Lecture 10 - Binary Response Models

Data Analysis in CJ (CJUS 6103)

Outline

I. Statistical Properties of Binary Variables
II. The Linear Probability Model
III. Limitations of the Linear Probability Model
IV. The Binary Response Model
V. Predicted Probabilities and Marginal Effects (Under Construction)
VI. More on the Logit Model (Under Construction)
VII. Other Distribution Functions (Under Construction)

Statistical Properties of Binary Variables

For a continuous random variable Y_{ij}, the formal notation for the expected value and variance are:

$$E(Y_i) = \int_{-\inf}^{+\inf} Y_i f(Y_i) dY_i = \mu$$

$$V(Y_i) = E[Y_i - E(Y_i)]^2 = \sigma^2$$

Now assume that we have a discrete random variable Y_i, which is a **dummy variable**. This refers to a special kind of binary variable that meets the following condition:

$$Y_i \in 0, 1$$

Another name for this kind of variable is a **Bernoulli random variable** - a variable with only two possible outcomes, with 1 classifying a *success* and 0 classifying a *failure*. With binary Y_i we can write the expected value and variance in a slightly different way. Use $f(\cdot)$ to represent the probability density function (technically in this case, the probability mass function) for a Bernoulli random variable. We then have the following:

$$\begin{aligned} E(Y_i) &= \sum [Y_i * f(Y_i] && (1)\\ &= 0 * Pf(Y_i = 0) + 1 * Pr(Y_i = 1) && (2)\\ &= Pr(Y_i = 1) && (3)\\ &= \pi && (4) \end{aligned}$$

$$
\begin{aligned}
V(Y_i) &= \sum [Y_i - E(Y_i)]^2 * f(Y_i) && (5)\\
&= (0-\pi)^2 * (1-\pi) + (1-\pi)^2 * \pi && (6)\\
&= (-\pi)(-\pi)(1-\pi) + (1-\pi)(1-\pi)\pi && (7)\\
&= (\pi^2)(1-\pi) + (\pi - \pi^2)(1-\pi) && (8)\\
&= (\pi - \pi^2 + \pi^2)(1-\pi) && (9)\\
&= \pi * (1-\pi) && (10)
\end{aligned}
$$

We will use π to denote $Pr(Y_i = 1)$ which is nothing more than μ for a Bernoulli random variable. Note that π in this sense does not refer to the constant which quantifies the ratio of a circle's circumference to its diameter, but refers to the first moment (arithmetic mean) of the distribution of Y_i.

The Linear Probability Model

Consider applying the ordinary least squares (OLS) estimator to a model in which the dependent variable is binary. This kind of model has a special name - the **linear probability model** (LPM). The population model we are interested in is, as usual:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \tag{11}$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{12}$$

for i=1,...,n respondents and j=1,...,k regressors. Or, for economy, it is also convenient to write:

$$Y_i = X\beta_i + \epsilon_i$$

where $X\beta_i$ is known as the **linear predictor**, or the sum of the intercept and all of the slope X regressor products (this requires a bit of matrix notation that I will not discuss here). Note that each slope represents the change in the $E(Y_i$ for a unit increase in each X_{ij}, holding all other regressors constant. If we take the expectation fo this linear equation, we obtain:

$$E(Y_i \mid X_{i1}, ..., X_{ik}) = \mu \mid X_{i1}, ..., X_{ik}$$

$$V(Y_i \mid X_{i1}, ..., X_{ik}) = \sigma_\epsilon^2$$

But recall from the previous section that, for binary Y_i:

$$E(Y_i) = \pi$$

$$V(Y_i) = \pi * (1 - \pi)$$

Now, if we take the conditional expectations of the LPM:

$$E(Y_i \mid X_{i1}, ..., X_{ik}) = Pr(Y_i = 1 \mid X_{i1}, ..., X_{ik}) \tag{13}$$

$$= \pi \mid X_{i1}, ..., X_{ik} \tag{14}$$

$$= X\beta_i \tag{15}$$

$$V(Y_i \mid X_{i1}, ..., X_{ik}) = Pr(Y_i = 1 \mid X_{i1}, ..., X_{ik}) * [1 - Pr(Y_i = 1 \mid X_{i1}, ..., X_{ik})] \tag{16}$$

$$= \pi \mid X_{i1}, ..., X_{ik} * (1 - \pi \mid X_{i1}, ..., X_{ik}) \tag{17}$$

$$= X\beta_i * (1 - X\beta_i) \tag{18}$$

Each slope still represents the change in $E(Y_i)$ for a unit increase in X_{ij}, holding all other regressors constant. However, because Y_i is binary, each slope now takes on a special meaning - β_j is the mean difference in $Pr(Y_i = 1)$ between subjects who differ by one unit in X_{ij}, holding all other regressors constant. The linear predictor $X\beta_i$ represents the predicted $Pr(Y_i = 1)$ for a given set of values of the regressors $X_{i1}, ..., X_{ij}$.

To provide an example of the linear probability model, we will use data from a study by Apel & Burrow (2011) [1]. The data are from the National Longitudinal Survey of Youth 1997. In this study, they examined what impact youth victimization had on the likelihood of *violent self help*. The key variables we will use for this example are:

Variable Name	Definition
selfhelp	=1 if youth was in a gang, carried a handgun, or assaulted someone
bully	=1 if youth was repeatedly bullied
male	=1 if youth is male
nonwhite	=1 if youth is African American or Latino
grades	middle school grades (1=mostly below Ds; 8=mostly As)
drugs	variety score of substance use (cigarettes, alcohol, or marijuana)
crime	variety score of crime (vandalism, minor theft, major theft, fencing, drug selling)

The dependent variable, *selfhelp*, is measured at the 1998 interview, and references behavior which occurred since the 1997 interview. All of the regressors are measured at the 1997 interview. Let's have a look at the descriptive statistics for each variable:

```
## # A tibble: 7 x 5
##   var      median   max  mean    sd
##   <chr>     <dbl> <dbl> <dbl> <dbl>
## 1 bully         0     1 0.207 0.406
## 2 crime         0     5 0.567 0.865
## 3 drugs         0     3 0.469 0.795
## 4 grades        6     8 5.84  1.34
## 5 male          1     1 0.514 0.500
## 6 nonwhite      0     1 0.463 0.499
## 7 selfhelp      0     1 0.174 0.379
```

As a starting point, we shall estimate an intercept-only model:

$$\text{SelfHelp}_i = \beta_0 + \epsilon_i$$

```
##
## Call:
## lm(formula = selfhelp ~ 1, data = self_help)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1739 -0.1739 -0.1739 -0.1739  0.8261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.173885   0.009712    17.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3791 on 1523 degrees of freedom
```

Notice that the estimate of the intercept is nothing more than the sample mean of *selfhelp*, and the residual standard error is its standard deviation. Recall that this is not a coincidence. Absent any additional

[1] Apel, Robert and John D. Burrow. 2011. Adolescent victimization and violent self-help. *Youth Violence and Juvenile Justice, 9*:112-133

information, the *best guess* for any random variable is the sample mean. Next, the model we will estimate is the following bivariate regression:

$$\text{SelfHelp}_i = \beta_0 + \beta_1 \text{Bully}_i + \epsilon_i$$

```
##
## Call:
## lm(formula = selfhelp ~ bully, data = self_help)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2753 -0.1474 -0.1474 -0.1474  0.8527
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14735    0.01081  13.632  < 2e-16 ***
## bully        0.12797    0.02374   5.391 8.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3757 on 1522 degrees of freedom
## Multiple R-squared:  0.01874,    Adjusted R-squared:  0.01809
## F-statistic: 29.06 on 1 and 1522 DF,  p-value: 8.122e-08
```

Because *bully* and *selfhelp* are both binary, the coefficient represents a contrast in the mean probability of *selfhelp* between youth who are bullied and youth who are not bullied. So you who are bullied have a probability of selfhelp that is 12.8 points higher, and significantly so (p<.001). Furthermore, the intercept in this model represents the mean self-help probability for youth who are not bullied. This means that the probability of self-help among non-bullied youth is 0.147, while the probability of self-help among bullied youth us significantly higher at 0.275 (0.147 + 0.128).

Let's incorporate some additional regressors as control variables. The population model to be estimated is now:

$$\text{SelfHelp}_i = \beta_0 + \beta_1 \text{Bully}_i + \beta_2 \text{Nonwhite}_i + \beta_3 \text{Grades}_i + \beta_5 \text{Drugs}_i + \beta_6 \text{Crime}_i + \epsilon_i$$

```
##
## Call:
## lm(formula = selfhelp ~ bully + male + nonwhite + grades + drugs +
##     crime, data = self_help)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68004 -0.18907 -0.10770 -0.04469  0.97914
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.272911   0.055279   4.937 8.81e-07 ***
## bully        0.073604   0.023551   3.125  0.00181 **
## male         0.023823   0.019332   1.232  0.21803
## nonwhite    -0.008256   0.020101  -0.411  0.68134
## grades      -0.031506   0.007753  -4.064 5.08e-05 ***
## drugs        0.054108   0.013224   4.092 4.51e-05 ***
```

```
## crime        0.063100   0.012300   5.130 3.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3626 on 1517 degrees of freedom
## Multiple R-squared:  0.08915,    Adjusted R-squared:  0.08555
## F-statistic: 24.75 on 6 and 1517 DF,  p-value: < 2.2e-16
```

The coefficient of interest is still the one for bully. It indicates that youth who are bullied are significantly more likely to use self-help later, controlling for other things that are likely to be correlated with both bullying and self-help. Specifically, compared to youth who are not bullied, they have a probability of self-help which is 7.4 points higher, all else equal. Note that other important correlates of self-help are *grades*, *drugs*, and *crime*.

Limitations of the Linear Probability Model

The OLS estimator has a number of properties that can make it less optimal than alternatives when the dependent variable is binary. Here we will consider four potential problems: (1) heteroscedasticity; (2) nonsensical predictions; (3) non-normality; and (4) non-linearity. Each problem will be examine in more detail below, as they concern the following model:

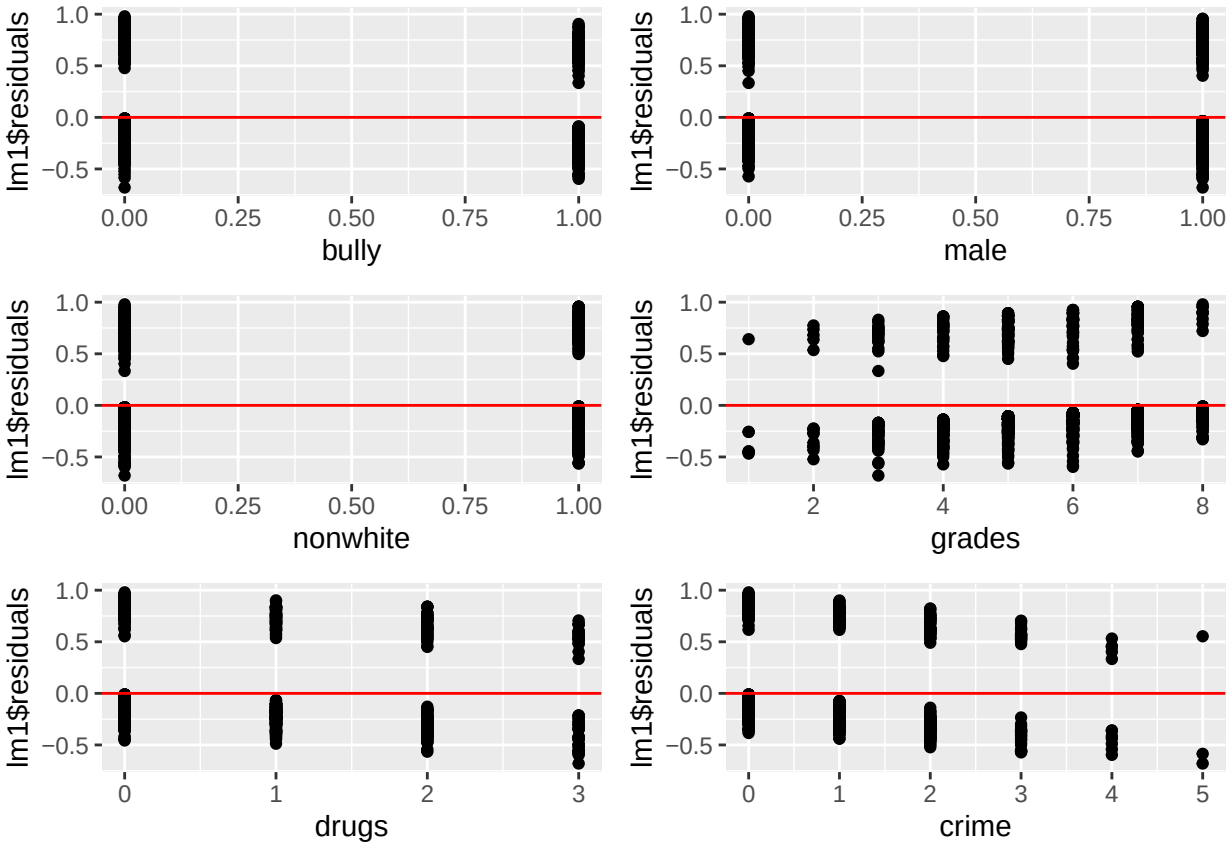
$$\text{SelfHelp}_i = \beta_0 + \beta_1 \text{Bully}_i + \beta_2 \text{Nonwhite}_i + \beta_3 \text{Grades}_i + \beta_5 \text{Drugs}_i + \beta_6 \text{Crime}_i + \epsilon_i$$

