HW11_Sampathirao_A

Anvita Sampathirao

8/5/2019

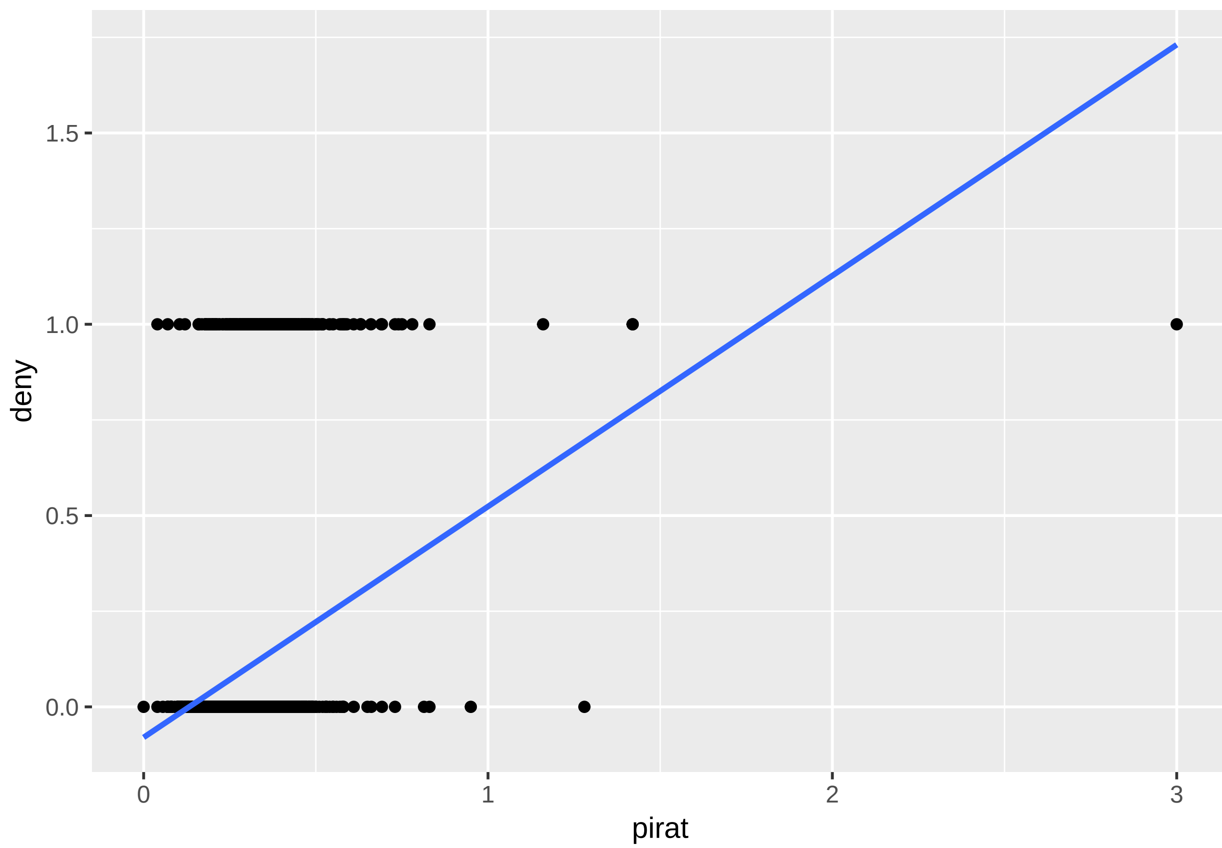
```r
suppressMessages(library(AER))
data(HMDA)
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```r
suppressMessages(library(stargazer))
HMDA$deny <- ifelse(HMDA$deny == "yes", 1, 0)
HMDA$afam <- ifelse(HMDA$afam == "yes", 1, 0)
HMDA$phist <- ifelse(HMDA$phist == "yes", 1, 0)
HMDA$selfemp <- ifelse(HMDA$selfemp == "yes", 1, 0)
HMDA$insurance <- ifelse(HMDA$insurance == "yes", 1, 0)
HMDA$condomin <- ifelse(HMDA$condomin == "yes", 1, 0)
HMDA$single <- ifelse(HMDA$single == "yes", 1, 0)
HMDA$hschool <- ifelse(HMDA$hschool == "yes", 1, 0)
HMDA$chist <- as.numeric(HMDA$chist)
HMDA$mhist <- as.numeric(HMDA$mhist)
```

