PSet 7

Spencer Papay - ssp2170

4/11/2017

Question 1

a) The linear probability model is a multiple regression model when Y is binary rather than continuous. Because the dependent variable is binary, the population regression function aligns to the probability that $Y = 1$, given X, the independent variable..

It is a multiple linear regression model ($Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_n X_{ni} + \mu_i$) applied to a binary (either 0 or 1) dependent variable Y_i. Because Y is binary, $E(Y|X_1, ..., X_n) = P(Y = 1|X_1, ..., X_n)$, meaning the linear probability model is $P(Y = 1|X_1, ..., X_n) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$

Coefficient β_i is the change in the probably that $Y = 1$ associated with a one-unit change in X_i, holding the other regressors constant. They are estimated using OLS, keeping in mind OLS standard errors and CIs and hypothesis tests.

b) Because probabilities cannot exceed 1, the probability that $Y = 1$ of a given change in X must be nonlinear. E.g. a change a variable from .1 to .2 may have a big effect on Y, but once the variable is large enough, increasing the variable may not have any effect. In the linear model, the effect of a change in a variable is constant, which leads to predict probabilities that can drop below 0 for very low values of variables and can exceed 1 for high values. However, certain models deal with this misspecifcation, such as the probit and logit regression models. (See S&W pg. 387).

c) $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_n X_{ni} + \mu_i$ Recall the variance of a Bernoulli random variable: $Var(Y) = P(Y = 1) \cdot (1 - P(Y = 1))$ Use to find the conditional variance of the error term: $Var(\mu_i|X_{1i}, ..., X_{ni}) = Var(Y_i - (\beta_0 + \beta_1 X_{1i} + ... + \beta_n X_{ni})|X_{1i}, ..., X_{ni})$

$Var(\mu_i|X_{1i}, ..., X_{ni}) = P(Y_i = 1|X_{1i}, ..., X_{ni}) \cdot (1 - P(Y_i = 1|X_{1i}, ..., X_{ni}))$

$Var(\mu_i|X_{1i}, ..., X_{ni}) = (\beta_0 + \beta_1 X_{1i} + ... + \beta_n X_{ni}) \cdot (1 - \beta_0 - \beta_1 X_{1i} - ... - \beta_n X_{ni})$

$\boxed{Var(\mu_i|X_{1i}, ..., X_{ni}) \neq \sigma_v^2}$

Question 2

a) $\boxed{L(\beta_0, \beta_1) = \prod_{i=1}^{n} F(\beta_0 + \beta_1 X)^{Y_i} \cdot (1 - F(\beta_0 + \beta_1 X))^{1-Y_i}}$

b) $\boxed{log(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} [Y_i log(F(\beta_0 + \beta_1 X)) + (1 - Y_i) log(1 - F(\beta_0 + \beta_1 X))]}$

c) $\partial_{\beta_n} log(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} [Y_i \partial_{\beta_n} log(F(\beta_0 + \beta_1 X)) + (1 - Y_i) \partial_{\beta_n} log(1 - F(\beta_0 + \beta_1 X))]$

For $K = 0, 1$ and $f(x) = \partial_x F(x)$:

$\frac{f(x)}{F(x)} = \frac{\frac{e^{-x}}{(1+e^{-x})^2}}{\frac{1}{(1+e^{-x})}} = \frac{e^{-x}}{1+e^{-x}} = \frac{1}{e^x+1} = F(-x) = 1 - F(x)$

and

$\frac{f(x)}{1-F(x)} = \frac{f(x)}{F(-x)} = \frac{\frac{e^{-x}}{(1+e^{-x})^2}}{\frac{1}{1+e^x}} = \frac{e^x+1}{(1+e^x)^2} = \frac{1}{1+e^{-x}} = F(x)$

$\therefore \partial_{e_n} log(L(\beta_x, \beta_1) = \sum_{i=1}^{n} \partial_{\beta_x}\beta_0 + \beta_1 X[Y_i(1 - F(\beta_0 + \beta_1 X) - (1 - X)F(\beta_0 + \beta_1 X)]$

$= \sum_{i=1}^{n} \partial_{\beta_x}\beta_0 + \beta_1 X[Y_i - Y_i F(\beta_0 + \beta_1 X - F(\beta_0 + \beta_1 X) + Y_i F(\beta_0 + \beta_1 X)]$

$= \sum_{i=1}^{n} \partial_{\beta_x}\beta_0 + \beta_1 X[Y_i - F(\beta_0 + \beta_1 X)]$

