------------------------------------------------------------------------------------------------
## lti                                                  -0.0004*               -0.001      -0.002
##                                                      (0.0002)               (0.001)     (0.002)
## applicant_race_1      0.132***    0.255***           0.021***               0.131***    0.254***
##                        (0.012)     (0.023)           (0.002)                (0.012)     (0.023)
## loan_type                                             0.001                 -0.004      -0.003
##                                                      (0.002)                (0.017)     (0.034)
## property_type        -1.082***   -1.959***          -0.263***              -1.085***   -1.963***
##                        (0.077)     (0.129)           (0.015)                (0.078)     (0.129)
## loan_purpose         -0.183***   -0.364***          -0.023***              -0.183***   -0.365***
##                        (0.009)     (0.019)           (0.001)                (0.009)     (0.019)
## owner_occupancy      -0.120***   -0.240***          -0.017***              -0.121***   -0.239***
##                        (0.030)     (0.059)           (0.004)                (0.030)     (0.060)
## applicant_sex        -0.045***   -0.092***          -0.006***              -0.045***   -0.093***
##                        (0.017)     (0.034)           (0.002)                (0.017)     (0.034)
## income                0.234***    0.488***           0.030***               0.232***    0.485***
##                        (0.014)     (0.029)           (0.002)                (0.014)     (0.029)
## Constant              2.309***    4.091***           1.134***               2.325***    4.111***
##                        (0.107)     (0.195)           (0.019)                (0.114)     (0.211)
## ------------------------------------------------------------------------------------------------
## Observations           55,864      55,864             55,864                 55,864      55,864
## R2                                                    0.024
## Adjusted R2                                           0.024
## Log Likelihood      -13,647.910 -13,654.020                             -13,647.410 -13,653.580
## Akaike Inf. Crit.   27,309.820  27,322.030                              27,312.810  27,325.160
## Residual Std. Error                            0.253 (df = 55855)
## F Statistic                                171.554*** (df = 8; 55855)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

f.

Do you think that the racial issues in the mortgage approval become less serious? Or it gets worse?

To see the impact of racial issues on mortgage approval, we need to compute the difference in the probability of the mortgage denial when the applicant are American African (3) and White (5), held other variables constant, with $L/Iratio = 0.3$.

```r
predictions <- predict(denyprobit3,
                       newdata = data.frame("applicant_race_1"=c(3,5),
                                            "loan_type" = c(0,0),
                                            "property_type" = c(0,0),
                                            "loan_purpose" = c(0,0),
                                            "owner_occupancy" = c(0,0),
                                            "income" = c(0,0),
                                            "applicant_sex" = c(0,0),
                                            "lti" = c(0.3,0.3)),
                       type = "response")
diff(predictions)
```

```
##           2
## 0.001840278
```

The result shows that racial issue can raise the denial probability up to 0.2%, the racial issue has become less serious when compared to the model in *Table 11.2*, with a difference of 7.1%.

g.

Make new sub-sample include all races (1~5) or exclude category 6 and 7 in applicant_race_1.
Then repeat (d) and (e) for new sub-sample and answer (f) with the new result.

```r
HMDA2017 %>%
  filter(applicant_race_1 == c(1:5))
```

```
## Warning: There was 1 warning in ‘filter()‘.
## i In argument: ‘applicant_race_1 == c(1:5)‘.
## Caused by warning in ‘applicant_race_1 == c(1:5)‘:
## ! longer object length is not a multiple of shorter object length
```

```r
HMDA2017$income = as.numeric(case_when(HMDA2017$applicant_income_000s<70 ~ "0",
                                HMDA2017$applicant_income_000s>70 & HMDA2017$applicant_income_000s<200 ~
                                HMDA2017$applicant_income_000s>200 ~ "2"))
```

1. **Linear Probability Model**

```r
denymod5 <- lm(deny ~ lti + applicant_race_1 + loan_type + property_type + loan_purpose + owner_occupanc
coeftest(denymod5, vcov. = vcovHC, type = "HC1")
```

2. **Probit Model**

```r
denyprobit5 <- glm(deny ~ lti + applicant_race_1 + loan_type + property_type + loan_purpose + owner_occu
                  family = binomial(link = "probit"),
                  data = HMDA2017)
coeftest(denyprobit5, vcov. = vcovHC, type = "HC1")
```

3. **Logit Model**

```r
denylogit5 <- glm(deny ~ lti + applicant_race_1 + loan_type + property_type + loan_purpose + owner_occu
                  family = binomial(link = "logit"),
                  data = HMDA2017)
coeftest(denylogit5, vcov. = vcovHC, type = "HC1")
```