#1

```r
plot1<- ggplot(data = HMDA, aes(x= pirat, y= deny)) +
  geom_point () +
  geom_smooth(method= "lm", formula = y~x, se= FALSE)
plot1
```



#2

```
LinReg1<- lm(deny ~ pirat, data = HMDA)
LinReg2<- lm(deny ~ pirat + afam, data = HMDA)
LinRegs<- list (LinReg1, LinReg2)
stargazer(LinRegs, type = "latex", intercept.bottom = FALSE, df = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 13, 2019 - 12:38:31 AM

In Model (1), - the estimator for pirat variable is statistically different than 0 and hence, a unit change in payments to income ratio will result in a loan most likely being denied by 60.4%

In Model (2), - the estimator for pirat variable is statistically different than 0 and a unit change in payments to income ratio will result in a in a loan most likely being denied by 55.9%

- the estimator for afam is also statistically different than 0. A loan is likely to get denied by 17.7% if a person applying for it is an african american as compared to not being an african american

Yes, afam is a relevant omitted variable in Model (1), as we note that the value of beta_pirat has changed from Model (1) to Model (2)

#3

```
PIavg<- mean(HMDA$pirat)
betas<- LinReg2$coefficients
P1.1<- as.numeric(betas[1] + (PIavg*betas[2]) + (1*betas[3])) #When an african american
```

Table 1:

	Dependent variable:	
	deny	
	(1)	(2)
Constant	−0.080***	−0.091***
	(0.021)	(0.021)
pirat	0.604***	0.559***
	(0.061)	(0.060)
afam		0.177***
		(0.018)
Observations	2,380	2,380
R^2	0.040	0.076
Adjusted R^2	0.039	0.075
Residual Std. Error	0.318	0.312
F Statistic	98.406***	97.760***
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
P1.0<- as.numeric(betas[1] + (PIavg*betas[2]) + (0*betas[3])) #When not an african american
P1.1/P1.0
```

```
## [1] 2.878033
```