$\therefore$ First Order Conditions = $\boxed{\sum_{i=1}^{n} \hat{\epsilon_i} = 0, \sum_{i=1}^{n} X_i \hat{\epsilon_i} = 0}$

where $\hat{\epsilon_i} = Y_i - F(\hat{\beta_0} + \hat{\beta_1} X)$

d) The first order condiitons in the model have same structure. With $\hat{\mu_i}$ and OLS residuals:

$$\boxed{\sum_{i=1}^{n} \hat{\mu_i} = 0, \sum_{i=1}^{n} X_i \hat{\mu_i} = 0}$$

Question 3

a) Probit model for Y using $(X_1, ..., X_4)$ models the probability that Y=1 using the cumulative standard normal distribution function. The model must follow the following:

i. $P(Y = 1|X)$ is increasing in X for $\beta_1 > 0$
ii. $0 \leq P(Y = 1|X) \leq 1$ for all x.

The parameter of interest is how different forms of prenatal care effect birthweight (e.g. the receiving of prenatal care, smoking during pregnancy, drinking, age of mother, etc.)

b) $CDF : \Phi(Z)$ $Z = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$ $P(Y = 1|X_1, ..., X_4) = \Phi(\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i})$

We could collect data through appropriate methods and conduct a regression. We could after solve for the CDF of Z ($\Phi(Z)$). This would be $P(Y = 1|X = x)$=area under the standard normal density curve to the left of Z.

Estimating, nonlinear least squares extends the concepts of OLS to models in which parameters enter nonlinearity.

$\min_{b_0, b_1, b_2, b_3, b_4} \sum_{i=1}^{n} (Y_i - \Phi(b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i}))^2$

This can be solved using software such as R.

c)

I would test the multicolinearity between prenatal care and smoking during pregnancy. To modify this, we can create a variable of interaction using regression software. We want to test the interaction of these two variables.

Here, this is ia case where $\frac{\Delta Y_1}{\Delta X_1}$ depends on X_2.

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 (X_1 \times X_2) + \mu_i$

We can isolate the variable of interest such that:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 (X_1 \times X_2) + \mu_i$

And we can compare:

a) $E(Y_i|X_1 = 0, X_2 = x_2) = \beta_0 + \beta_2 x_2$

b) $E(Y_i|X_1 = 1, X_2 = x_2) = \beta_0 + \beta_1 + \beta_2 x_2 + \beta_3 x_2$

$(a) - (b) = \beta_1 + \beta_5 x_2$

d) If we were to use the model from part b, we would create separate regression models and potentially reject the joint null hypothesis but not either individual null hypothesis, which would imply that our friend is correct. If, however, we reject both the joint and the two individual nulls, our friend is wrong. This is all referring to the coefficient of interaction.

Question 4

a)

```
mean(AmEx$cardhldr)
```

```
## [1] 0.7828467
```

78.28% are cardholders, and this sample is not representative of the greater US population.

b)

```
lpm1<-lm(cardhldr~income+age+selfempl+ownrent+acadmos, data=AmEx)
selpm1 <- sqrt(diag(vcovHC(lpm1)))
logit1<-glm(cardhldr~income+age+selfempl+ownrent+acadmos, family = binomial(link="logit"), data=AmEx)
selogit1 <- sqrt(diag(vcovHC(logit1)))
probit1<-glm(cardhldr~income+age+selfempl+ownrent+acadmos, family = binomial(link="probit"), data=AmEx)
seprobit1 <- sqrt(diag(vcovHC(probit1)))
```