Comparison Models

The summary result for *Question 1g* is shown as below:

```
##
## Table 5: Summary Statistics
## =======================================================================
##                                      Dependent variable:
##                        ------------------------------------------------
##                                             deny
##                                  OLS                  probit    logistic
##                                  (1)                    (2)        (3)
## -----------------------------------------------------------------------
## lti                           -0.0005**              -0.001     -0.002
##                               (0.0002)               (0.001)    (0.002)
## applicant_race_1               0.005***              0.041***   0.079***
##                               (0.001)                (0.007)    (0.014)
## loan_type                     -0.004*               -0.036**   -0.066**
##                               (0.002)                (0.015)    (0.029)
## property_type                 -0.267***             -1.079***  -1.947***
##                               (0.014)                (0.071)    (0.118)
## loan_purpose                  -0.026***             -0.197***  -0.392***
##                               (0.001)                (0.008)    (0.016)
## owner_occupancy               -0.021***             -0.150***  -0.289***
##                               (0.003)                (0.024)    (0.047)
## applicant_sex                 -0.011***             -0.076***  -0.153***
##                               (0.001)                (0.010)    (0.021)
## income                         0.031***              0.227***   0.469***
##                               (0.002)                (0.012)    (0.024)
## Constant                       1.234***              2.875***   5.159***
##                               (0.016)                (0.090)    (0.162)
## -----------------------------------------------------------------------
## Observations                   71,777                 71,777     71,777
## R2                             0.023
## Adjusted R2                    0.022
## Log Likelihood                                     -18,402.940 -18,409.850
## Akaike Inf. Crit.                                  36,823.890  36,837.690
## Residual Std. Error      0.261 (df = 71768)
## F Statistic          206.615*** (df = 8; 71768)
## =======================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

To examine the impact of racial issue in 2017 when considering all kinds of race, we compute the difference in the probability of mortgage denial. I use the *probit model* with $L/Iratio = 0.3$:

```
predictions <- predict(denyprobit5,
                       newdata = data.frame("applicant_race_1"=c(3,5),
                                            "loan_type" = c(0,0),
                                            "property_type" = c(0,0),
                                            "loan_purpose" = c(0,0),
                                            "owner_occupancy" = c(0,0),
                                            "income" = c(0,0),
                                            "applicant_sex" = c(0,0),
                                            "lti" = c(0.3,0.3)),
                       type = "response")
diff(predictions)
```

```
##            2
## 0.0003228508
```

The difference is shown to be even smaller when including all kinds of race, 0.003%. Thus, in 2017, the racial issue has been less serious in the mortgage when compared to 1990.

Question 2:

a.

Is the data set a balanced panel?

```
df <- read.csv("income_democracy.csv")
is.pbalanced(df, df = c('country','year'))
```

```
## [1] FALSE
```

The data set is a **unbalanced panel**. From the data, we see that there are 9 time periods for each countries. However, for some countries, like Benin, Cameroon Central, or African Republic, there are only 8 time periods recorded. It means that they are missing one entity for one time period.

b.

i.

```
summary(df$dem_ind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1667  0.5000  0.4991  0.8333  1.0000     103
```

```
sd(df$dem_ind,na.rm = TRUE)
```

```
## [1] 0.3713367
```

What are the minimum and maximum value of Dem_ind in the data set?
- The minimum value is 0.00.
- The maximum value is 1.00.

What are the mean and standard deviation of Dem_ind in the data set?
- The mean is 0.4991.
- The standard deviation is 0902.

What are the 10th, 25th, 50th, 75th, and 90th percentile of its distribution?
- The 10th percentile is 0.00.
- The 25th percentile is 0.167.
- The 50th percentile is 0.50.
- The 75th percentile is 0.833.
- The 90th percentile is 1.00.

ii.

```
b2 <- subset(df,country == "United States")
head(b2$dem_ind)
```

```
## [1] 0.95 0.92 1.00 1.00 1.00 1.00
```

What are the value of Dem_ind for the United States in 2000? 1.

```
mean(b2$dem_ind)
```

```
## [1] 0.9855556
```

Averaged over all years in the data set? 0.98556.

iii.

```
b3 <- subset(df,country == "Libya")
head(b3$dem_ind)
```

```
## [1] 0.3100000 0.3400000 0.0000000 0.0000000 0.1666667 0.1666667
```

What is the value of Dem_ind for Libya in 2000? 0.167.

```
mean(b3$dem_ind)
```

```
## [1] 0.1092593
```

Averaged over all years in the data set? 0.109.

iv.

List five countries with an average value of Dem_ind greater than 0.95; less than 0.10; and between 0.3 and 0.7.