Heteroscedastic Errors

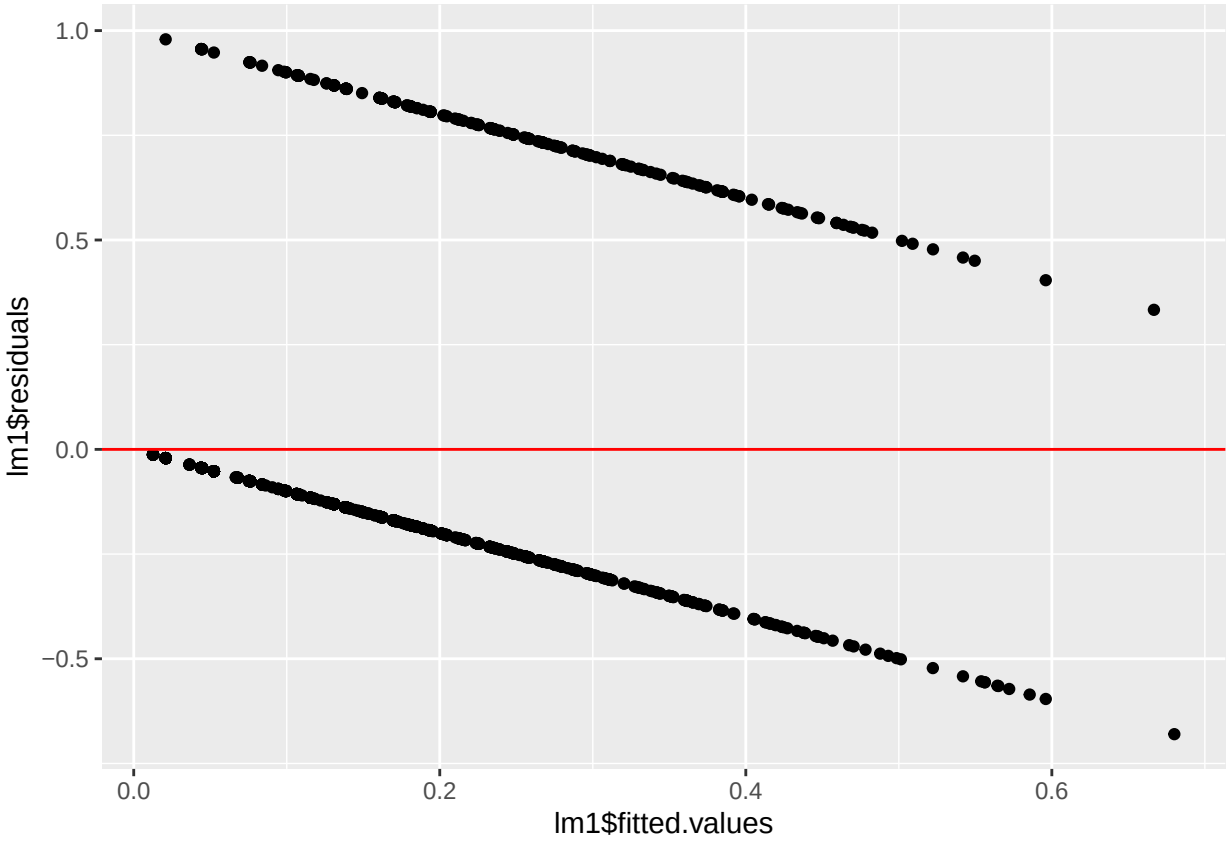
Since we showed that the conditional variance of Y_i depends on the value(s) of the regressor(s), the LPM by definition has heteroscedastic errors. To see this, recall the formula for the conditional variance:

$$V(Y_i \mid X_{i1}, ..., X_{ik}) = X\beta_i \cdot (1 - X\beta_i)$$

Recall that heteroscedasticity means that the OLS estimator is inefficient. In other words, although the coefficients are still unbiased, the standard errors are biased. This can produce misleading tests of statistical significance. Let's observe this by plotting the model residuals against each of the regressors.



We can also plot the residuals, $Y_i - \hat{Y}_i$, against the fitted values, $\hat{Y}_i$:



As if the visual evidence was not enough, a Breusch-Pagan test provides confirmation of heteroscedastic residuals:

```
##
##  studentized Breusch-Pagan test
##
## data:  lm1
## BP = 135.15, df = 6, p-value < 2.2e-16

##
##  Score Test for Heteroskedasticity
##  ---------------------------------
##  Ho: Variance is homogenous
##  Ha: Variance is not homogenous
##
##  Variables: fitted values of selfhelp
##
##          Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    133.2199
##  Prob > Chi2   =    8.093383e-31
```

Notice that, in each plot, the heteroscedasticity is of a peculiar form - it is not in a classic fan shape as we may expect in applications with a continuous dependent variable. The residuals appear to move in an echelon pattern. There appear to be two parallel lines - the top line represents those individuals who have

a value of 1 on *selfhelp*, while the bottom line represents those individuals who have a value of 0. This pattern makes perfect sense when you inspect the x-axis. Someone with a value of 1 on *selfhelp* who has a predicted probability close to 1.0 must have a much smaller residual than a counterpart who has a predicted probability close to 0.0.

Despite the heteroscedasticity, a couple of solutions do exist. One early solution proposed by Goldberger (1964) was to estimate the LPM in three steps [2] First, we need to estimate the LPM as we did above. Second, we need to take the fitted values from this model, and compute a person-specific **weight** in the following manner:

$$W_i = \hat{Y}_i \cdot (1 - \hat{Y}_i)$$

. Third, we need to re-estimate the LMP again, but weighting the regression by the inverse of W_i - a procedure that is known as **weighted least squares** (WLS):

$$\frac{Y_i}{\sqrt{W_i}} = \beta_0^{'} \frac{1}{\sqrt{W_i}} + \cdots + \beta_k^{'} \frac{X_{ik}}{\sqrt{W_i}} + \epsilon_i \frac{1}{\sqrt{W_i}}$$

Before computing the individual weights and re-estimating the model, however, we need to make sure that all fitted values are greater than 0.0 and less than 1.0 (i.e., recode out-of-range values), or else we will end up dividing by zero and getting an error message. To perform WLS on the LPM in R, we can specify Goldberger's weight using the **weights** option in the lm() function.

```
yhat<-lm1$fitted.values
yhat[yhat<=.001]<-.001
yhat[yhat>=.999]<-.999
gb_weight<-yhat*(1-yhat)
```

```
##
## Call:
## lm(formula = selfhelp ~ bully + male + nonwhite + grades + drugs +
##     crime, data = self_help, weights = 1/gb_weight)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4655 -0.4828 -0.3310 -0.1886  6.9031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.242434   0.050692   4.782 1.90e-06 ***
## bully        0.076796   0.025636   2.996 0.002783 **
## male         0.031522   0.016909   1.864 0.062476 .
## nonwhite     0.003087   0.016211   0.190 0.848988
## grades      -0.028630   0.006727  -4.256 2.21e-05 ***
## drugs        0.056800   0.015024   3.781 0.000163 ***
## crime        0.065023   0.014057   4.626 4.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9878 on 1517 degrees of freedom
## Multiple R-squared:  0.1008, Adjusted R-squared:  0.09723
## F-statistic: 28.34 on 6 and 1517 DF,  p-value: < 2.2e-16
```

An alternative, and simpler, approach is to obtain robust (i.e., heteroscedasticity-consistent) standard errors, which have the appeal of adjusting for heteroscedasticity of arbitrary form:

[2] Goldberger, Arthur. 1964. *Econometric Theory*. New York: John Wiley and Sons.

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.2729112  0.0596695  4.5737 5.181e-06 ***
## bully        0.0736042  0.0273701  2.6892 0.0072405 **
## male         0.0238232  0.0194872  1.2225 0.2217062
## nonwhite    -0.0082559  0.0198821 -0.4152 0.6780216
## grades      -0.0315062  0.0082701 -3.8097 0.0001447 ***
## drugs        0.0541084  0.0159202  3.3987 0.0006945 ***
## crime        0.0630998  0.0150215  4.2006 2.816e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's compare results from the three models:

Regressor	LPM w/ No Correction	Goldberger LPM WLS Model	LPM w/Robust Standard Errors
Bullied	.074(.024)**	.077(.026)**	.074(.027)**
Male	.024(.019)	.032(.017)	.024(.019)
Non-White	-.008(.020)	.003(.016)	-.008(.020)
Grades	-.032(.008)***	-.029(.007)***	-.032(.008)***
Drugs	.054(.013)***	.057(.015)***	.054(.016)***
Crime	.063(.012)***	.065(.014)***	.063(.015)***
Constant	.273(.055)***	.242(.051)***	.273(.060)***

*p<.05; **p<.01; ***p<.001

Notice that the tendency is for the standard errors to et a little bit larger when some form of heteroscedasticity adjustment is made, while the coefficients are largely unchanged. Most importantly, the pattern of statistical significance is little affected. In the scheme of things, the mechanical heteroscedasticity of the LPM is the least worrisome shortcoming. Instead, the three remaining limitations pose a bigger problem for the LPM.

Out-of-Range Predictions

The LPM can produce fitted values that are negative or greater than one. In the previous example, we know that all of the fitted values were in the 0-1 interval, because none of them were out of range when we performed Goldberger's WLS (not shown during creation of this variable). Let's add some additional regressors to the model to get a fuller specification:

Variable Name	Definition
age97	age in years at the 1997 interview
hhsize97	number of household residents
bbio97	=1 if youth lives with both biological parents
ccity97	=1 if youth lives in the central city of an MSA
house97	=1 if youth lives in a stable dwelling (e.g., house, condo, farm, ranch)
piat97	achievement score on the mathematics section of the P.I.A.T.
thomewk	number of hours per week spent on homework
schpos	index of positive school attitudes (e.g., "teachers are good")
schprob	variety score of school problem behavior (tardiness, absences, suspensions)

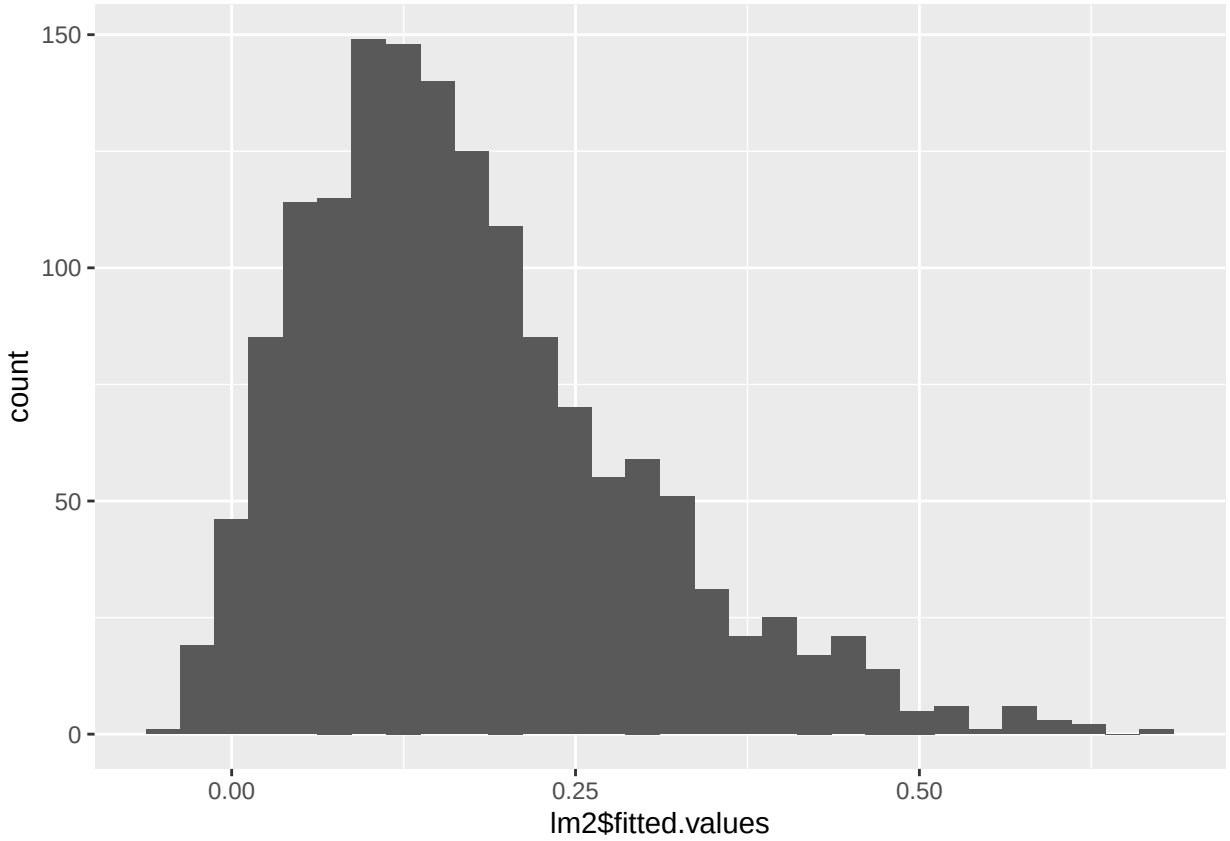
```
## # A tibble: 9 x 5
##   var      median   max   mean     sd
```

```
##   <chr>     <dbl> <dbl>  <dbl>  <dbl>
## 1 age97      12.8  13.6 12.8    0.303
## 2 bbio97      1     1    0.519  0.500
## 3 ccity97     0     1    0.314  0.464
## 4 hhsize97    4    11    4.63   1.47
## 5 house97     1     1    0.793  0.405
## 6 piat97     51.5 100   50.9   33.3
## 7 schpos      5.5   7    5.24   1.38
## 8 schprob     1     3    1.22   0.788
## 9 thomewk     1.5  10    1.87   1.83
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.13594888  0.40408478  0.3364 0.7365885
## bully        0.06542653  0.02750406  2.3788 0.0174932 *
## male         0.02869649  0.01954664  1.4681 0.1422847
## nonwhite    -0.03887894  0.02221210 -1.7503 0.0802614 .
## grades      -0.02617269  0.00852366 -3.0706 0.0021745 **
## drugs        0.05001932  0.01592271  3.1414 0.0017142 **
## crime        0.05515847  0.01521193  3.6260 0.0002974 ***
## age97        0.01220917  0.03143017  0.3885 0.6977351
## hhsize97     0.00685888  0.00680558  1.0078 0.3136970
## bbio97      -0.01559346  0.02019448 -0.7722 0.4401379
## ccity97      0.01321219  0.02140303  0.6173 0.5371271
## house97     -0.02914047  0.02564440 -1.1363 0.2559996
## piat97      -0.00025190  0.00032291 -0.7801 0.4354599
## thomewk      0.00558925  0.00529895  1.0548 0.2916931
## schpos      -0.01334021  0.00777251 -1.7163 0.0863065 .
## schprob      0.03047722  0.01340899  2.2729 0.0231727 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can then summarize and plot the fitted values from this model to check for any out of range predicted values:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.04208  0.08730  0.15025  0.17388  0.23993  0.68040
```



