Moh Jaiswal 500916860 Professor Laura Turner

Assignment 4

Question 1

Linear Probability

This model is flawed as it assumes the conditional probability function to be linear, it needs to assume that it is normally distributed (bell curve shape). It is because of this flaw, the results for this model are not reliable. For example: If the piratio is approaching 0, the probability figure can also be negative. There are several other approaches to solve for this problem, that use non-linear functions on a linear model.

Probit

$E(Y|X) = P(Y = 1|X) = \Phi(\beta 0 + \beta 1X)$

Here $\beta 0 + \beta 1X$ serves as our Z standardised value, such that the Probit coefficient β1 is the change in z associated with a one unit change in X. Although the effect on z of a change in X is linear, the link between z and the dependent variable Y is nonlinear since Φ is a nonlinear function of X.

All the variables have a positive relationship between z score and an increment by 1 unit, except for female (- 0.1398). Here income has no affect on the z score.

Logit

Understanding the coefficients for a logit model is not as straightforward as linear regression. Here we are dealing with the probability of something happening rather than a direct impact of IV on DV.

a) logit positive value = logistic > 1 = increase in the probability of the event when you have a positive change in the IV

b) logit negative value = logistic < 1 = decrease in the probability of the event when you have a positive change in the IV

Both logit and probit models produce very similar estimates of the probability that a mortgage application will be denied depending on the applicants payment-to-income ratio.

I would prefer logit due it being more efficient and overall a better predicting model when working with multiple dummy variables. It also relies on logisticial values, which can never be negative, giving us an accurate picture of the significance each variable plays on the model.

```r
rm(list=ls())
options(warn=-1)
options(scipen = 20)
library(haven)
HMDA <- read_dta("hmda_sw.dta")
HMDA$black <- ifelse(HMDA$s13==3,1,0)
HMDA$piratio <- HMDA$s46/100
HMDA$deny <- ifelse(HMDA$s7==3,1,0) #
HMDA$female <- ifelse(HMDA$s15=="2", 1, 0)
HMDA$marital <- ifelse(HMDA$s23a=="M",1,0)


model1 <- lm(deny ~ black + piratio + s17 + female, data=HMDA)
summ(model1, digits=4, robust="HC1")

model2 <- glm(deny ~ black + piratio + s17 + female, family = binomial(link = "probit"), data = HMDA)
summ(model2, digits=4, type="HC1")

model3 <- glm(deny ~ black + piratio + s17 + female, family = binomial(link = "logit"), data = HMDA)
summ(model3, digits=4)
```

```
> summ(model1, digits=4, robust="HC1")
MODEL INFO:
Observations: 2380
Dependent Variable: deny
Type: OLS linear regression

MODEL FIT:
F(4,2375) = 52.1550, p = 0.0000
R² = 0.0807
Adj. R² = 0.0792

Standard errors: Robust, type = HC1
---------------------------------------------------------
                        Est.      S.E.      t val.      p
----------------- --------- -------- ---------- --------
(Intercept)          -0.0875    0.0286    -3.0575   0.0023
black                 0.1840    0.0252     7.3095   0.0000
piratio               0.5571    0.0882     6.3145   0.0000
s17                   0.0000    0.0000     2.2394   0.0252
female               -0.0270    0.0159    -1.6971   0.0898
```

```
> summ(model2, digits=4, type="HC1")
MODEL INFO:
Observations: 2380
Dependent Variable: deny
Type: Generalized linear model
  Family: binomial
  Link function: probit

MODEL FIT:
χ²(4) = 159.8006, p = 0.0000
Pseudo-R² (Cragg-Uhler) = 0.1250
Pseudo-R² (McFadden) = 0.0916
AIC = 1594.3700, BIC = 1623.2443

Standard errors: MLE
----------------------------------------------------------
                        Est.      S.E.      z val.       p
----------------- --------- -------- ---------- --------
(Intercept)          -2.2464    0.1375   -16.3424   0.0000
black                 0.7465    0.0853     8.7467   0.0000
piratio               2.7287    0.3804     7.1737   0.0000
s17                   0.0000    0.0000     2.8400   0.0045
female               -0.1398    0.0884    -1.5815   0.1138
```

```
> summ(model3, digits=4)
MODEL INFO:
Observations: 2380
Dependent Variable: deny
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
χ²(4) = 163.2743, p = 0.0000
Pseudo-R² (Cragg-Uhler) = 0.1276
Pseudo-R² (McFadden) = 0.0936
AIC = 1590.8963, BIC = 1619.7706

Standard errors: MLE
----------------------------------------------------------
                        Est.      S.E.      z val.       p
----------------- --------- -------- ---------- --------
(Intercept)          -4.1105    0.2700   -15.2265   0.0000
black                 1.3502    0.1504     8.9747   0.0000
piratio               5.3715    0.7292     7.3661   0.0000
s17                   0.0000    0.0000     3.0376   0.0024
female               -0.2845    0.1680    -1.6938   0.0903
```

Question 2 (*Appendix: Figures 2.0 and onwards*)

Coefficient Analysis:
piratio: 5.2699 (The probability of being denied increases as the PI ratio increases, significantly)
black: 1.3338 (The probability of being denied increases as the PI ratio increases, but slightly)
s17: 0 (The probability of being denied does not change as the price of the house increases, this might be due to income being inversely related to piratio)
female: - 0.5227 (The probability of being denied increases if the applicant is female, albeit marginally)
martial: - 0.4978 (The probability of being denied increases if the applicant is married, it has a similar effect that being a female has)