c)

```
age <- AmEx$age
income <- AmEx$income
selfempl <- AmEx$selfempl
ownrent <- AmEx$ownrent
acadmos <- AmEx$acadmos
cpredictdata <- data.frame(age=as.vector(quantile(age,.5)),
                           income=as.vector(quantile(income,.5)),selfempl=as.numeric(0),
                           ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdataif1<-data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.5)),selfempl=as.numeric(1),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.
#predict1 <-

  predict.lm(lpm1,cpredictdataif1,type="response")-
  predict.lm(lpm1,cpredictdata,type="response")
```

```
##          1
## -0.1019954
```

```
cpredictdata2 <- data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.2)),selfempl=as.numeric(0),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.
cpredictdata2if1<-data.frame(age=as.vector(quantile(age,.5)),
                             income=as.vector(quantile(income,.2)),selfempl=as.numeric(1),
                             ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,
#predict2 <-
  predict.lm(lpm1,cpredictdata2if1,type="response")-
  predict.lm(lpm1,cpredictdata2,type="response")
```

```
##          1
## -0.1019954
```

```
cpredictdata3 <- data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.8)),selfempl=as.numeric(0),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdata3if1<-data.frame(age=as.vector(quantile(age,.5)),
                             income=as.vector(quantile(income,.8)),selfempl=as.numeric(1),
                             ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,
#predict3 <-
  predict.lm(lpm1,cpredictdata3if1,type="response")-
  predict.lm(lpm1,cpredictdata3,type="response")
```

```
##          1
## -0.1019954
```

```
cpredictdata <- data.frame(age=as.vector(quantile(age,.5)),
                           income=as.vector(quantile(income,.5)),selfempl=as.numeric(0),
                           ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdataif1<-data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.5)),selfempl=as.numeric(1),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
#predict4 <-
  predict.glm(logit1,cpredictdataif1,type="response")-
  predict.glm(logit1,cpredictdata,type="response")
```

```
##          1
## -0.1222479
```

```
cpredictdata <- data.frame(age=as.vector(quantile(age,.5)),
                           income=as.vector(quantile(income,.5)),selfempl=as.numeric(0),
                           ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdataif1<-data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.5)),selfempl=as.numeric(1),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.
#predict5 <-
  predict.glm(logit1,cpredictdataif1,type="response")-
  predict.glm(logit1,cpredictdata,type="response")
```

```
##          1
## -0.1222479
```

```
cpredictdata2 <- data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.2)),selfempl=as.numeric(0),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.
cpredictdata2if1<-data.frame(age=as.vector(quantile(age,.5)),
                             income=as.vector(quantile(income,.2)),selfempl=as.numeric(1),
                             ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,
#predict6 <-
  predict.glm(logit1,cpredictdata2if1,type="response")-
  predict.glm(logit1,cpredictdata2,type="response")
```

```
##          1
## -0.1316777
```

```
cpredictdata <- data.frame(age=as.vector(quantile(age,.5)),
                           income=as.vector(quantile(income,.5)),selfempl=as.numeric(0),
                           ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdataif1<-data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.5)),selfempl=as.numeric(1),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
#predict7 <-
  predict.glm(probit1,cpredictdataif1,type="response")-
  predict.glm(probit1,cpredictdata,type="response")
```

```
##          1
## -0.1208986
```

```
cpredictdata2 <- data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.2)),selfempl=as.numeric(0),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdata2if1<-data.frame(age=as.vector(quantile(age,.5)),
                             income=as.vector(quantile(income,.2)),selfempl=as.numeric(1),
                             ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,
#predict8 <-
  predict.glm(probit1,cpredictdata2if1,type="response")-
  predict.glm(probit1,cpredictdata2,type="response")
```

```
##          1
## -0.1273752
```

```
cpredictdata3 <- data.frame(age=as.vector(quantile(age,.5)),
                            income=as.vector(quantile(income,.8)),selfempl=as.numeric(0),
                            ownrent=as.vector(quantile(ownrent,0)),acadmos=as.vector(quantile(acadmos,.5
cpredictdata3if1<-data.frame(age=as.vector(quantile(age,.5)),
                             income=as.vector(quantile(income,.8)),selfempl=as.numeric(1),
                             ownrent=as.vector(quantile(ownrent,0)),acadmos=
                               as.vector(quantile(acadmos,.5)))
#predict9 <-
  predict.glm(probit1,cpredictdata3if1,type="response")-
  predict.glm(probit1,cpredictdata3,type="response")
```

```
##          1
## -0.1068485
```

Stargazer output:

LPM 50%	Logit 50%	Probit 50%
-0.102	-0.122	-0.121

LPM 80%	Logit 80%	Probit 80%
-0.102	-0.102	-0.107

LPM 20%	Logit 20%	Probit 20%
-0.102	-0.132	-0.127

d)

```
lpm2<-lm(cardhldr~income+age+selfempl+ownrent+acadmos+minordrg+majordrg, data=AmEx)
selpm2 <- sqrt(diag(vcovHC(lpm2)))
logit2<-glm(cardhldr~income+age+selfempl+ownrent+acadmos+minordrg+majordrg, family = binomial(link="logi
selogit2 <- sqrt(diag(vcovHC(logit2)))
probit2<-glm(cardhldr~income+age+selfempl+ownrent+acadmos+minordrg+majordrg, family = binomial(link="pro
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
seprobit2 <- sqrt(diag(vcovHC(probit2)))
```