As you can see, we appear to have several observations with a predicted probability of self-help which is below 0. We can flag these cases with a dummy variable just to see how many there are:

```
outrange<-ifelse(lm2$fitted.values<0,1,0)
table(outrange)
```

```
## outrange
##    0    1
## 1486   38
```

So 38 cases (or 2.5% of the sample) have predicted probabilities that are outside the 0-1 interval. Note that out-of-range predicted probabilities are more common when there are many regressors in the model that are not statistically significant. One solution would therefore be to trim the non-significant regressors from the model, although this would be a questionable solution if there are theoretical reasons for retaining the non-significant regressors. A second solution would be to truncate the $\hat{Y} < 0$ at 0.0 and $\hat{Y} > 1$ at 1.0, but this also seems like an arbitrary solution to out-of-range predictions. Just because there are nonsensical predictions does not mean that the LPM is biased. Nevertheless, if one is most interested in obtaining predicted probabilities, the LPM might not be the preferred model.

Residual Non-Normality

Since Y_i can only take on two values (0,1), the residual similarly can only take on two values. Let's start with what we already know about the residual:

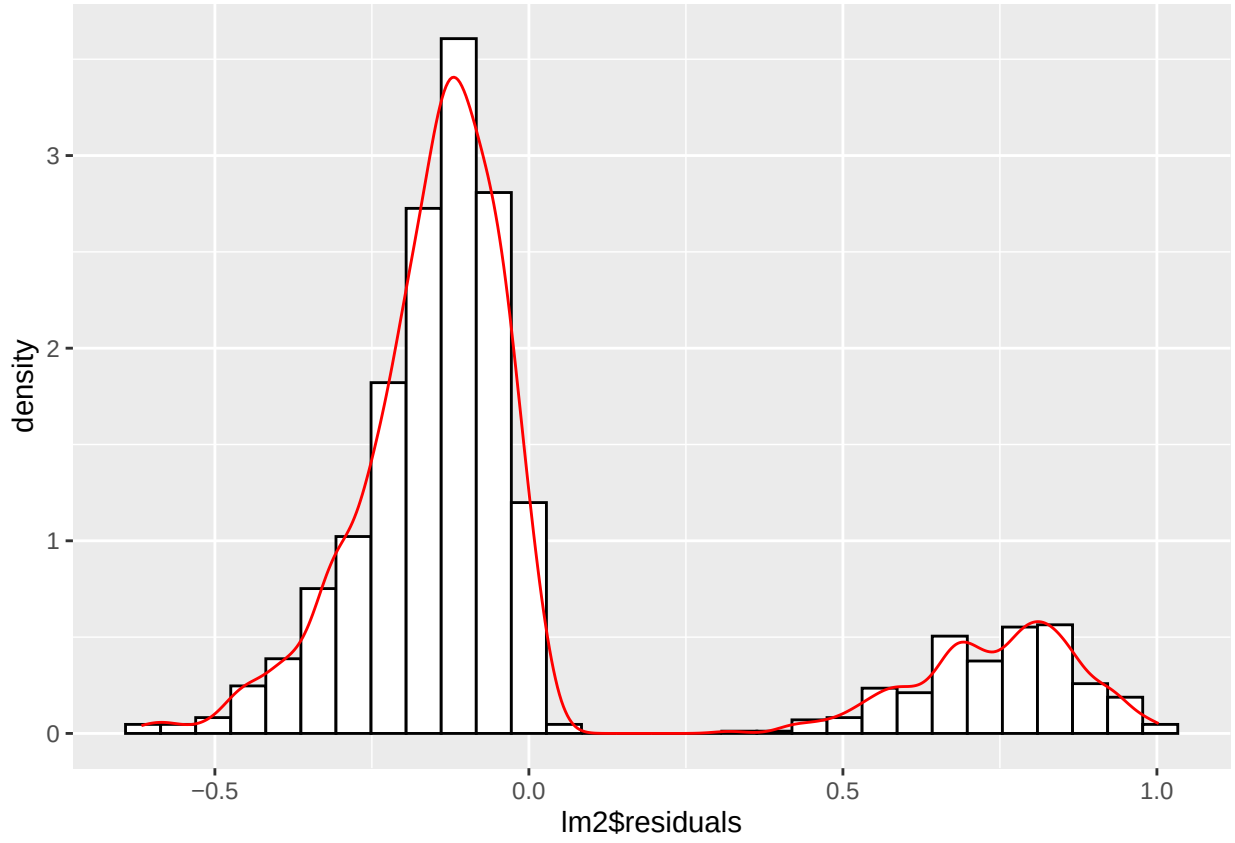
$$\begin{aligned} \epsilon_i &= Y_i - \hat{Y}_i && (19) \\ &= Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik} && (20) \\ & Y_i - X\beta_i && (21) \end{aligned}$$

where $X\beta_i$ is taken to be the linear predictor for a specific subject in the sample. In an LPM, it is necessarily the case that:

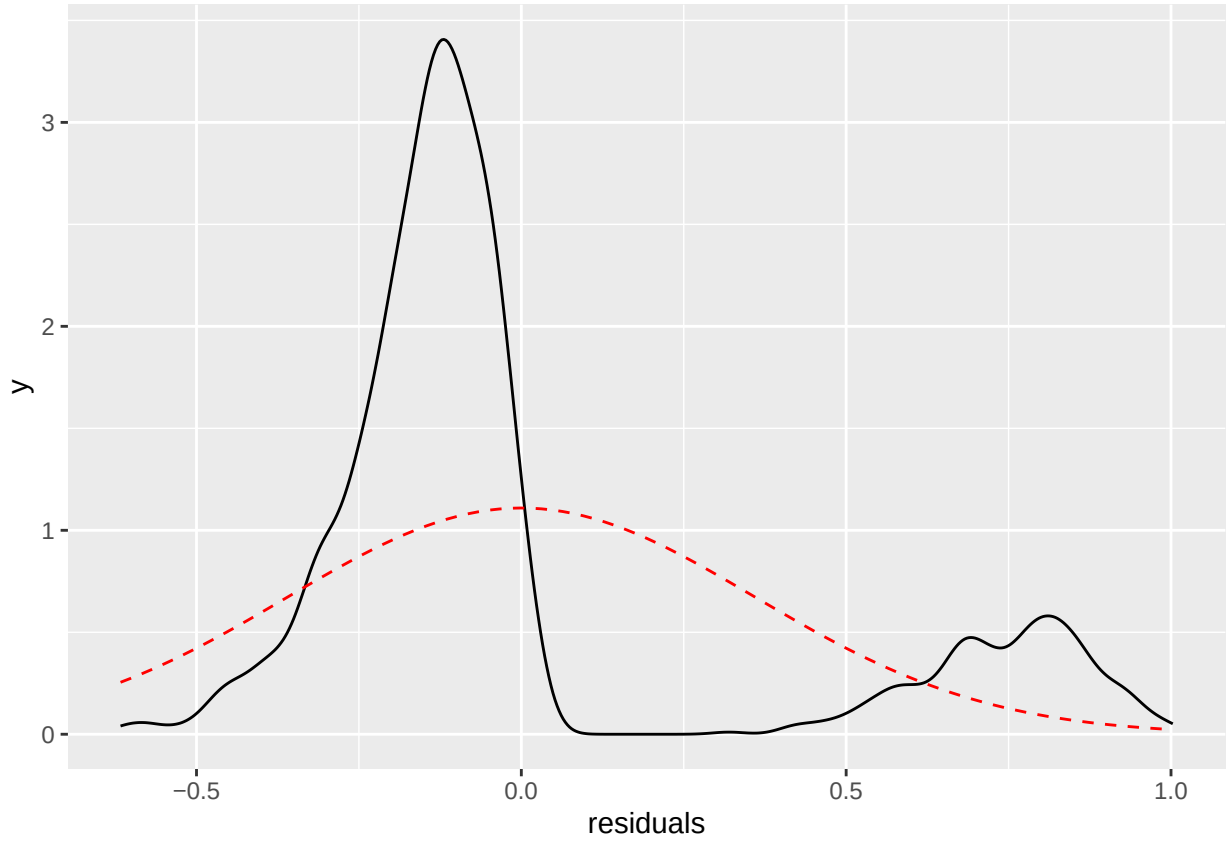
$$\epsilon_i = \begin{cases} -X\beta_i & \text{if } Y_i = 0 \\ 1 - X\beta_i & \text{if } Y_i = 1 \end{cases} \tag{22}$$

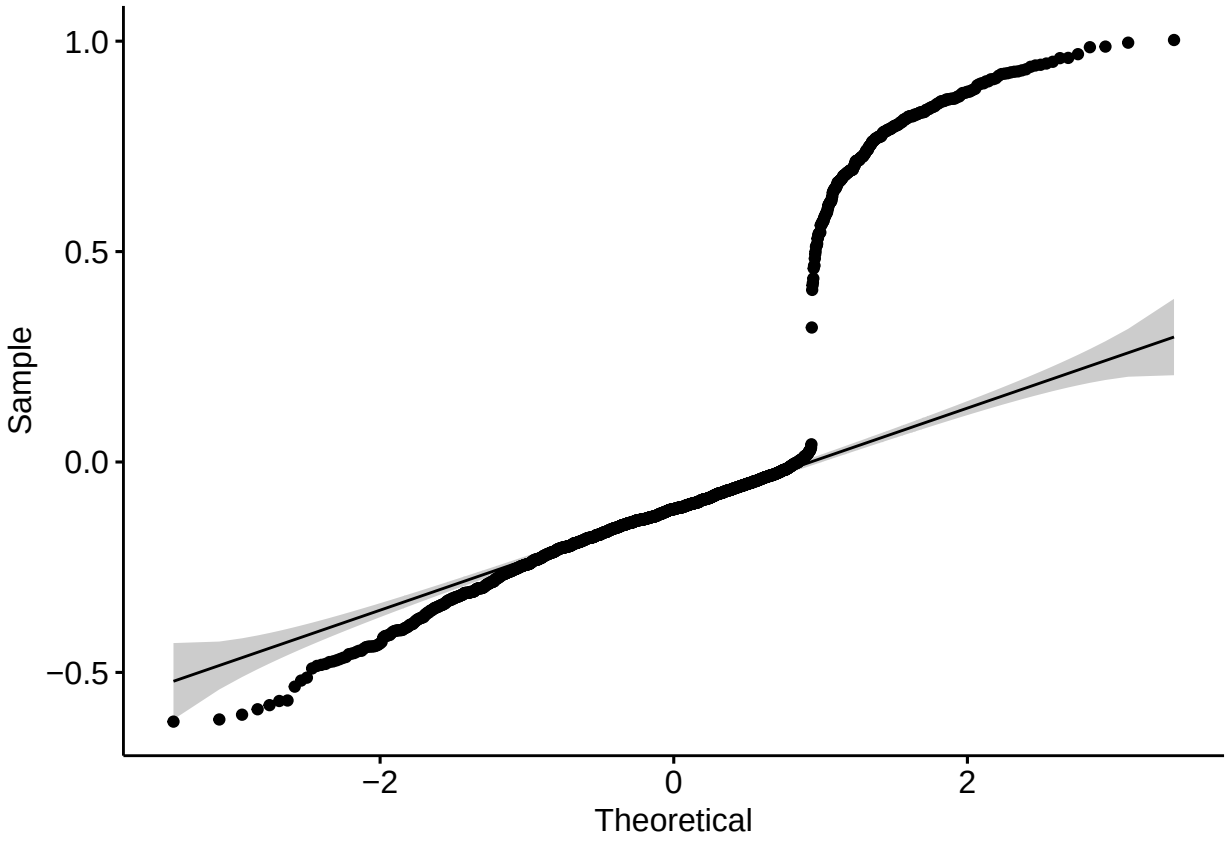
Consequently, the residuals cannot (and will never) be normally distributed). Let's see this empirically:

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



We can confirm the non-normality of the residuals with some diagnostics, although the non-normality is so obvious that diagnostics are practically unnecessary. Let's have a look at a kernel density plot, a standardized normal probability plot, and a Shapiro-Wilk test:






```
##
##  Shapiro-Wilk normality test
##
## data:  lm2$residuals
## W = 0.74876, p-value < 2.2e-16
```

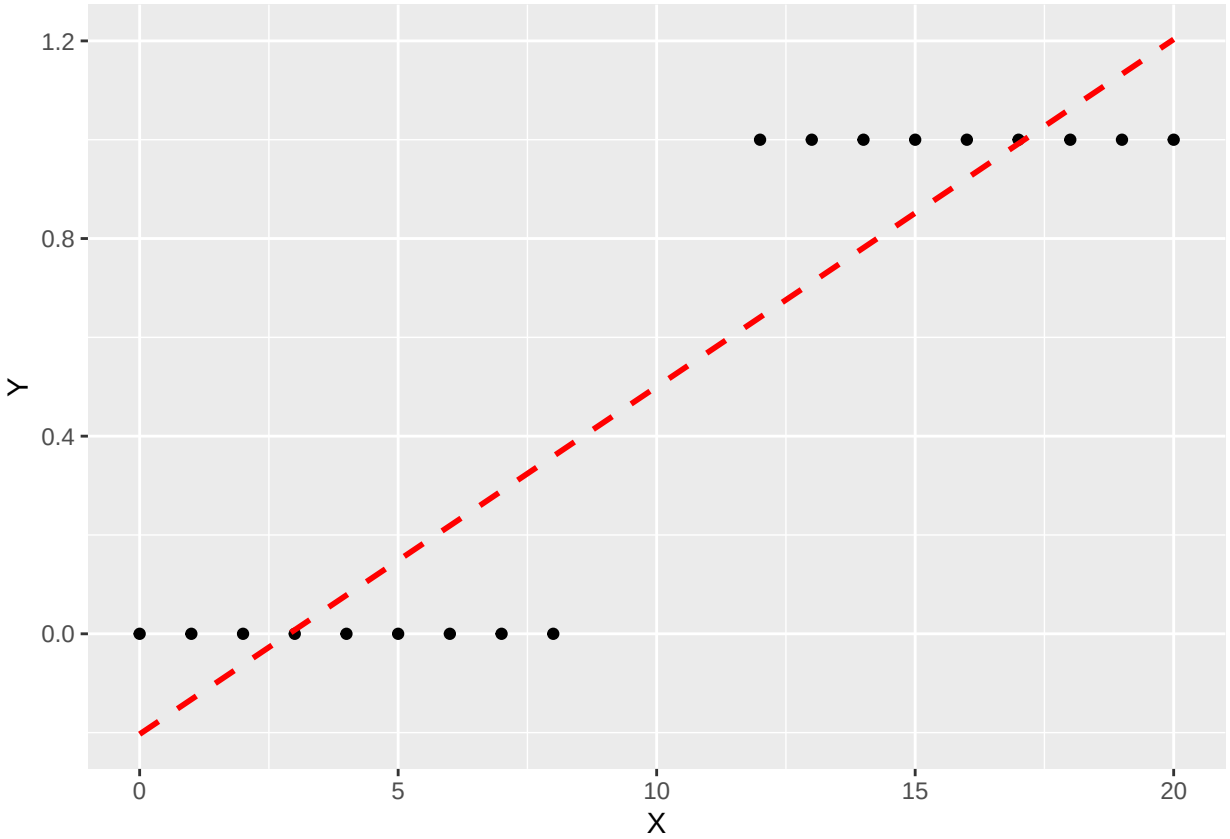
Recall that residual non-normality does not introduce bias into the LPM, it only affects our ability to conduct hypothesis tests. Our tests assume that the residuals are approximately normal to a degree that we can appeal to known probability distributions (i.e., the t- or z-distribution) in order to provide credible p-values for our statistical tests.

Inherent Non-Linearity

Perhaps the most serious shortcoming of the LPM is the inherent non-linearity of models with binary dependent variables. The LPM requires that the effect of X_i on Y_i be constant over the range of X_i. When the outcome is a probability, however, it is reasonable to expect $X_{i} to have some diminishing influence as $Pr(Y_i = 1)$ approaches 0 or 1. This is simply because there is less room for large changes in $Pr(Y_i = 1)$ at the endpoints. This is a very important limitation of the LPM model which can be illustrated by inputting some hypothetical data:

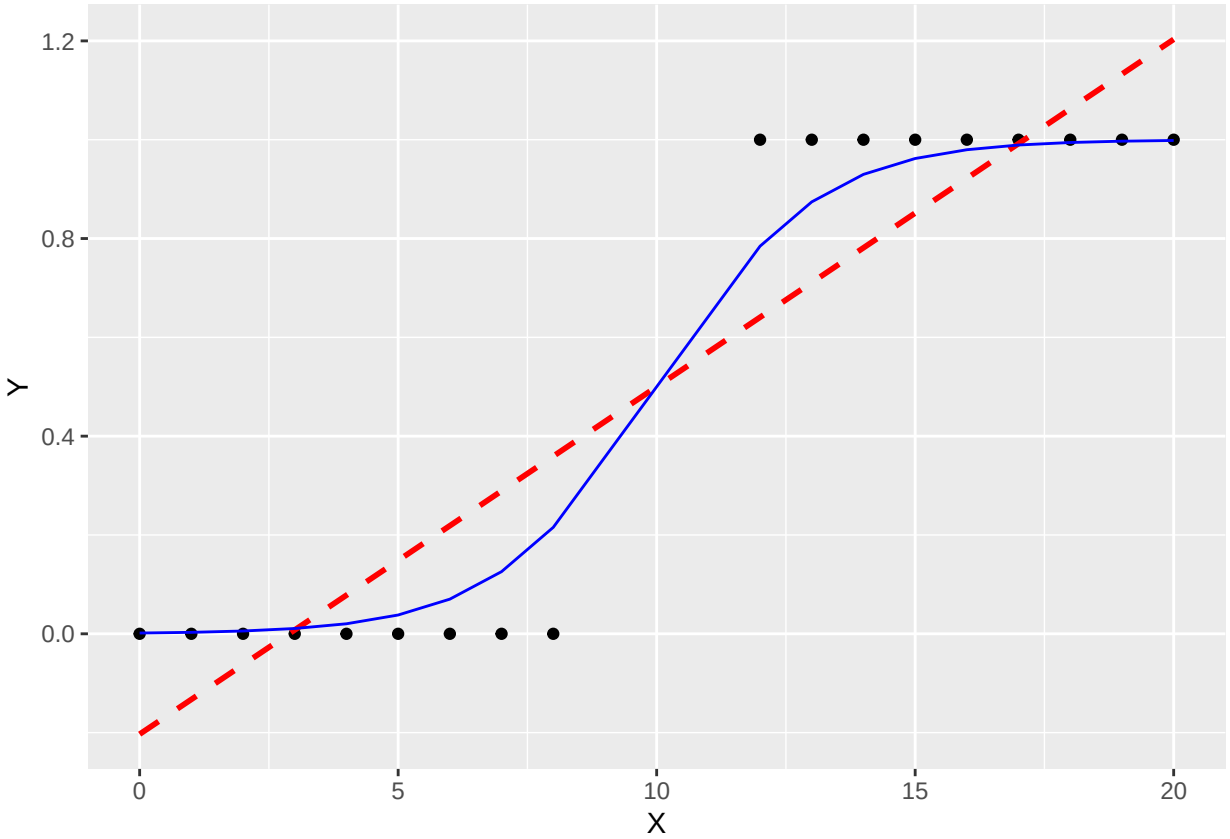
```
Y<-c(0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1)
X<-c(0,1,2,3,4,5,6,7,8,12,13,14,15,16,17,18,19,20)
```

Let's look at a basic scatterplot of these data, and fit a regression line to them:



Notice first that we have the problem of nonsensical prediction here, because really low data points (X_i = 0,1,2) and really high data points (X_i = 18,19,20) have fitted values that are below 0.0 and above 1.0, respectively. We can tell this because the regression line (the dashed red line) dips below 0.0 on the left side of the scatterplot and peaks above 1.0 on the right side. Secondarily, though, wouldn't it be nice if we could make this line bend at the endpoints so that we could provide a better fit to the data points and simultaneously keep the fitted values within the 0-1 interval? It will be necessary to write some additional code to illustrate how we can do this, but it effectively represents a logistic (or log-odds) transformation of Y_i:

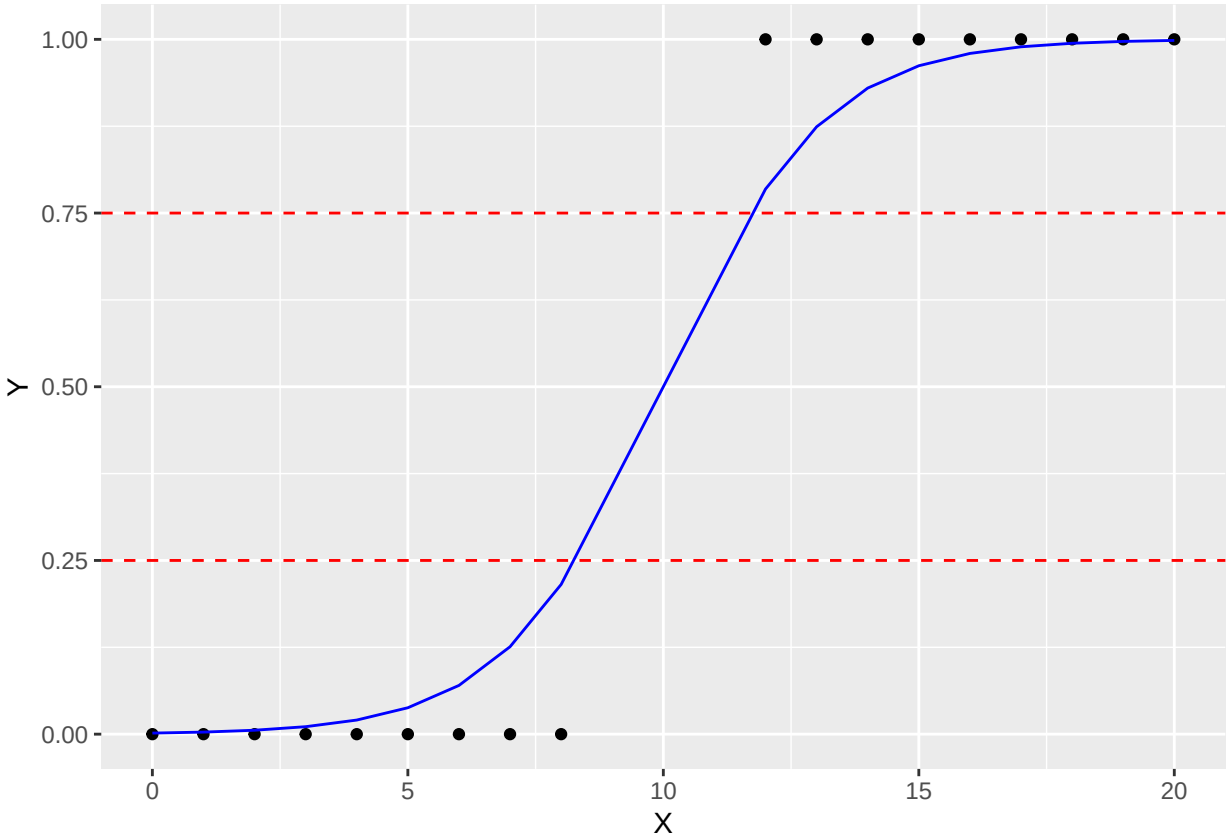
```
newY<-Y
newY[newY<.01]<-.01
newY[newY>.99]<-.99
logitY<-log(newY/(1-newY))
lm3<-lm(logitY~X)
probY<-exp(lm3$fitted.values)/(1+exp(lm3$fitted.values))
```



This is much better. Almost all of the data point fall on the curved line as opposed to the straight line. Additionally, the curved line does not fall below 0.0 or above 1.0, eliminating out-of-range predictions. This line is what we call an S-curve (technically, the curved line is known as a **sigmoid**). To get the line to bend in this manner, we had to perform a transformation which allowed Y_i to be non-linear in X_i. So we had to forgo the simplicity and ease of interpretation of the LPM in order to achieve a better fit to the data points. Details on this (and other) non-linear transformations will be provided in the sections to come.

As the straight line in the figure above indicates, the LPM assumes that X_i has a *constant absolute effect* on $Pr(Y_i = 1)$. In other words, a one-unit increase in X_i produces an increase in $Pr(Y_i = 1)$ that is uniform, regardless of where on the X_i continuum the increase occurs. On the other hand, the line representing the S-curve assumes that X_i has a *constant proportional effect* on $Pr(Y_i = 1)$. It is still a linear model, but we have taken a non-linear transformation of the dependent variable. This allows the absolute increase in $Pr(Y_i = 1)$ for a one-unit increase in X_i to differ depending on where the increase occurs. As shown in the figure, the increase in $Pr(Y_i = 1)$ is smallest at the endpoints of the X_i continuum, and is largest in the middle of the distribution of X_i.