#4

```
#Finding out the order in which variables are to be added by strength of correlation
corData<- cor(HMDA)
corData<- corData[, colnames(corData) %in% c("deny", "afam")]
corData
```

```
##                  deny        afam
## deny       1.00000000  0.20516050
## pirat      0.19934276  0.07667742
## hirat      0.13307622  0.04605482
## lvrat      0.16222276  0.16207609
## chist      0.26236869  0.22115730
## mhist      0.10964947  0.11991007
## phist      0.27292927  0.16155178
## unemp      0.04359448 -0.06623347
## selfemp    0.05177546 -0.05418954
## insurance  0.35215423  0.08691480
## condomin   0.03958091  0.18390619
## afam       0.20516050  1.00000000
## single     0.07660881  0.10257731
## hschool   -0.06452406 -0.05156050
```

```r
a<- corData[,1] * corData[,2]
sort(a, decreasing = TRUE)
```

```
##        deny        afam       chist       phist   insurance
##  0.205160498  0.205160498  0.058024752  0.044092209  0.030607413
##       lvrat       pirat       mhist      single    condomin
##  0.026292431  0.015285089  0.013148075  0.007858325  0.007279174
##       hirat     hschool     selfemp       unemp
##  0.006128802  0.003326893 -0.002805688 -0.002887413
```

```r
#Running Logit Regressions
LogitReg1 <- glm(deny ~ pirat + afam,
                 data = HMDA, family = "binomial")
LogitReg2 <- glm(deny ~ pirat + afam + chist,
                 data = HMDA, family = "binomial")
LogitReg3 <- glm(deny ~ pirat + afam + chist + phist,
                 data = HMDA, family = "binomial")
LogitReg4 <- glm(deny ~ pirat + afam + chist + phist + insurance,
                 data = HMDA, family = "binomial")
LogitReg5 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat,
                 data = HMDA, family = "binomial")
LogitReg6 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                 + mhist, data = HMDA, family = "binomial")
LogitReg7 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                 + mhist + single, data = HMDA, family = "binomial")
LogitReg8 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                 + mhist + single + condomin,
                 data = HMDA, family = "binomial")
LogitReg9 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                 + mhist + single + condomin + hirat,
                 data = HMDA, family = "binomial")
LogitReg10 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                  + mhist + single + condomin + hirat + hschool,
                  data = HMDA, family = "binomial")
LogitReg11 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                  + mhist + single + condomin + hirat + hschool + selfemp,
                  data = HMDA, family = "binomial")
LogitReg12 <- glm(deny ~ pirat + afam + chist + phist + insurance + lvrat
                  + mhist + single + condomin + hirat + hschool + selfemp + unemp,
                  data = HMDA, family = "binomial")

LogitRegs1 <- list(LogitReg1, LogitReg2, LogitReg3, LogitReg4, LogitReg5,
                  LogitReg6)
LogitRegs2 <- list(LogitReg7, LogitReg8, LogitReg9, LogitReg10,
                  LogitReg11, LogitReg12)

stargazer(LogitRegs1, type = "latex", intercept.bottom = FALSE, df= FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 13, 2019 - 12:38:31 AM

Table 2:

	Dependent variable:					
	deny					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	−4.126***	−4.849***	−4.751***	−4.959***	−6.263***	−6.663***
	(0.268)	(0.288)	(0.292)	(0.304)	(0.488)	(0.533)
pirat	5.370***	5.133***	4.913***	4.819***	4.705***	4.745***
	(0.728)	(0.732)	(0.749)	(0.768)	(0.777)	(0.779)
afam	1.273***	0.950***	0.864***	0.815***	0.708***	0.688***
	(0.146)	(0.155)	(0.160)	(0.171)	(0.174)	(0.174)
chist		0.340***	0.277***	0.305***	0.298***	0.286***
		(0.034)	(0.037)	(0.039)	(0.039)	(0.039)
phist			1.290***	1.289***	1.250***	1.247***
			(0.192)	(0.202)	(0.202)	(0.203)
insurance				4.708***	4.524***	4.513***
				(0.545)	(0.552)	(0.551)
lvrat					1.807***	1.695***
					(0.497)	(0.500)
mhist						0.283**
						(0.140)
Observations	2,380	2,380	2,380	2,380	2,380	2,380
Log Likelihood	−795.695	−748.766	−727.667	−654.791	−647.895	−645.863
Akaike Inf. Crit.	1,597.390	1,505.532	1,465.334	1,321.582	1,309.791	1,307.725

Note: *p<0.1; **p<0.05; ***p<0.01

```r
stargazer(LogitRegs2, type = "latex", intercept.bottom = FALSE, df= FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Aug 13, 2019 - 12:38:32 AM