Conclusion:
The results point toward a high focus on piratio being the most important predictor for the loan, which is a sensible metric to underwrite credit approvals.

```
> # Question 2
>
> model4 <- glm(deny ~ black + piratio + s17 + female + marital, family = binomial(link = "logit"), data = HMDA)
> summ(model4, digits=4)
MODEL INFO:
Observations: 2380
Dependent Variable: deny
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
χ²(5) = 175.0084, p = 0.0000
Pseudo-R² (Cragg-Uhler) = 0.1365
Pseudo-R² (McFadden) = 0.1003
AIC = 1581.1622, BIC = 1615.8114

Standard errors: MLE
--------------------------------------------------------
                        Est.     S.E.     z val.       p
----------------- --------- -------- ---------- --------
(Intercept)          -3.7417   0.2862   -13.0757   0.0000
black                 1.3338   0.1512     8.8205   0.0000
piratio               5.2699   0.7277     7.2423   0.0000
s17                   0.0000   0.0000     3.1338   0.0017
female               -0.5227   0.1817    -2.8761   0.0040
marital              -0.4978   0.1443    -3.4488   0.0006
--------------------------------------------------------
```

Question 3 (*Appendix: Figures 3.0 and onwards*)

The probability of being denied with a 10% increase to piratio in the case of using linear probability model predictions, the difference in denial is 1.818%.

The probability of being denied with a 10% increase to piratio in the case of probit predictions, the difference in denial probabilities is 2.323%.

The probability of being denied with a 10% increase to piratio in the case of logit predictions, the difference in denial probabilities is 2.427%.

This model holds all other variables where the predicted values are based on the fact that the applicant is a black female with median income. Normally there should be a positive relation between probability of being denied and income, however, all three results tell us otherwise.

```r
# Question 3

model5 <- glm(deny ~ black + piratio + s17 + female + marital, family = binomial(link = "probit"), data = HMDA)
summ(model5, digits=4)

model6 <- lm(deny ~ black + piratio + s17 + female + marital, data=HMDA)
summ(model6, digits=4, robust="HC1")

# Logit

prediction10 <- predict(model4,
                        newdata = data.frame("black" = 1,
                                             "piratio" = median(HMDA$piratio),
                                             "female" = 1,
                                             "s17" = median(HMDA$s17),
                                             "marital" = 1),
                        type = "response")

prediction11 <- predict(model4,
                        newdata = data.frame("black" = 1,
                                             "piratio" = 1.1*median(HMDA$piratio),
                                             "female" = 1,
                                             "s17" = median(HMDA$s17),
                                             "marital" = 1),
                        type = "response")

logdif <- prediction11 - prediction10
logdif

# Probit

prediction20 <- predict(model5,
                        newdata = data.frame("black" = 1,
                                             "piratio" = median(HMDA$piratio),
                                             "female" = 1,
                                             "s17" = median(HMDA$s17),
                                             "marital" = 1),
                        type = "response")

prediction21 <- predict(model5,
                        newdata = data.frame("black" = 1,
                                             "piratio" = 1.1*median(HMDA$piratio),
                                             "female" = 1,
                                             "s17" = median(HMDA$s17),
                                             "marital" = 1),
                        type = "response")

prodif <- prediction21 - prediction20
prodif
```

```r
# Linear probability


prediction30 <- predict(model6,
                        newdata = data.frame("black" = 1,
                                             "piratio" = median(HMDA$piratio),
                                             "female" = 1,
                                             "s17" = median(HMDA$s17),
                                             "marital" = 1),
                        type = "response")

prediction31 <- predict(model6,
                        newdata = data.frame("black" = 1,
                                             "piratio" = 1.1*median(HMDA$piratio),
                                             "female" = 1,
                                             "s17" = median(HMDA$s17),
                                             "marital" = 1),
                        type = "response")


OLSdif = prediction31-prediction30
OLSdif
```

```
> summ(model5, digits=4)
MODEL INFO:
Observations: 2380
Dependent Variable: deny
Type: Generalized linear model
  Family: binomial
  Link function: probit

MODEL FIT:
χ²(5) = 172.3594, p = 0.0000
Pseudo-R² (Cragg-Uhler) = 0.1345
Pseudo-R² (McFadden) = 0.0988
AIC = 1583.8112, BIC = 1618.4603

Standard errors: MLE
--------------------------------------------------------
                      Est.     S.E.     z val.        p
----------------- --------- -------- ---------- --------
(Intercept)        -2.0520   0.1480   -13.8615   0.0000
black               0.7397   0.0857     8.6333   0.0000
piratio             2.6978   0.3826     7.0522   0.0000
s17                 0.0000   0.0000     2.9281   0.0034
female             -0.2714   0.0962    -2.8219   0.0048
marital            -0.2728   0.0768    -3.5539   0.0004
```

```
> summ(model6, digits=4, robust="HC1")
MODEL INFO:
Observations: 2380
Dependent Variable: deny
Type: OLS linear regression

MODEL FIT:
F(5,2374) = 44.5979, p = 0.0000
R² = 0.0859
Adj. R² = 0.0839

Standard errors: Robust, type = HC1
-----------------------------------------------------
                     Est.     S.E.     t val.       p
----------------- -------- -------- ---------- --------
(Intercept)       -0.0482   0.0306   -1.5740   0.1156
black              0.1814   0.0252    7.2068   0.0000
piratio            0.5508   0.0856    6.4383   0.0000
s17                0.0000   0.0000    2.2968   0.0217
female            -0.0522   0.0176   -2.9603   0.0031
marital           -0.0521   0.0151   -3.4545   0.0006
```

```
> logdif <- prediction11 - prediction10
> logdif
      1
0.02427
```

```
> prodif <- prediction21 - prediction20
> prodif
      1
0.02323
```

```
> OLSdif = prediction31-prediction30
> OLSdif
      1
0.01818
```