Lecture 12: Categorical Dependent Variables and Interactions

Thomas Chadefaux

Quantitative Methods I

1. Specifying the regression model

2. Qualitative and Limited Dependent Variable Models

Specifying the regression model

What variables to include

- Tests are usually performed under the assumption that the model has been correctly specified.
- For instance, the computed standard errors of estimated coefficients and their P-values depend on this assumption.

How many variables should I include?

- Assume that a set of explanatory variables has been selected as possible determinants of the variable y.
- Even if one is interested in the effect of only one of these explanatory variables—say, x_2—it is important not to exclude the other variables a priori.
- The reason is that variation in the other variables may cause variations in the variable y, and, if these variables are excluded from the model, then all the variation in y will be attributed to the variable x_2 alone.

How many variables should I include?

- On the other hand, the list of possibly influential variables may be very long. If all these variables are included, it may be impossible to estimate the model (if the number of parameters becomes larger than the number of observations) or the estimates may become very inefficient
- The question then is how many variables to include in the model.

Omitting Relevant Variables: Omitted Variable Bias

- Need to control for important variables, or else our estimates will be biased.
- But the variance of b_R will be smaller than the variance of b_1.
  - Why? Intuitively, we have more data to estimate b_R.

Omitting Relevant Variables: Omitted Variable Bias

- Venn Diagram
- Note that omitted variable bias is not a problem for our estimate of b_1 if (it always leads to bias for our constant estimate):
  - X_1 and X_2 are not correlated.
  - $\beta_2 = 0$.

Omitted variable bias

- Sometimes we know that we have omitted a variable but have no way of including it, possibly because it is difficult to measure or observe.
- You should at least think about the likely direction of the bias. So there is a trade-off between bias and efficiency.

What about irrelevant variables?

- Adding irrelevant variables leads to inefficiency
  - it eats up a degree of freedom.
  - Most of the time, there will be some collinearity with other Xs, and hence the variance of the *b*s will increase.
- But the estimates will remain unbiased.

Trade-off between bias and efficiency

- Which estimator of b_1 should be preferred, b_1 or b_R?
- If $\beta_2 = 0$, then clearly we need to use b_R to avoid the loss of efficiency.
- BUT usually $\beta_2 \neq 0$. If we remove the estimator b_2, we obtain a more efficient b_1, but also one that is biased.

So how to choose between the two models?

- F-test
- AIC and BIC
- Out-of-sample predictions. Possible evaluation criteria are
  - the root mean squared error (RMSE)

    $$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2}$$

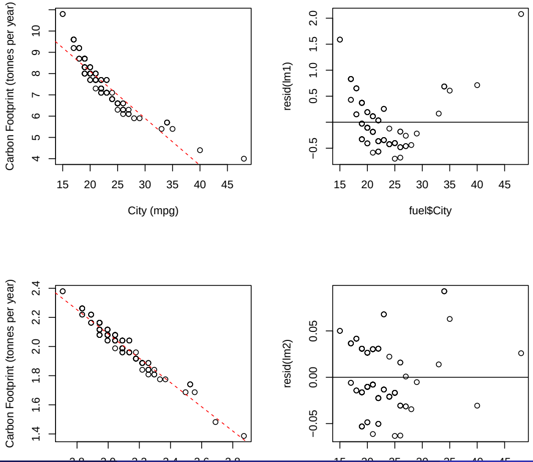
  - the mean absolute error (MAE).MAE =

    $$\frac{1}{n}\sum_{1}^{n}|y_i - \hat{y}_i|$$

- Iterative variable selection methods: Brute force, bottom-up or top-bottom.

Non-linearities

- First note that the scaling of variables is of no intrinsic importance
- An important transformation is to take the logarithm of a variable.



Dummy Variables

- Your independent variables may not be quantitative, but rather qualitative. In that case, you will need dummy variables.
- E.g., the effect of education on income. You think the effect is the same for men and women, but that men tend to get more on average for the same level of education.
  - i.e., the slope is the same, but the intercept is not. One way of formulating this argument is :

$$Y_i = \alpha + \beta_1 X_1 + \gamma D_i + \varepsilon_i,$$

where D_i takes value 0 for women and 1 for men.