```r
    # 1) Remove rows with NA in Column "dem_ind"
df.c <- df[complete.cases(df[ , c('dem_ind')]), ]

    # 2) Compute average dem_ind for each country
df.c$ave_dem <- ave(df.c$dem_ind, df.c$country)

    # 3) Countries with high averaged dem_ind: ave_dem>0.95
list.high <- subset(df.c, ave_dem > 0.95)$country
unique(list.high)
```

```
##  [1] "Australia"           "Austria"             "Belgium"
##  [4] "Belize"              "Barbados"            "Canada"
##  [7] "Switzerland"         "Costa Rica"          "Czech Republic"
## [10] "Germany"             "Germany, West"       "Denmark"
## [13] "France"              "United Kingdom"      "Ireland"
## [16] "Iceland"             "Italy"               "Japan"
## [19] "Kiribati"            "St. Kitts and Nevis" "St. Lucia"
## [22] "Lithuania"           "Luxembourg"          "Malta"
## [25] "Netherlands"         "Norway"              "New Zealand"
## [28] "Slovakia"            "Slovenia"            "Sweden"
## [31] "United States"
```

```r
    # 4) Countries with low averaged dem_ind: ave_dem<0.1
list.low <- subset(df.c, ave_dem < 0.1)$country
unique(list.low)
```

```
##  [1] "Afghanistan"       "Angola"            "Burundi"
##  [4] "Brunei"            "China"             "Cuba"
##  [7] "Germany, East"     "Eritrea"           "Equatorial Guinea"
## [10] "Iraq"              "Myanmar"           "Korea, Dem. Rep."
## [13] "Rwanda"            "Saudi Arabia"      "Turkmenistan"
## [16] "Uzbekistan"        "Vietnam"           "Congo, Dem. Rep."
```

```r
    # 5) Countries with mid averaged dem_ind: 0.3<ave_dem<0.7
list.mid <- subset(df.c, 0.3 < ave_dem & ave_dem < 0.7)$country
unique(list.mid)
```

```
##  [1] "Argentina"              "Armenia"                "Antigua"
##  [4] "Bangladesh"             "Bulgaria"               "Bosnia and Herzegovina"
##  [7] "Bolivia"                "Brazil"                 "Chile"
## [10] "Comoros"                "Cape Verde"             "Dominican Republic"
## [13] "Ecuador"                "Spain"                  "Ethiopia 1993-"
## [16] "Fiji"                   "Georgia"                "Ghana"
## [19] "Gambia, The"            "Guinea-Bissau"          "Guatemala"
## [22] "Guyana"                 "Honduras"               "Hungary"
## [25] "Jordan"                 "Korea, Rep."            "Kuwait"
## [28] "Lebanon"                "Lesotho"                "Morocco"
## [31] "Madagascar"             "Maldives"               "Mexico"
## [34] "Macedonia, FYR"         "Mozambique"             "Malaysia"
## [37] "Nigeria"                "Nicaragua"              "Nepal"
## [40] "Pakistan-post-1972"     "Pakistan-pre-1972"      "Panama"
```

```
## [43] "Peru"                     "Philippines"              "Poland"
## [46] "Paraguay"                 "Russia"                   "Senegal"
## [49] "Singapore"                "El Salvador"              "Sao Tome and Principe"
## [52] "Suriname"                 "Seychelles"               "Thailand"
## [55] "Tonga"                    "Turkey"                   "Taiwan"
## [58] "Ukraine"                  "Yemen"                    "Yugoslavia - post 1991"
## [61] "South Africa"             "Zambia"                   "Zimbabwe"
```

c.

i.

How large is the estimated coefficient of Log_GDPPC?
Is the coefficient statistically significant?

```r
  # OLS regression with clustered standard errors
fit.c <- lm(dem_ind ~ log_gdppc, data = df)
summary(fit.c)
```

```
##
## Call:
## lm(formula = dem_ind ~ log_gdppc, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72854 -0.19534  0.02586  0.19123  0.72698
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.354828   0.070919  -19.10   <2e-16 ***
## log_gdppc    0.235673   0.008626   27.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2719 on 956 degrees of freedom
##   (411 observations deleted due to missingness)
## Multiple R-squared:  0.4385, Adjusted R-squared:  0.4379
## F-statistic: 746.5 on 1 and 956 DF,  p-value: < 2.2e-16
```

At 0.01% significance level, $\beta_1 = 0.24$, the coefficient of log_GDPpc is statistically significant based on the result.

ii.

If per capital income in a country increases by 20%, by how much is Dem_ind predicted to increase?

```r
predictions <- predict(fit.c,
                       newdata = data.frame("log_gdppc" = c(0,0.2)),
                       type = "response")
diff(predictions)
```

```
##          2
## 0.04713462
```

- If per capital income in a country increases by 20%, *Dem_ind* is predicted to increased by 4.71%.

What is the 95% confidence interval for the prediction?