With binary outcomes, then, there tends to be inherent non-linearity that the LPM cannot easily accommodate. This is the single shortcoming that best justifies the adoption of a non-linear probability model. That being said, the LPM should not necessarily be abandoned in every application with a binary dependent variable. Its shortcomings can be overcome by making adjustments such as those considered in the preceding sections (e.g., robust standard errors). The tradeoff is the ease of interpretation of the coefficients in the LPM compared to the models evaluated in the next section. As a general rule. the LMP (with proper adjustments) is a perfectly fine model when the mean of Y_i lies between about 0.25 and 0.75:



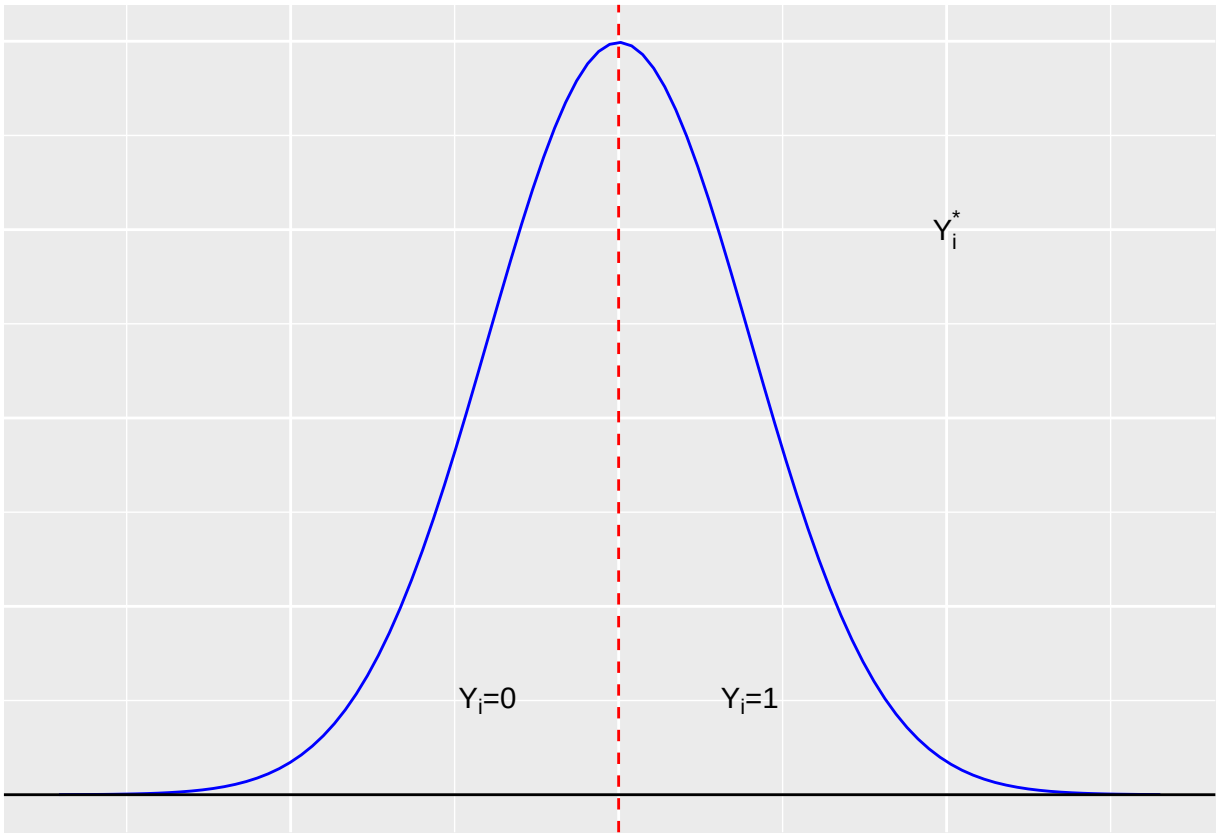
You can see in the figure above that the S-curve is roughly linear between these two probabilities. When the mean of Y_i lies outside the 0.25 - 0.75 interval, on the other hand, it might be worthwhile to consider some of the models explained in the remainder of this document.

The Binary Response Model

Logic of the Binary Response Model

To understand the logic of binary response models, suppose that there exists an underlying response variable Y_i^* that generates the observed (and binary) Y_i. Think of this underlying variable as some kind of **latent propensity** for experiencing the outcome event. This Y_i^* is continuous but unobserved, with a vertical line at some threshold which is often assumed to be 0. What we observe instead is a dummy variable, Y_i, such that:

$$Y_i = \begin{cases} 0 \text{ if } Y_i^* < 0 \\ 1 \text{ if } Y_i^* > 0 \end{cases} \tag{23}$$



Thus, the vertical line represents the threshold on the latent response variable beyond which an observation is *assigned* a value of 1 on the observed response variable. (NOTE: In principle, any value τ may be chosen for this threshold. It can be shown that the only consequence of selecting $\tau \neq 0$ is a rescaling of the intercept in the binary response model.) Notice that the distribution can be shifted to the left or the right depending on the relative number of 0s and 1s in the data. For example, with proportionately more 0s than 1s in the sample, the distribution will be shifted to the left so that there are more negative values underneath the curve.

Regressor variables are incorporated into the latent variable model in the following manner:

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \tag{24}$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{25}$$

$$= X\beta_i + \epsilon_i \tag{26}$$

To economize the notation, $X\beta_i$ will be used to denote the linear predictor for each observation. Notice that this model is actually *linear in the latent variable* whereas it will be non-linear in the observed variable, as we will soon see. The relationship between Y_i and Y_i^* can actually be illustrated in the following way - let's begin with the obvious, the unconditional expectation:

$$E(Y_i) = Pr(Y_i = 1) \tag{27}$$

$$= Pr(Y_i^*) \tag{28}$$

With the inclusion of regressors, the conditional expectation becomes:

$$Pr(Y_i^* \mid X_{i1}, ..., X_{ik} > 0) = Pr(X\beta_i + \epsilon_i > 0) \tag{29}$$

$$= Pr(\epsilon_i - X\beta_i) \tag{30}$$

Since Y_i^* is symmetrical, we can rewrite this:

$$Pr(\epsilon_i > -X\beta_i) = Pr(\epsilon_i < X\beta_i) \tag{31}$$

The inequality on the right hand side is the notation for a cumulative distribution function (c.d.f.) of the residual, evaluated at $X\beta_i$. In other words, the c.d.f. of the residual is the entire area under a continuous curve from $-\inf$ to $X\beta_i$. This is in contrast to the probability density function (p.d.f.), which is the height of a curve evaluated at $X\beta_i$. So the formal notation for the binary response model, with appropriate conditioning on the regressors, is:

$$Pr(Y_i = 1 \mid X_{i1}, ..., X_{ik}) = Pr(\epsilon_i < X\beta_i) \tag{32}$$

$$= F(X\beta_i) \tag{33}$$

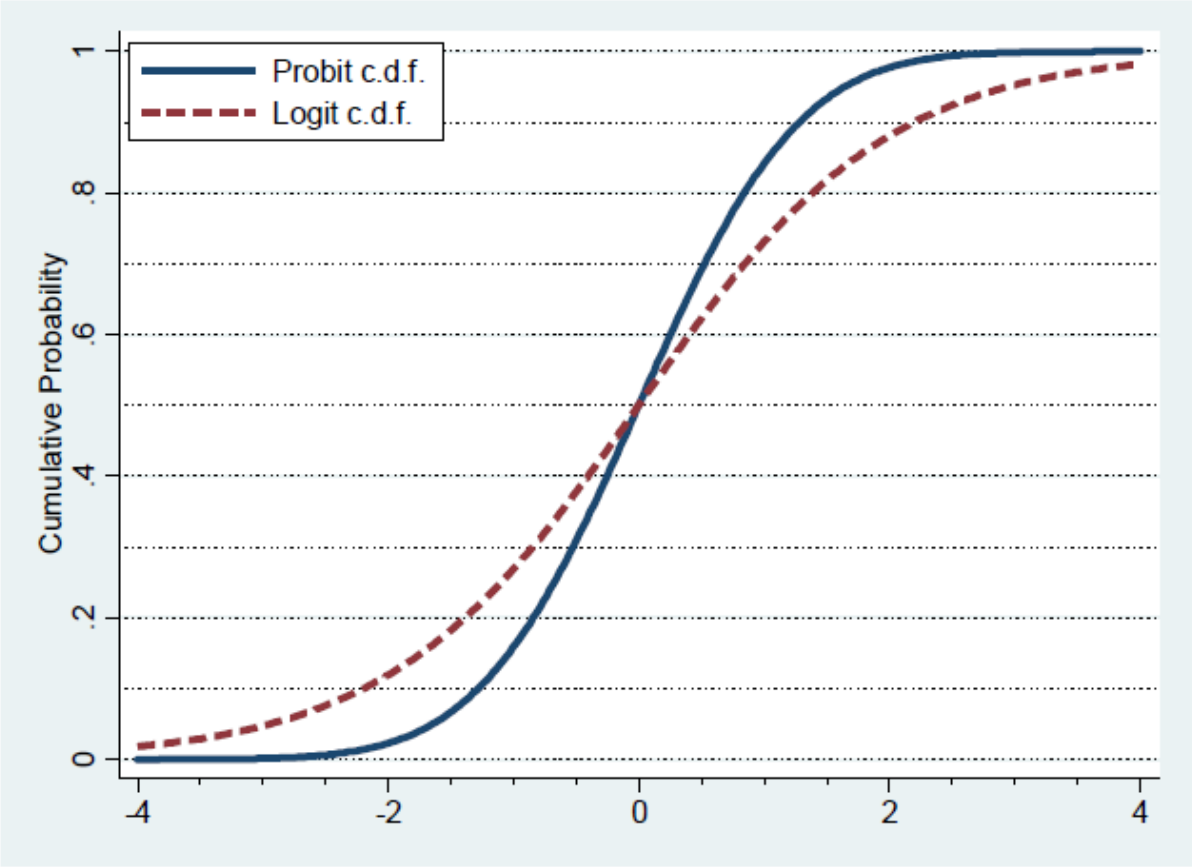
where $F(\cdot)$ is the notation for a c.d.f. Now, there are two useful c.d.f.'s that we may choose - the standard normal distribution function and the logistic distribution function. The standard normal c.d.f. is represented by $\Phi(\cdot)$, and the logistic c.d.f. is represented by $\Lambda(\cdot)$. The former gives rise to the probit model, and the latter gives rise to the logit model. The two c.d.f's can be formalized as follows:

$$\text{Probit}: (Y_i = 1 \mid X_{i1}, ..., X_{ik}) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{X\beta_i} exp(-\frac{1}{2}\epsilon_i) d\epsilon_i$$

$$\text{Logit}: \Lambda(X\beta_i) = \frac{exp(X\beta_i)}{1 + exp(X\beta_i)} \tag{34}$$

$$= \frac{1}{1 + exp(-X\beta_i)} \tag{35}$$

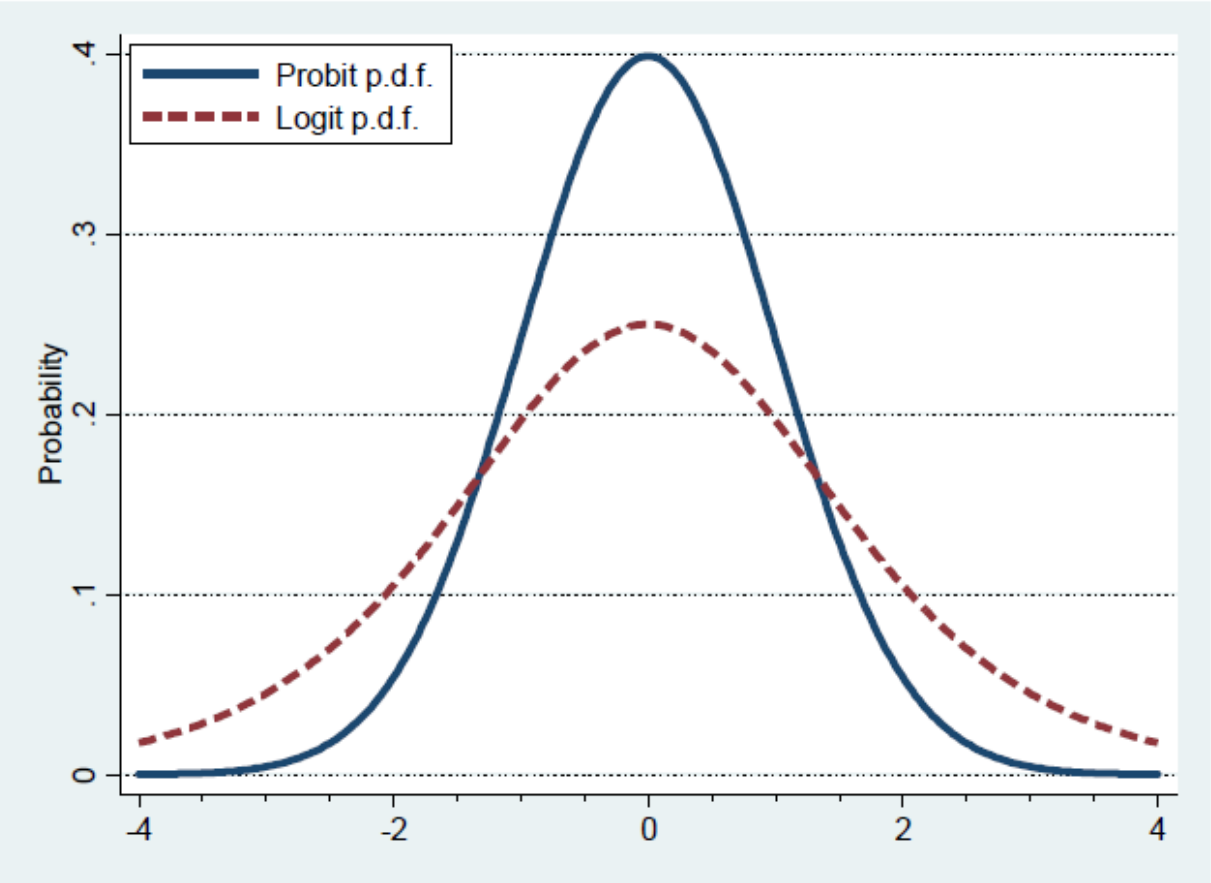
The c.d.f.'s for the probit and logit models look something like this:



Although they are less important for our purposes, the corresponding p.d.f.'s are parameterized and shaped as follows:

$$\text{Probit}: \Phi(X\beta_i) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\epsilon_i\right)$$

$$\text{Logit}: \Lambda(X\beta_i) = \frac{exp(-X\beta_i)}{[1 + exp(-X\beta_i)]^2}$$



Notice that the two distributions are very similar, with the exception that the logistic distribution has somewhat thicker tails, while the density does not peak as high. In practice, the distributions yield essentially the same predicted probabilities, and in fact the point estimates will tend to differ by only a scale factor. In fact, it can be shown:

$$1.60 * \beta_{Probit} \leq \beta_{Logit} \leq 1.81 * \beta_{Probit}$$

$$0.56 * \beta_{Logit} \leq \beta_{Probit} \leq 0.62 * \beta_{Logit}$$

An Empirical Illustration

To see how the binary response model works, let's begin with an empirical illustration in which the response variable is the self-help indicator utilized in earlier examples, *selfhelp*. The frequency distribution for this response variable is:

```
table(self_help$selfhelp)
```

```
##
##    0    1
## 1259  265
```

To acquire some intuition about what the coefficients from a binary response model mean, let's first estimate an intercept-only probit model:

```
summary(glm(selfhelp~1, data=self_help, family=binomial(link='probit')))
```

```
##
## Call:
## glm(formula = selfhelp ~ 1, family = binomial(link = "probit"),
##     data = self_help)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6181  -0.6181  -0.6181  -0.6181   1.8705
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.93893    0.03782  -24.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1408.2  on 1523  degrees of freedom
## Residual deviance: 1408.2  on 1523  degrees of freedom
## AIC: 1410.2
##
## Number of Fisher Scoring iterations: 3
```

Notice that the intercept represents a z-score, which means that we can evaluate the standard normal c.d.f. at the intercept in order to obtain a probability. Let's see what the probability yields:

$$\Phi(-0.939) = 0.174$$

Notice that this is nothing more that the sample mean of *selfhelp*, or the proportion of the sample that reports involvement in self-help behavior. I simply used the pnorm() function (where 0.939 = q) above to produce the cumulative probability at part of the normal curve.

Now let's estimate the intercept-only logit model:

```
summary(glm(selfhelp~1, data=self_help, family=binomial(link='logit')))
```

```
##
## Call:
## glm(formula = selfhelp ~ 1, family = binomial(link = "logit"),
##     data = self_help)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6181  -0.6181  -0.6181  -0.6181   1.8705
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.55834    0.06759  -23.06   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1408.2  on 1523  degrees of freedom
## Residual deviance: 1408.2  on 1523  degrees of freedom
## AIC: 1410.2
##
## Number of Fisher Scoring iterations: 4
```

We can evaluate the cumulative logistic distribution at the intercept in order to obtain the mean self-help probability for the sample. This can be performed by plugging the intercept into the formula:

$$\Lambda(-1.558) = \frac{exp(-1.558)}{1 + exp(-1.558)} = 0.174$$

In R, we can use the invlogit() function from the *boot* package to compute this value directly:

```
inv.logit(-1.558)
```

```
## [1] 0.1739338
```

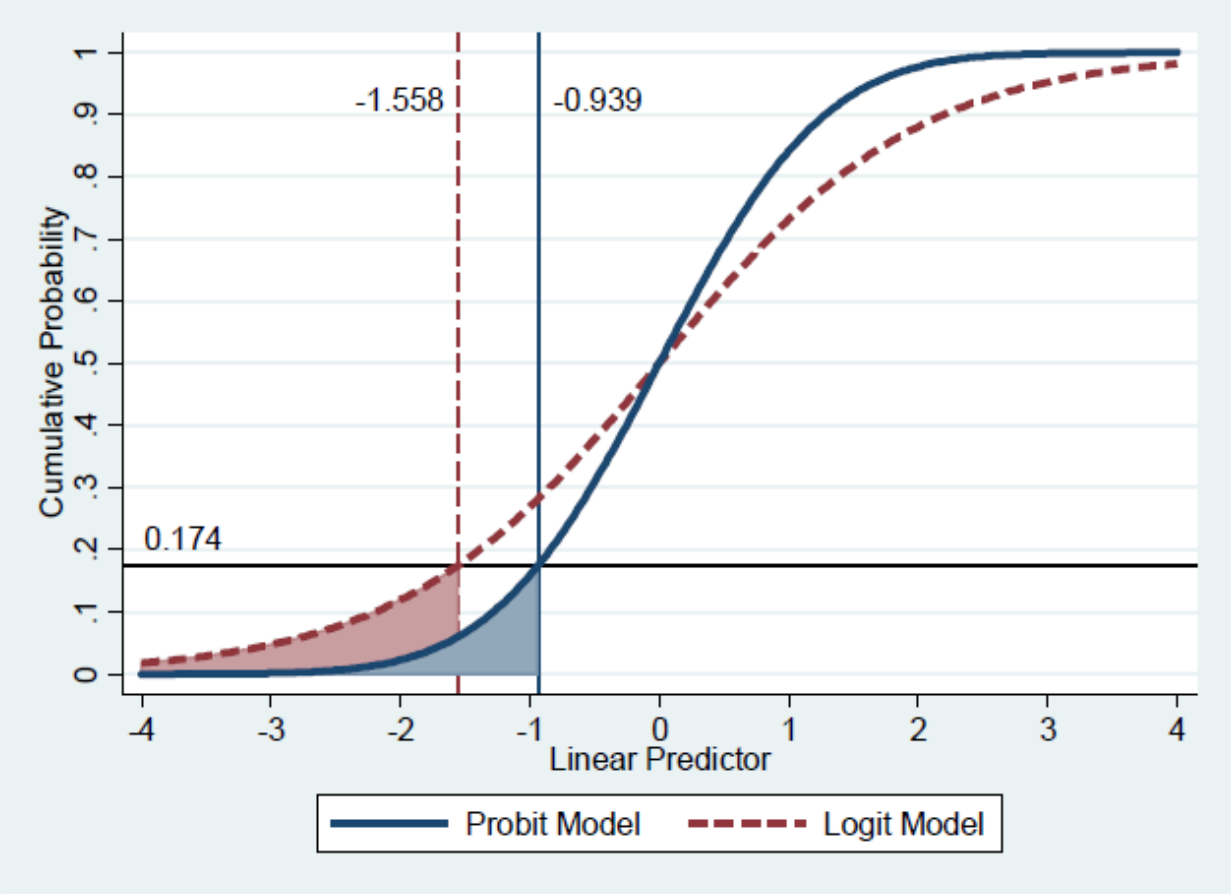
Notice that if we compute the ratio of the probit and logit coefficients, we obtain:

$$\frac{\hat{\beta}_{Probit}}{\hat{\beta}_{Logit}} = \frac{-0.939}{-1.558} = 0.603$$

$$\frac{\hat{\beta}_{Logit}}{\hat{\beta}_{Probit}} = \frac{-1.558}{-0.939} = 1.659$$

Thus, the probit coefficient lies within the 0.56 to 0.62 interval relative to the logit coefficient, whereas the logit coefficient lies within the 1.60 to 1.81 interval relative to the probit coefficient.

We can visualize this probability on both the standard normal c.d.f. and the logistic c.d.f.:



The horizontal line identifies the response probability of 0.174, which we have already obtained by evaluating the standard normal c.d.f. at -0.939 and the logistic c.d.f. at -1.558, both of which are shown as vertical lines in the figure. Note that, when we include a regressor, we are asking how $Pr(Y_i = 1)$ response to an incremental increase in X_i. If the slope corresponding to X_i is positive, the vertical line will be shifted to the right and will thus increase the predicted $Pr(Y_i = 1)$. On the other hand, if the slope corresponding to X_i is negative, the vertical line will be shifted to the left and will thus decrease the predicted $Pr(Y_i)$.

Let's now consider adding a single regressor to the model. We will again examine *bully*, the dummy variable for having been the victim of repeated bullying. But first, let's get a cross-tabulation of *selfhelp* by *bully*:

```
with(self_help, table(selfhelp, bully))
```

```
##          bully
## selfhelp    0    1
##        0 1030  229
##        1  178   87
```

This indicates that the likelihood of self-help among bullied youth is 27.5% ($\frac{87}{229}$) compared to 14.7% ($\frac{178}{1030}$) among non-bullied youth. The probit and logit models should be capable of perfectly replicating these probabilities. The formal equations for the models we are about to estimate is as follows:

$$\Phi^{-1}[Pr(\text{SelfHelp}_i = 1)] = \beta_0 + \beta_1 \text{Bully}_i + \epsilon_i$$

$$\Lambda^{-1}[Pr(\text{SelfHelp}_i = 1)] = \beta_0 + \beta_1 \text{Bully}_i + \epsilon_i$$

```r
probit_bully<-glm(selfhelp~bully, data=self_help, family=binomial(link='probit'))
summary(probit_bully)
```

```
##
## Call:
## glm(formula = selfhelp ~ bully, family = binomial(link = "probit"),
##     data = self_help)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8025  -0.5646  -0.5646  -0.5646   1.9570
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.04786    0.04426 -23.673  < 2e-16 ***
## bully        0.45105    0.08731   5.166 2.39e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1408.2  on 1523  degrees of freedom
## Residual deviance: 1382.0  on 1522  degrees of freedom
## AIC: 1386
##
## Number of Fisher Scoring iterations: 4
```

```r
pnorm(probit_bully$coefficients[1])
```

```
## (Intercept)
##    0.147351
```

```r
pnorm(probit_bully$coefficients[1]+probit_bully$coefficients[2])
```

```
## (Intercept)
##   0.2753165
```

```r
logit_bully<-glm(selfhelp~bully, data=self_help, family=binomial(link='logit'))
summary(logit_bully)
```

```
##
## Call:
## glm(formula = selfhelp ~ bully, family = binomial(link = "logit"),
##     data = self_help)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8025  -0.5646  -0.5646  -0.5646   1.9570
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.75553    0.08117 -21.627  < 2e-16 ***
## bully        0.78772    0.14983   5.257 1.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1408.2  on 1523  degrees of freedom
## Residual deviance: 1382.0  on 1522  degrees of freedom
## AIC: 1386
##
## Number of Fisher Scoring iterations: 4
```

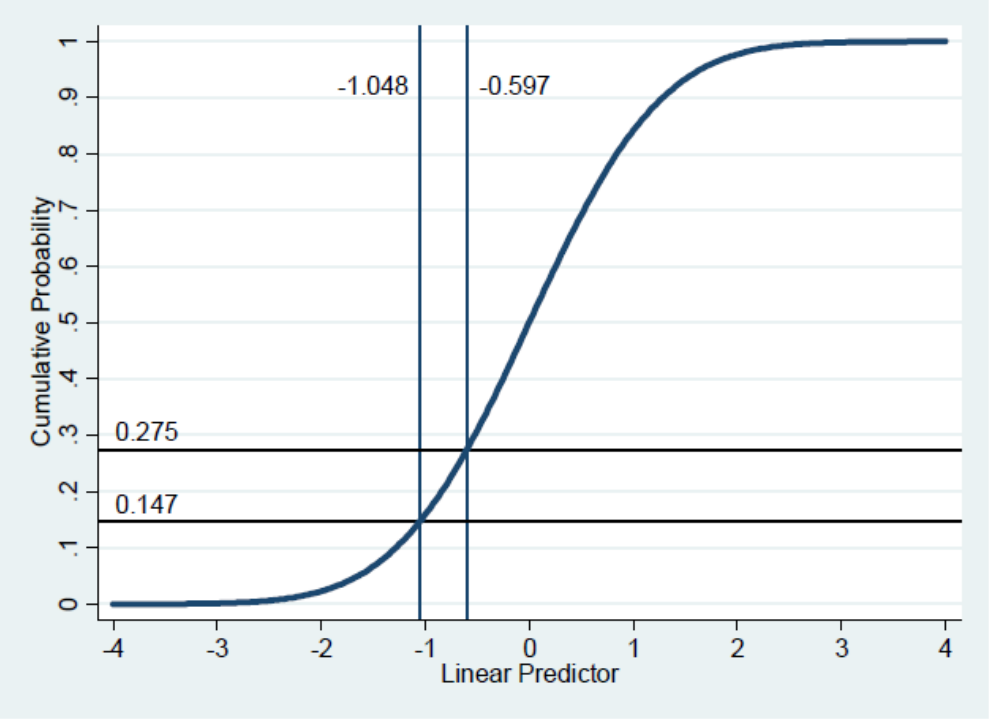
```
inv.logit(logit_bully$coefficients[1])
```