We see that the value of Beta_afam is changing in every regression until Model (12). Hence, we have to test for multicollinearity.

```r
#Check for multi collinearity
aux1_lr12 <- lm(pirat ~ afam + chist + phist + insurance + lvrat +
                  mhist + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux2_lr12 <- lm(afam ~ pirat + chist + phist + insurance + lvrat +
                  mhist + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux3_lr12 <- lm(chist ~ pirat + afam + phist + insurance + lvrat +
                  mhist + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux4_lr12 <- lm(phist ~ pirat + afam + chist + insurance + lvrat +
                  mhist + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux5_lr12 <- lm(insurance ~ pirat + afam + chist + phist + lvrat +
                  mhist + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux6_lr12 <- lm(lvrat ~ pirat + afam + chist + phist + insurance +
                  mhist + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux7_lr12 <- lm(mhist ~ pirat + afam + chist + phist + insurance +
                  lvrat + single + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux8_lr12 <- lm(single ~ pirat + afam + chist + phist + insurance +
                  lvrat + mhist + condomin + hirat + hschool + selfemp + unemp,
                data = HMDA)
aux9_lr12<- lm(condomin ~ pirat + afam + chist + phist + insurance +
                 lvrat + mhist + single + hirat + hschool + selfemp + unemp,
               data = HMDA)
aux10_lr12<- lm(hirat ~ pirat + afam + chist + phist + insurance +
                  lvrat + mhist + single + condomin + hschool + selfemp + unemp,
                data = HMDA)
aux11_lr12<- lm(hschool ~ pirat + afam + chist + phist + insurance +
                  lvrat + mhist + single + condomin + hirat + selfemp + unemp,
                data = HMDA)
aux12_lr12<- lm(selfemp ~ pirat + afam + chist + phist + insurance +
                  lvrat + mhist + single + condomin + hirat + hschool + unemp,
                data = HMDA)
aux13_lr12 <- lm(unemp ~ pirat + afam + chist + phist + insurance +
                   lvrat + mhist + single + condomin + hirat + hschool + selfemp,
                 data = HMDA)
aux1_r2 <- summary(aux1_lr12)$r.squared
aux2_r2 <- summary(aux2_lr12)$r.squared
aux3_r2 <- summary(aux3_lr12)$r.squared
aux4_r2 <- summary(aux4_lr12)$r.squared
aux5_r2 <- summary(aux5_lr12)$r.squared
aux6_r2 <- summary(aux6_lr12)$r.squared
```

Table 3:

	Dependent variable:					