```
  # results with clustered standard errors
result.c <- coeftest(fit.c, vcovCL(fit.c, cluster=df$country, type="HC1"))
result.c
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -1.354828   0.100421 -13.491 < 2.2e-16 ***
## log_gdppc    0.235673   0.011837  19.910 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
  # 95% CI for the slope coefficient
b <- fit.c$coefficients["log_gdppc"]  #estimated slope coefficient
ucl <- b + 1.96*result.c[2,2]
lcl <- b - 1.96*result.c[2,2]
lcl; ucl
```

```
## log_gdppc
## 0.2124726
```

```
## log_gdppc
## 0.2588736
```

- The 95% confidence interval for the prediction is $[0.212; 0.259]$.

Is the predicted increase in *Dem_ind* large or small? (Explain what you mean by large and small.)
- The predicted increase in *Dem_ind* is small, because on average, 1% increase in *log_GDPpc* will lead to only 0.002 percentage points increase in *Dem_ind*.

iii.

Why is it important to use clustered standard errors for the regression?
Do the results change if you do not use clustered standard errors?
The clustered standard error for *Dem_ind* is 0.012; while the unclustered standard error is smaller (0.009). This is because the unclustered standard error ignores the correlations between the country entities.
The result still stays the same, as the model generated for the clustered and unclustered standard errors are the same.

d.

i.

Suggest a variable that varies across countries but plausibly varies little – or not at all – over time and that

could cause omitted variable bias in the regression in (c).
- Some variables that can vary across countries but little over time can be religion, cultures, social structures, etc. These variables can affect the country's demography while also correlates with the economic development, thus affect the per capital income.

ii.
Estimate the regression in (c), allowing for country fixed effect.

```
fit.d.ii <- plm(dem_ind ~ log_gdppc,
                data = df,
                df = c("country", "year"),
                effect = "individual",
                model = "within")
coeftest(fit.d.ii, vcovHC(fit.d.ii, cluster="group", type="HC1"))
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## log_gdppc 0.083741   0.031421  2.6652 0.007849 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the country fixed effect, the estimated coefficient (β_1) falls to 0.084 with a clustered standard error of 0.031. The estimated effect of a country fixed effect is smaller compared to the model in (c), however, it is still statistically significant at the 1% significance level.

iii. & iv.
Exclude the data for Azerbaijan, and rerun the regression.
Do the result changes? Why or why not?

```
df.no.Azer <- subset(df, country!="Azerbaijan")  # Exclude data for Azerbaijan
fit.d.iii <- plm(dem_ind ~ log_gdppc, data = df.no.Azer,
                 df = c("country", "year"), effect = "individual",
                 model="within")
coeftest(fit.d.iii, vcovHC(fit.d.iii, cluster = "group", type = "HC0"))
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## log_gdppc 0.083741   0.031404  2.6666 0.007817 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Excluding the data for Azerbaijan does not change the result much, because Azerbaijan only has available data for 2000, these data has been absorbed by the country-specific fixed effect.

v.
Assume there are additional demographic controls in the data set. Should these variables be included in the

regression?
If so, how do the results change when they are included?

In the data set, there are some additional demographic controls that can be included as control variables such as *log_pop*, *age_1 - age_5*, or *educ*.
Including these control variables will have the following effects on the results:

```
fit.d.v <- plm(dem_ind ~ log_gdppc + age_2 + age_3 +
              age_4 + age_5 + educ + log_pop, data = df,
              effect="twoways", model="within")
coeftest(fit.d.v, vcovHC(fit.d.v, cluster="group", type="HC0"))
```

```
##
## t test of coefficients:
##
##              Estimate  Std. Error t value Pr(>|t|)
## log_gdppc  0.02520125  0.05312453  0.4744 0.635410
## age_2     -0.52551573  0.60034016 -0.8754 0.381745
## age_3     -2.48123518  0.88025459 -2.8188 0.004988 **
## age_4      0.29781155  1.27765323  0.2331 0.815773
## age_5      0.60540584  1.27632866  0.4743 0.635444
## educ      -0.00040127  0.02288646 -0.0175 0.986017
## log_pop   -0.06922950  0.12261279 -0.5646 0.572555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When other demographic controls are included, the estimated coefficient on *log_GDPpc* falls further to 0.03 with a standard error of 0.05. Therefore, after controlling for omitted variables - particularly country-fixed effects - *log_GDPpc* is not statistically significant in changing *Dem_ind*. It means that with other control variables included, there is little evidence of an income effect on the demand for democracy.

e.

The income effect on the demand for democracy is evidently strongest when the regression does not include the country fixed effect. With 1% increase in per capital income will increase 0.2%.
With the country fixed effect, the income effect on the demand for democracy is smaller, yet still significant. When controlling other demographics variables such as population, age, and education, there is little evidence of an income effect on the demand for democracy.