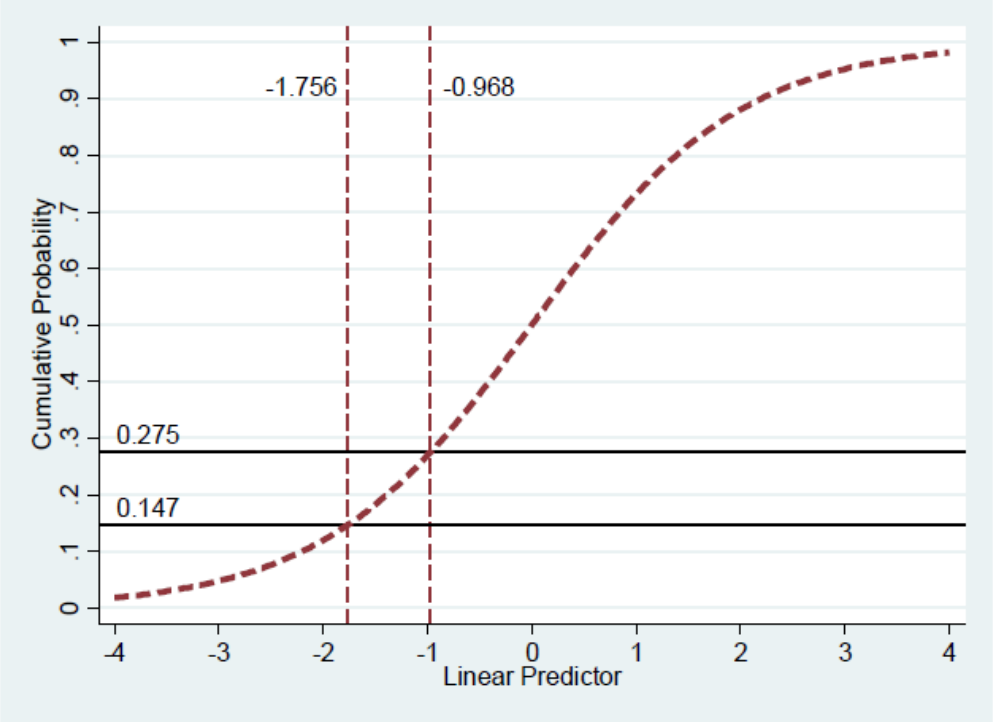
```
## (Intercept)
##    0.147351
```

```
inv.logit(logit_bully$coefficients[1]+logit_bully$coefficients[2])
```

```
## (Intercept)
##   0.2753165
```

Notice that the coefficients from these two models produce virtually identical predicted probabilities of self-help conditional on bullying. Indeed, they perfectly match those produced by the cross-tabulation shown above. We can plot the coefficients and predicted probabilities on the standard normal and logistic distribution functions:





Because the coefficient for bullying is positive, notice that it shifts respondents with *bully*=1 to the right, further along the S-curve. This effectively assigns a higher predicted probability to bullied youth compared to their non-bullied counterparts.

Before we examine the binary response model more closely, let's estimate the full specification with all of the regressors of interest. We'll estimate the probit and logit models in sequences, with the population models and output as follows:

$$\Phi^{-1}[Pr(\text{SelfHelp}_i = 1)] = \beta_0 + \beta_1 \text{Bully}_i + \beta_2 \text{NonWhite}_i + \beta_3 \text{Grades}_i + \beta_5 \text{Drugs}_i + \beta_6 \text{Crime}_i + \epsilon_i$$

$$\Lambda^{-1}[Pr(\text{SelfHelp}_i = 1)] = \beta_0 + \beta_1 \text{Bully}_i + \beta_2 \text{NonWhite}_i + \beta_3 \text{Grades}_i + \beta_5 \text{Drugs}_i + \beta_6 \text{Crime}_i + \epsilon_i$$

```
probit_full<-glm(selfhelp~bully+male+nonwhite+grades+drugs+crime,
                 data=self_help, family=binomial(link='probit'))
summary(probit_full)
```

```
##
## Call:
## glm(formula = selfhelp ~ bully + male + nonwhite + grades + drugs +
##     crime, family = binomial(link = "probit"), data = self_help)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7544  -0.6099  -0.4753  -0.3689   2.4433
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.58234    0.22281  -2.614 0.008960 **
## bully        0.28839    0.09250   3.118 0.001822 **
```

```
## male         0.13169    0.08241   1.598 0.110021
## nonwhite    -0.01076    0.08498  -0.127 0.899247
## grades      -0.13216    0.03188  -4.145 3.39e-05 ***
## drugs        0.19230    0.05092   3.776 0.000159 ***
## crime        0.21216    0.04715   4.499 6.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1408.2  on 1523  degrees of freedom
## Residual deviance: 1280.1  on 1517  degrees of freedom
## AIC: 1294.1
##
## Number of Fisher Scoring iterations: 5
```

```
logit_full<-glm(selfhelp~bully+male+nonwhite+grades+drugs+crime,
                data=self_help, family=binomial(link='logit'))
summary(logit_full)
```

```
##
## Call:
## glm(formula = selfhelp ~ bully + male + nonwhite + grades + drugs +
##     crime, family = binomial(link = "logit"), data = self_help)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7857  -0.6067  -0.4822  -0.3872   2.3830
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.94087    0.38718  -2.430 0.015096 *
## bully        0.49422    0.16012   3.087 0.002025 **
## male         0.22620    0.14808   1.528 0.126633
## nonwhite    -0.02439    0.15140  -0.161 0.872027
## grades      -0.22979    0.05590  -4.111 3.94e-05 ***
## drugs        0.32639    0.08717   3.744 0.000181 ***
## crime        0.35846    0.08048   4.454 8.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1408.2  on 1523  degrees of freedom
## Residual deviance: 1283.9  on 1517  degrees of freedom
## AIC: 1297.9
##
## Number of Fisher Scoring iterations: 4
```

For easier comparison, I'll put these estimates and those from the LPM (with heteroscedasticity-robust standard errors) in one table:

Notice that, in terms of the pattern of statistical significance, the results from all three models are identical. Also notice that the ratios of the logit to probit coefficients are almost all in the 1.60 to 1.81 interval, or to consider the inverse, the ratios of probit to logit coefficients are almost all in the 0.56 to 0.62 interval.

Regressor	LPM	Probit	Logit	b_L/b_P	b_P/b_L
Bullied	.074(.027)**	.288(.092)**	.494(.160)**	1.72	0.59
Male	.024(.019)	.132(.082)	.226(.148)	1.71	0.58
Non-White	-.008(.020)	-.011(.085)	-.024(.151)	2.18	0.46
Grades	-.032(.008)***	-.132(.032)***	-.230(.056)***	1.74	0.57
Drugs	.054(.016)***	.192(.051)***	.326(.087)***	1.70	0.59
Crime	.063(.015)***	.212(.047)***	.358(.080)***	1.69	0.59
Constant	.273(.060)***	-.582(.223)**	-.941(.387)*	1.62	0.62

*p<.05; **p<.01; ***p<.001

Recall that the coefficients from the LPM are interpreted as the difference in $Pr(Y_i = 1)$ between two hypothetical subjects who differ by one unit in X_{ij}. So if we take the regressor *bully* we see that youth who have been repeatedly bullied exhibit a self-help probability that is 7.4 points higher than youth who have not been bullied. The coefficients from the probit and logit models, on the other hand, are interpreted as the impact of a unit increase in the regressor on the latent, continuous response variable Y_i^* (not on the observed, discrete response variable Y_i). So in the probit model, youth who were bullied have a value on latent self-help that is 0.288 unites higher than non-bullied youth, and the corresponding difference in the logit model is 0.494. The problem is that these values, and the very notion of *latent self-help* are meaningless in practical terms. We need some way to transform the coefficients into a more meaningful metric. Predicted probabilities and marginal effects are just such transformations.

Predicted Probabilities and Marginal Effects

Recall the formula to estimate the response probability in a binary response model:

$$Pr(Y_i = 1 \mid X_{i1}, ..., X_{ik}) = F(X\beta_i)$$

where $X\beta_i$ is the linear predictor and $F(\cdot)$ denotes either the standard normal of logistic distribution function:

$$\text{Probit Predicted Probability: } (Y_i = 1 \mid X_{i1}, ..., X_{ik}) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{X\beta_i} exp(-\frac{1}{2}\epsilon_i) d\epsilon_i$$

$$\text{Logit Predicted Probability: } (Y_i = 1 \mid X_{i1}, ..., X_{ik}) = \frac{exp(X\beta_i)}{1 + exp(X\beta_i)}$$

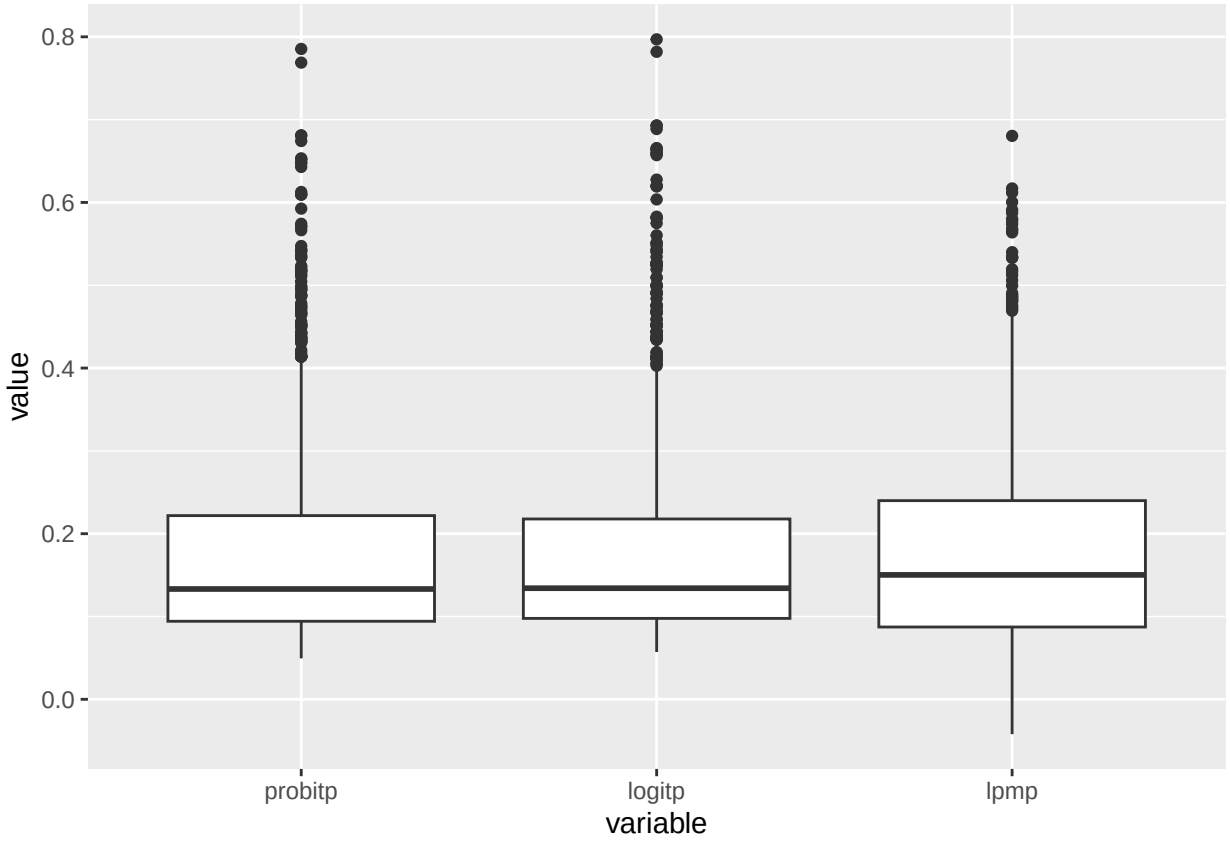
These formulas allow us to estimate a predicted probability for each respondent in the sample based on his or her actual values for each of the regressors, or otherwise to estimate a predicted probability for a hypothetical respondent with values for each of the regressors specified in the model. Let's begin by inspecting the predicted probabilities from the probit model, the logit model, and for comparative purposes, the linear model (LPM). We will use the fully specified models we estimated at the end of the last section.