e)

```
sum((predict(lpm1,type="response")>=.5)&(AmEx$cardhldr==1))+sum((predict(lpm1,type= "response")<.5)&(AmE
```

```
## [1] 9867
```

```
sum((predict(logit1,type="response")>=.5)&(AmEx$cardhldr==1))+sum((predict(logit1,type= "response")<.5)&
```

```
## [1] 9866
```

```
sum((predict(probit1,type="response")>=.5)&(AmEx$cardhldr==1))+sum((predict(probit1,type= "response")<.5
```

```
## [1] 9867
```

```
sum((predict(lpm2,type="response")>=.5)&(AmEx$cardhldr==1))+sum((predict(lpm2,type= "response")<.5)&(AmE
```

```
## [1] 10420
```

```
sum((predict(logit2,type="response")>=.5)&(AmEx$cardhldr==1))+sum((predict(logit2,type= "response")<.5)&
```

```
## [1] 10577
```

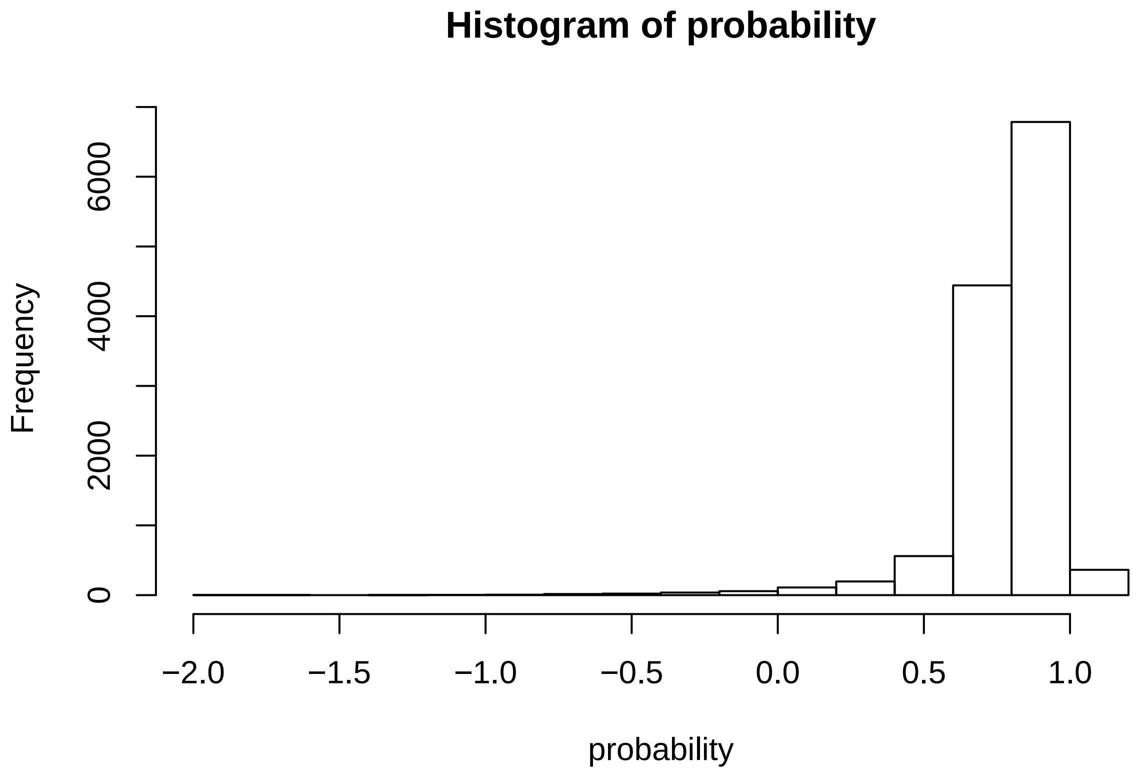
```
sum((predict(probit2,type="response")>=.5)&(AmEx$cardhldr==1))+sum((predict(probit2,type= "response")<.5
```

```
## [1] 10561
```

The highest numbers are logit2 model with 10577 and probit2 model with 10561, so these two models therefore have the highest explanatory power

f)