	deny					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-6.733^{***}	-6.724^{***}	-6.686^{***}	-5.592^{***}	-5.744^{***}	-6.047^{***}
	(0.535)	(0.535)	(0.538)	(0.674)	(0.673)	(0.697)
pirat	4.684^{***}	4.671^{***}	5.075^{***}	5.097^{***}	4.811^{***}	4.798^{***}
	(0.773)	(0.774)	(1.025)	(1.020)	(1.034)	(1.035)
afam	0.645^{***}	0.659^{***}	0.662^{***}	0.639^{***}	0.673^{***}	0.704^{***}
	(0.175)	(0.178)	(0.178)	(0.179)	(0.180)	(0.180)
chist	0.290^{***}	0.290^{***}	0.288^{***}	0.290^{***}	0.295^{***}	0.296^{***}
	(0.039)	(0.039)	(0.039)	(0.040)	(0.040)	(0.040)
phist	1.267^{***}	1.267^{***}	1.265^{***}	1.268^{***}	1.240^{***}	1.231^{***}
	(0.204)	(0.204)	(0.204)	(0.204)	(0.205)	(0.205)
insurance	4.510^{***}	4.504^{***}	4.499^{***}	4.528^{***}	4.548^{***}	4.545^{***}
	(0.552)	(0.553)	(0.553)	(0.554)	(0.554)	(0.554)
lvrat	1.682^{***}	1.689^{***}	1.699^{***}	1.702^{***}	1.790^{***}	1.804^{***}
	(0.499)	(0.499)	(0.499)	(0.499)	(0.498)	(0.499)
mhist	0.242^{*}	0.243^{*}	0.249^{*}	0.240^{*}	0.263^{*}	0.242^{*}
	(0.143)	(0.143)	(0.142)	(0.143)	(0.142)	(0.143)
single	0.387^{**}	0.404^{***}	0.411^{***}	0.453^{***}	0.453^{***}	0.450^{***}
	(0.151)	(0.155)	(0.156)	(0.157)	(0.157)	(0.157)
condomin		-0.075	-0.079	-0.109	-0.097	-0.066
		(0.168)	(0.168)	(0.169)	(0.169)	(0.170)
hirat			-0.737	-0.788	-0.496	-0.483
			(1.233)	(1.231)	(1.235)	(1.237)
hschool				-1.116^{***}	-1.166^{***}	-1.080^{**}
				(0.421)	(0.420)	(0.421)
selfemp					0.693^{***}	0.647^{***}
					(0.211)	(0.213)
unemp						0.060^{*}
						(0.034)
Observations	2,380	2,380	2,380	2,380	2,380	2,380
Log Likelihood	-642.576	-642.477	-642.299	-639.234	-634.259	-632.772
Akaike Inf. Crit.	1,303.152	1,304.953	1,306.598	1,302.469	1,294.518	1,293.544
Note:						*p<0.1; **p<0.05; ***p<0.01

```r
aux7_r2 <- summary(aux7_lr12)$r.squared
aux8_r2 <- summary(aux8_lr12)$r.squared
aux9_r2 <- summary(aux9_lr12)$r.squared
aux10_r2 <- summary(aux10_lr12)$r.squared
aux11_r2 <- summary(aux11_lr12)$r.squared
aux12_r2 <- summary(aux12_lr12)$r.squared
aux13_r2 <- summary(aux13_lr12)$r.squared
vif1 <- 1/ (1 - aux1_r2)
vif2 <- 1/ (1 - aux2_r2)
vif3 <- 1/ (1 - aux3_r2)
vif4 <- 1/ (1 - aux4_r2)
vif5 <- 1/ (1 - aux5_r2)
vif6 <- 1/ (1 - aux6_r2)
vif7 <- 1/ (1 - aux7_r2)
vif8 <- 1/ (1 - aux8_r2)
vif9 <- 1/ (1 - aux9_r2)
vif10 <- 1/ (1 - aux10_r2)
vif11 <- 1/ (1 - aux11_r2)
vif12 <- 1/ (1 - aux12_r2)
vif13 <- 1/ (1 - aux13_r2)
vifs <- list(vif1, vif2, vif3, vif4, vif5, vif6, vif7,
             vif8, vif9, vif10, vif11, vif12, vif13)
vifs > 10
```