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## i Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 7 x 5
##   var        mean    sd     min   max
##   <chr>     <dbl> <dbl>   <dbl> <dbl>
## 1 logitp    0.174 0.116  0.0571 0.797
## 2 logitxb  -1.72  0.722 -2.80   1.37
## 3 lpmp      0.174 0.120 -0.0421 0.680
## 4 lpmxb     0.174 0.120 -0.0421 0.680
## 5 probitp   0.173 0.117  0.0494 0.785
## 6 probitxb -1.02  0.423 -1.65   0.791
## 7 selfhelp  0.174 0.379  0      1
```

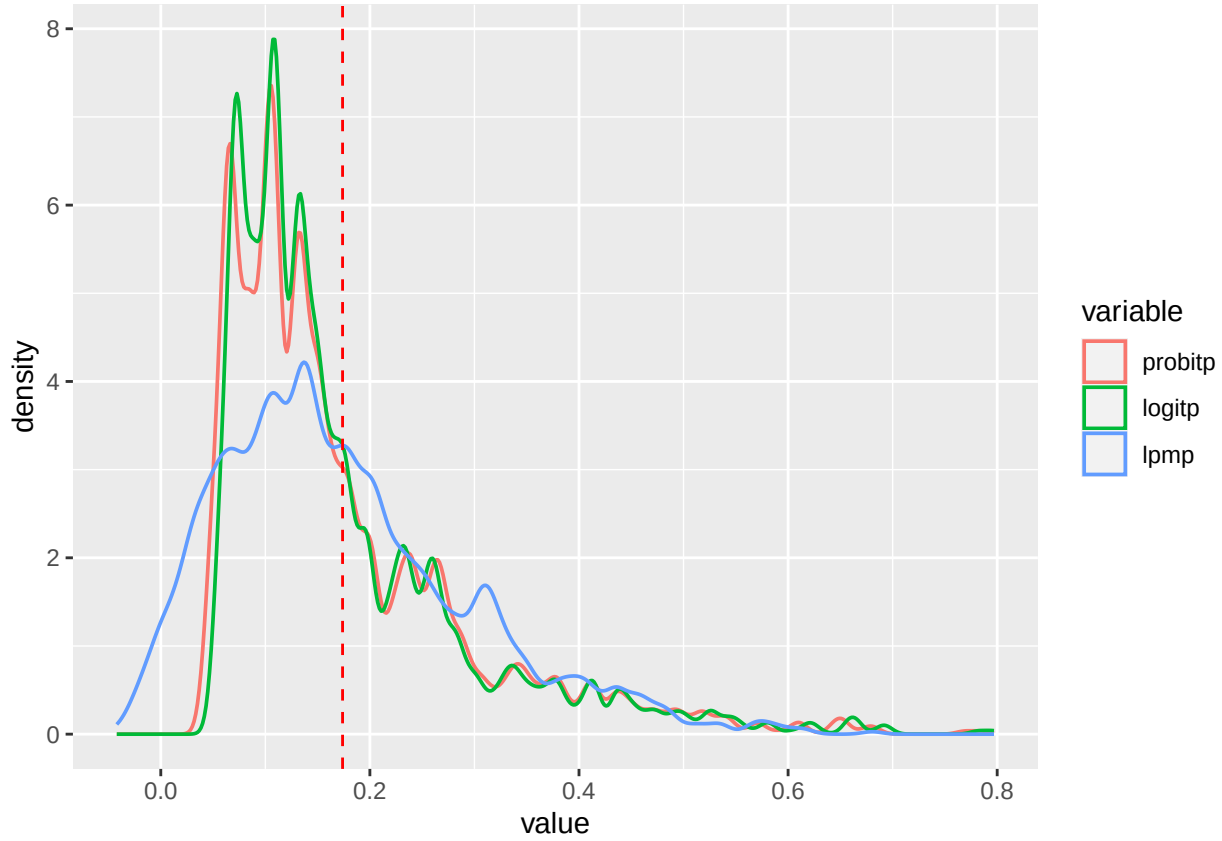
The mean predicted probability from each model is virtually indistinguishable, and it should come as no surprise that the mean model-predicted probabilities are almost identical to the sample mean of the dependent variable, *selfhelp*. So let's examine the three predicted probability distributions together:

```
ggplot(df_boxplot_long, aes(x=variable, y=value)) +
  geom_boxplot()
```



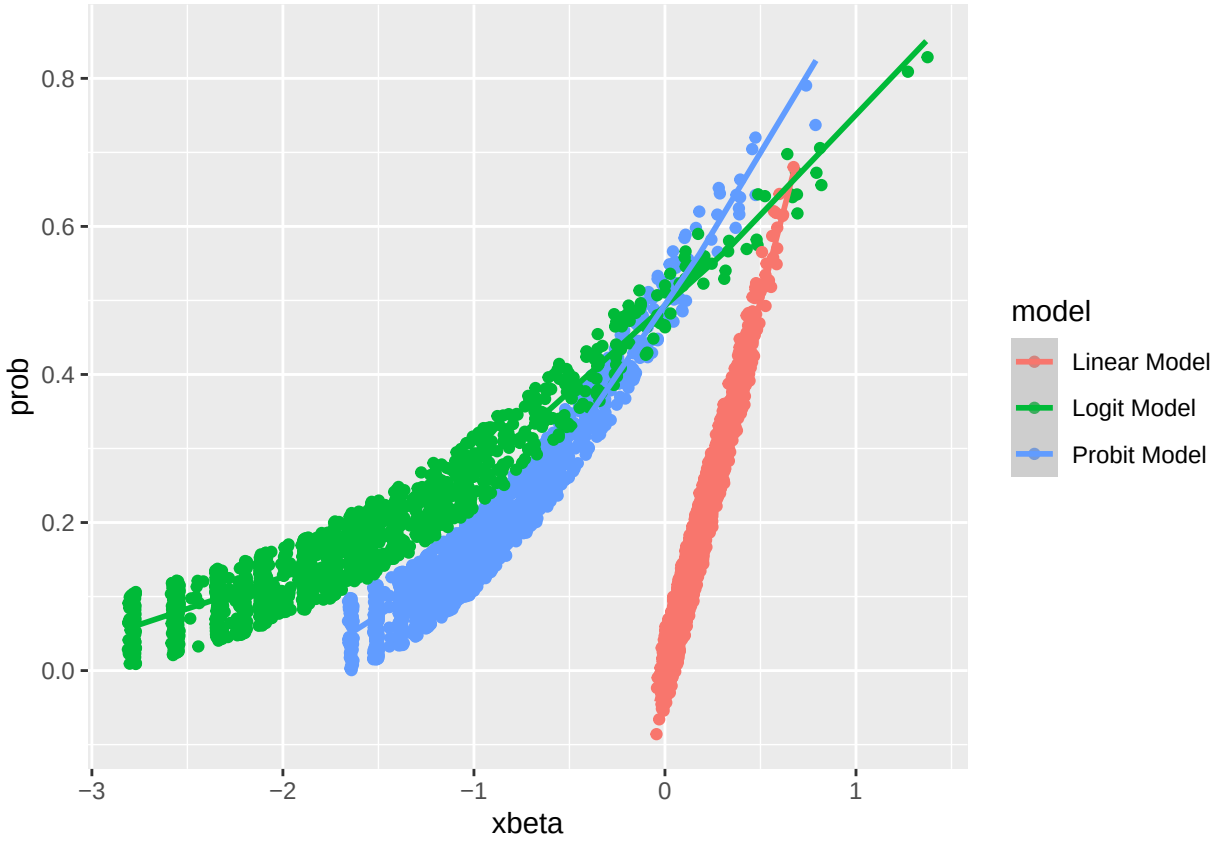

```
ggplot(df_boxplot_long, aes(x=value, color=variable)) +
  geom_density(adjust=.40, size=.65) +
  geom_vline(xintercept=mean(self_help$selfhelp), color='red',
             linetype='dashed')
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



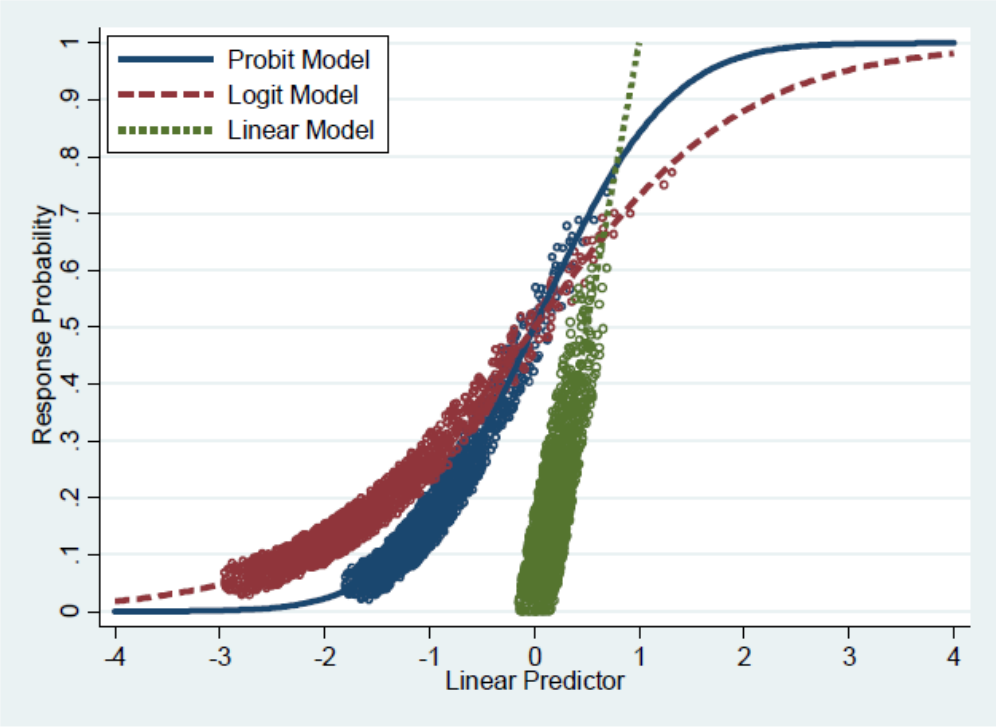
Although there are some modest differences, what seem more notable are the similarities in the distributions. The probit and logit distributions, in particular, are difficult to distinguish from each other. Indeed, the predicted probabilities from the latter two models are correlated higher than 0.999. In a final comparative graph, we can plot the predicted probabilities from each model against their respective linear predictors.

```
ggplot(df_xbp, aes(x=xbeta, y=prob, color=model)) +
  geom_point(position=position_jitter(width=.01, height=.05)) +
  geom_smooth(method="loess", n=5)
```



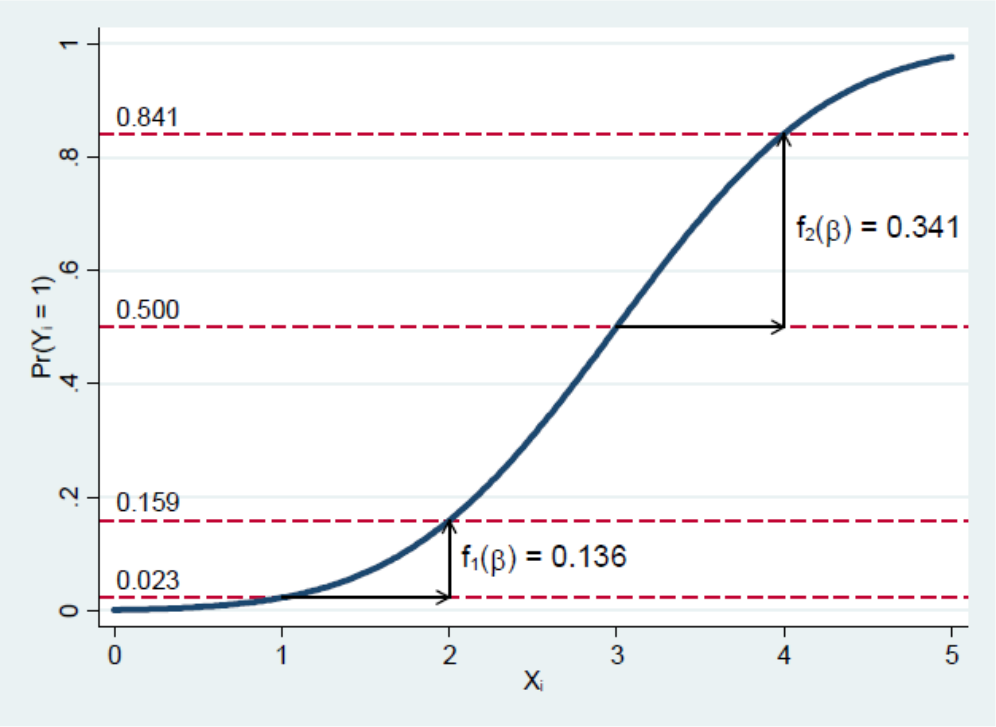
NOTE: The predicted probabilities actually line up perfectly on the fitted lines but I used a position_jitter() option in the ggplot command to introduce a small perturbation so that it is possible to see roughly how many subjects cluster at the same probability. Notice the curvilinearity in the predicted probabilities estimated from the probit and logit models. This is by design, because they track the S-curve of the standard normal and logistic c.d.f's, respectively. Recall that the predicted probabilities from these models map onto their linear predictors by way of a non-linear function, $F(X\beta_i)$. The predicted probabilities estimated from the linear mode, of course, exhibit no curvilinearity at all, because the predicted probability is exactly equal to the linear predictor from the model.

If we back the graph out a little bit further, in fact, we can see precisely where our sample falls on the full probit, logit, and linear continua:



Now that we see the ease with which we can use the model coefficients to estimate a predicted probability for each respondent based on his or her own "profile" on the regressors, we can also use the coefficients to estimate the behavior of $Pr(Y_i)$ with respect to incremental increases in X_i for an average respondent. Recall that, because the probit and logit models are inherently non-linear, changes in $Pr(Y_i)$ must be estimated from specific values for all of the regressors (including the regressor for which a marginal effect is being calculated). The reason is easy enough to show with a graph of a hypothetical S-curve. The probit model and fitted curve are as follows:

$$Pr(Y_i = 1 \mid X_i) = \Phi(-3.0 + 1.0X_i)$$



In this graph, we will consider the case in which we evaluate a one-unit increase from $X_i = 1$ to $X_i = 2$, and then a one-unit increase from $X_i = 3$ to $X_i = 4$. In the instance on the left, the height of the vertical arrow is 0.136, indicating that the predicted $Pr(Y_i = 1)$ increases by 0.136 when X_i increases from 1 to 2 (0.159 - 0.023 = 0.136). In the instance on the right, the height of the vertical arrow is 0.341, indicating that $Pr(Y_i = 1)$ increases by 0.341 when X_i increases from 3 to 4 (0.841 - 0.500 = 0.341). In both cases, X_i advances by one unit, it just happens to be the case that the regression curve is fairly shallow at $X_i = 1$ but much steeper at $X_i = 3$. Therefore, the practical impact of an incremental increase in X_i on $Pr(Y_i = 1)$ is sensitive to where the incremental increase occurs.

Under Construction