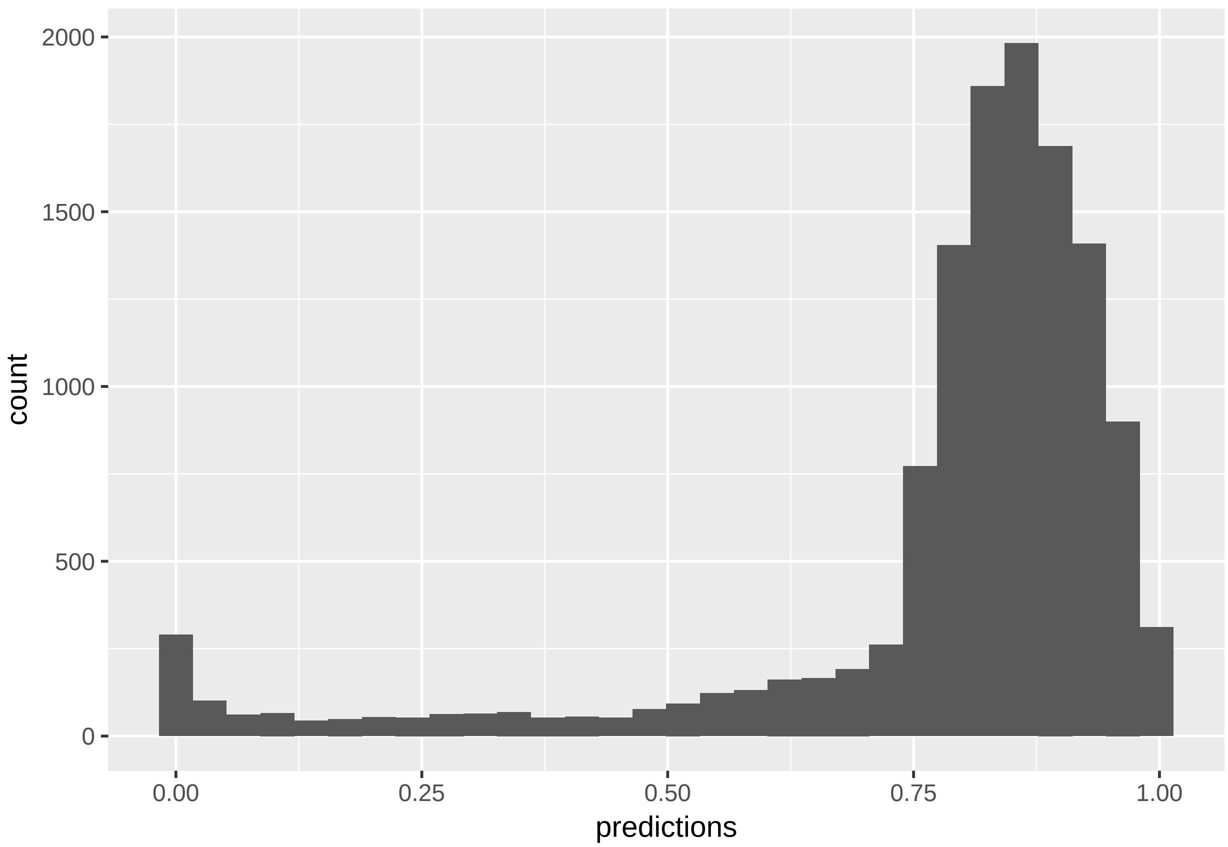
```
probability <- fitted(lpm2)
hist(probability)
```