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

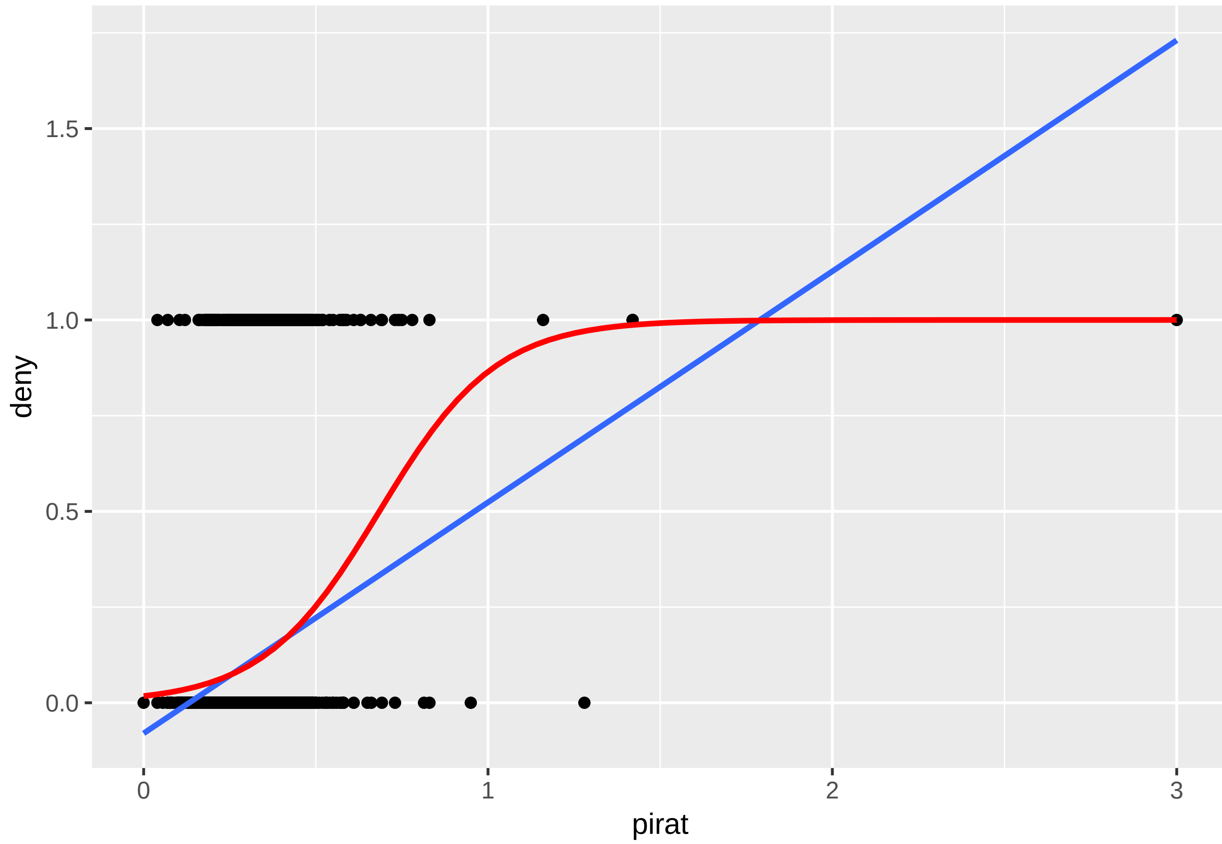
```r
vifs > 5
```

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

We can note that for all regressors VIF is less than 5 we can be confident that imperfect multicollienarity is not an issue in regression (12). And, if is not an issue in regression (12) - which includes the larger number of independent variables - then it won't be an issue in models (1) to (11). Hence, Model (12) is the model with the least bias.

#5

```r
plot2 <- plot1 + geom_smooth(method = "glm",
                             formula = y ~ x,
                             method.args = list(family = "binomial"),
                             color = "red", se = FALSE)
plot2
```



The logistic regression model is a more appropriate specification as it fits the values of deny better than the linear regressor which extends beyond the bounds of deny variable.

#6

```
betalog <- LogitReg12$coefficients
exp(betalog[3])
```

```
##     afam
## 2.021119
```

Because the relation is non-linear, the odds of a loan getting denied for an african american is 2.02 times the odds of a loan getting denied for a non african american.

#7

```
chistavg <- mean(HMDA$chist)
phistavg <- mean(HMDA$phist)
insuranceavg <- mean(HMDA$insurance)
lvratavg <- mean(HMDA$lvrat)
mhistavg <- mean(HMDA$mhist)
singleavg <- mean(HMDA$single)
condominavg<- mean(HMDA$condomin)
hiratavg <- mean(HMDA$hirat)
hschoolavg <- mean(HMDA$hschool)
selfempavg <- mean(HMDA$selfemp)
unempavg <- mean(HMDA$unemp)
```

```r
invlogit <- function(x){
  yloghat <- betalog[1] + (betalog[2]*PIavg) + (betalog[3]*x) +
    (betalog[4]*chistavg) + (betalog[5]*phistavg) + (betalog[6]*insuranceavg) +
    (betalog[7]*lvratavg) + (betalog[8]*mhistavg) + (betalog[9]*singleavg) +
    (betalog[10]*condominavg) + (betalog[11]*hiratavg) + (betalog[12]*hschoolavg) +
    (betalog[13]*selfempavg) + (betalog[14]*unempavg)
  yhat <- exp(yloghat)/ (1 + exp(yloghat))
  return(yhat)
}
P7.1.1 <- invlogit(1) #When an african american
P7.1.0 <- invlogit(0) #When not an african american
as.numeric(P7.1.1/P7.1.0)
```

```
## [1] 1.89095
```

The odds of a loan getting denied when an applicant is an african american versus when an applicant is not an african american is 1.89. This is a better estimator than the odds determined with the linear regression model because the range of of the odds ratio is bounded by (0,1)- that fits all values within the domain of deny variable.