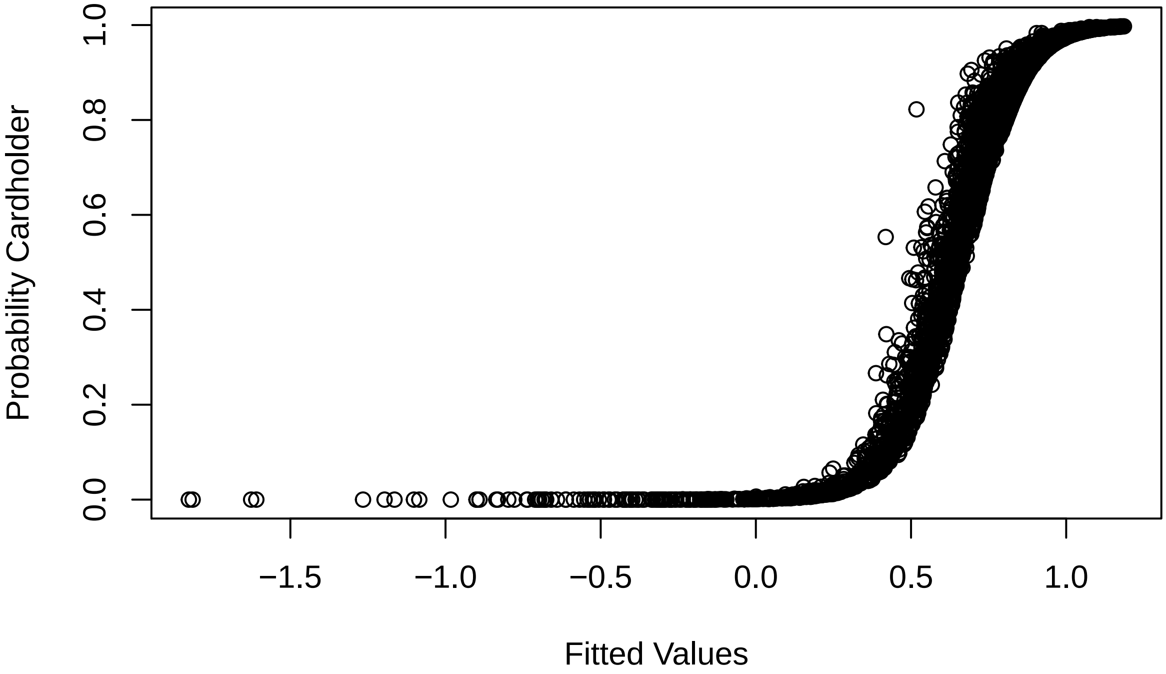
g) The range should be between 0 and 1. 147/12604 predicted values are below zero and 363/12604 are above 1. Because we can't interpret these predicted values as probabilities, the LPM is mis-specified and not the right fit for what we're trying to do.

h)

```
predictions<-predict.glm(logit2,type="response")
qplot(predictions,geom="histogram")
```



```
predictions2<-predict.lm(lpm2,type="response")
plot(predictions2, predictions, xlab = "Fitted Values", ylab = "Probability Cardholder")
```



The above plots indicate 0 to .1 the LPM and Logit models predict the same value but from .1 to .8 the LPM predicts a higher value than the Logit model.

i)

See the following regressors

1) logit2(.25) $-$ 1.234 $= -.3085$ but lpm2 $= -0.117$

2) probit2(.4) $-$ 0.706 $= -.2824$ but lpm2 $= -0.117$

These are not great approximations, but this makes sense because the range of probabilities is much larger than [0.3, 0.7].

Table 4:

	Dependent variable:					
	cardhldr					
	OLS		*logistic*		*probit*	
	(1)	(2)	(3)	(4)	(5)	(6)
income	0.00000***	0.00000***	0.00003***	0.00005***	0.00002***	0.00002***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
age	−0.001***	−0.0002	−0.008***	0.001	−0.004***	0.001
	(0.0004)	(0.0004)	(0.002)	(0.003)	(0.001)	(0.002)
selfempl	−0.102***	−0.100***	−0.585***	−0.770***	−0.348***	−0.445***
	(0.017)	(0.016)	(0.091)	(0.106)	(0.054)	(0.061)
ownrent	0.050***	0.046***	0.289***	0.319***	0.174***	0.191***
	(0.008)	(0.007)	(0.051)	(0.059)	(0.029)	(0.032)
acadmos	0.00002	0.0002***	0.0001	0.002***	0.0001	0.001***
	(0.0001)	(0.0001)	(0.0004)	(0.0005)	(0.0002)	(0.0003)
minordrg		−0.032***		−0.070*		−0.036*
		(0.005)		(0.036)		(0.020)
majordrg		−0.117***		−1.234***		−0.706***
		(0.004)		(0.037)		(0.021)
Constant	0.691***	0.712***	0.603***	0.420***	0.416***	0.361***
	(0.013)	(0.012)	(0.085)	(0.107)	(0.049)	(0.060)
Observations	12,604	12,604	12,604	12,604	12,604	12,604
R^2	0.027	0.212				
Adjusted R^2	0.027	0.211				
Log Likelihood			−6,404.021	−5,066.369	−6,408.237	−5,080.812
Akaike Inf. Crit.			12,820.040	10,148.740	12,828.470	10,177.620
Residual Std. Error	0.407 (df = 12598)	0.366 (df = 12596)				
Note:						*p<0.1; **p<0.05; ***p<0.01