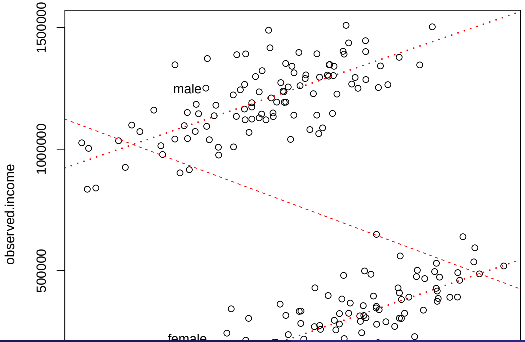
Dummy variables

Thus for women, the model becomes:

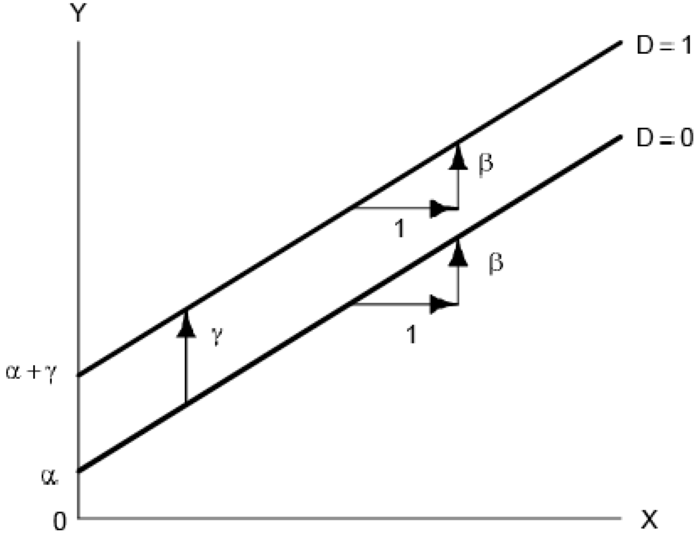
$$Y_i = \alpha + \beta_1 X_1 + \gamma(0) + \varepsilon_i = \alpha + \beta_1 X_1 + \varepsilon_i$$

whereas for men

$$Y_i = \alpha + \beta_1 X_1 + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta_1 X_1 + \varepsilon_i$$



dummies



Dummies: Interpretation

- γ gives the difference in intercepts for the two regression lines—that is, the expected income advantage accruing to men when education is held constant. If men were disadvantaged relative to women, then λ would be negative.
- Note that it doesn't really matter which group is coded 1 or 0 (just make sure you interpret correctly. . .).
- Of course, we can add more dummy variable, for example for ethnicity, etc.

Polytomous variables

- Your ‘dummies’ need not be binary. Suppose, for example, that you are interested in the effect of GDP on a country’s level of democracy but think that its geographic location matters. In particular, you want to control for the region a country belongs to. Say, Europe, America, Middle East, Asia, Australia. That’s 5 possible values. Then you would create 4 (yes, 4, not 5. Why?) dummy variables:

  $$Democracy_i = \alpha + \beta_1 GDP_i + \beta_2 Europe_i + \beta_3 America_i + \beta_4 MiddleEast_i$$

Interaction Effects

- Now you might not think that the effect of gender on income is simply a shift in intercept. Instead, gender might affect the slope. In other words, one more year of education for a man might have a different effect on income than an additional year of education for a woman.
- To model this type of 'interaction', we need a model with different slopes for men and women. - Consider the following model:

$$Y_i = \alpha + \beta_1 education_i + \beta_2 gender_i + \beta_3 (education_i \times gender_i) + \varepsilon_i \tag{1}$$

Interaction Effects

So for males, the model is:

$$Y_i = (\alpha + \beta_2) + (\beta_1 + \beta_3) education_i + \varepsilon_i \tag{2}$$

and for females:

$$Y_i = \alpha + \beta_1 education_i + \varepsilon_i \tag{3}$$

Interaction Effects

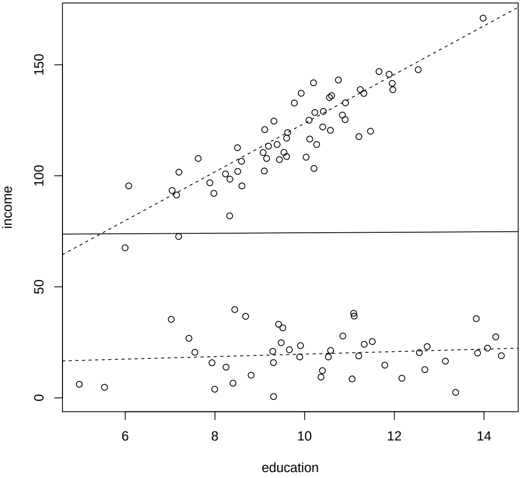


Figure 2

Interaction Effects

Note that the slope for education for males ($\beta_1 + \beta_3$) is different than the slope for females (β_1), AND that the intercepts are different, just like it was for dummies.

VERY IMPORTANT: You MUST be careful with the interpretation of interaction terms. Suppose your model is

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \tag{4}$$

Then the effect of X on Y holding Z constant is:

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z$$

In other words, the effect of X on Y depends on the value of Z! You *can no longer* interpret β_1 as the effect of X_1 on Y.

Qualitative and Limited Dependent Variable Models

Introduction

- So far we have considered the case of continuous dependent variables. Income, % of votes, etc., are continuous variables. But often your *dependent* variable will be binary or at least limited to a small number of options. For example, whether:
    - A country enters (or wins) a war or not
    - A politician wins an election
    - A citizen votes
    - A bill is adopted
    - Someone gets a BA/MA/PhD
- We have already encountered the case of binary variables in the context of independent variables. We called these dummy variables, and they could be for example gender (male/female), whether someone went to college or not, etc.

Why Not Use the Linear Probability Model?

Suppose y_i denotes whether a country goes to war or not. I.e., $y_i = 0$ if it does not enter the war, and $y_i = 1$ if it does. Suppose that we are interested in the probability that a country goes to war, i.e., $P(y_i = 1)$. We think the level of trade dependence of that country might explain the decision to go to war. So we estimate the following model:

$$y_i = \beta_0 + \beta_1 trade_i + \varepsilon_i$$

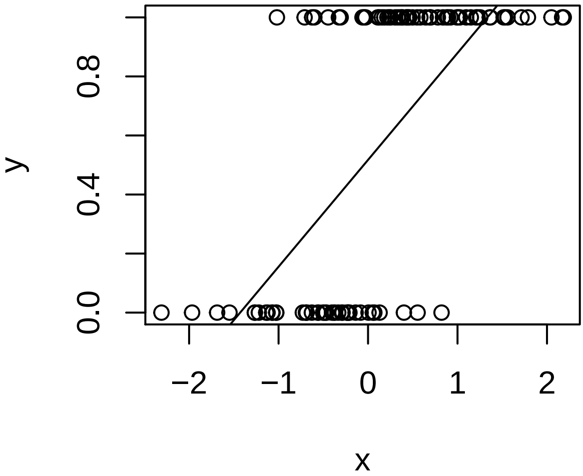
Why Not Use the Linear Probability Model?

```
setwd('~/Dropbox/Documents/Academia/Teaching/TCD/PO7005_Q
set.seed(123)
n <- 100
x <- rnorm(n, mean=0)
y <- round(1 / (1 + exp(-(x + rnorm(n, sd = 0.5)))))

#--- First use the linear model
lm1 <- lm(y ~ x)
#
```

LPM results

```
plot(x, y, lty = 2)
abline(lm1)
```



LPM results

Three problems with using a linear model:

1. Nonsensical predictions. For example, it makes no sense to predict a probability of war > 1.

```
predict(lm1, newdata= data.frame(x = 2))
```

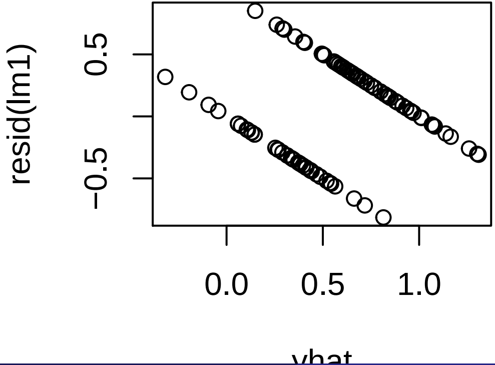
```
##        1
## 1.241055
```

LPM results

Three problems with using a linear model:

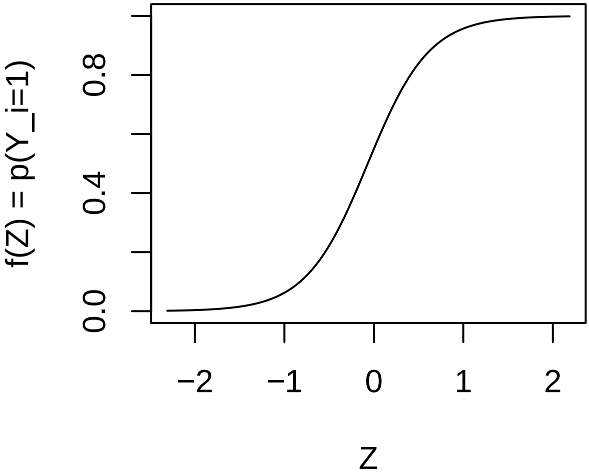
2. Second, the distribution of our error terms is clearly heteroskedastic and not normal

```
yhat <- predict(lm1)
plot(yhat, resid(lm1))
```



Sigmoid function (logit)

Example of a sigmoid function: the logistic function



Logit and Probit

- let Z be a linear function of the Xs. For example,

  $$Z = \beta_0 + \beta_1, X$$

  and then we can define $p(Y_i = 1)$ as a sigmoid function of Z.
- There are two popular versions of this 'sigmoid' curve:
  - The logistic function, used for the logit
  - The cumulative normal distribution, used in probit estimation. N
- Neither has any particular advantage, and it does not really matter which you use. Here I will only cover the logit, but the probit is pretty much the same.

Logit

In logit estimation the probability of the occurrence of the event is determined by the function

$$p(Y_i = 1|X_i) = f(Z) = \frac{1}{1 + e^{-Z}} = \frac{1}{1 + e^{-(X\beta)}} \tag{5}$$

Logit

- Note that this is simply choosing a different link from X to p. In the OLS case, the link was linear, i.e.,

$$p(Y_i = 1|X_i) = f(Z) = Z = X\beta.$$

- With the logit (and similarly with the probit), the link is a sigmoid function, such that
  - as Z increases (to infinity), e^{-Z} tends to 0, and hence $p(Y_i = 1|X_i)$ tends to 1.
  - Similarly, as Z decreases (to minus infinity), e^{-Z} tends to infinity, and hence $p(Y_i = 1|X_i)$ tends to 0.
  - As a result, our predictions are bounded by 0 and 1.
- Clearly, we cannot estimate this directly by OLS, as $p(Y_i = 1|X_i)$ is nonlinear not only in X, but also in the βs.

(Probit)

The probit model assumes that the transformation function F is the cumulative density function (cdf) of the standard normal distribution. The response probabilities are then

$$\begin{aligned} p(Y_i = 1|X_i) &= \Phi(X\beta) \\ &= \int_{-\infty}^{X\beta} \phi(t)dt \\ &= \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{/}2t^2} dt, \end{aligned}$$

where $\phi(\cdot)$ is the pdf and $\Phi(\cdot)$ is the cdf of the standard normal distribution.

Interpretation of logit in terms of odds ratio

We can transform (5) in a way in which the relationship becomes linear. First, rewrite eqn 5 as:

$$
\begin{aligned}
p(Y_i = 1|X_i) = f(Z) = \frac{1}{1+e^{-Z}} &= \frac{1}{1+\frac{1}{e^Z}} \\
&= \frac{1}{e^Z + \frac{1}{e^Z}} = \frac{e^Z}{1+e^Z}
\end{aligned}
$$

Interpretation of logit in terms of odds ratio

Now consider the *odds ratio* of the outcome being one. The odds ratio are simply the probability that $y = 1$ divided by the probability that $y = 0$. So the odds ratio are:

$$\frac{p(Y_i = 1|X_i)}{1 - p(Y_i = 1|X_i)} = \frac{\frac{e^Z}{1+e^Z}}{1 - \frac{e^Z}{1+e^Z}} = \frac{\frac{e^Z}{1+e^Z}}{\frac{1}{1+e^Z}} = e^Z$$

Now we simply take the log of (**??**):

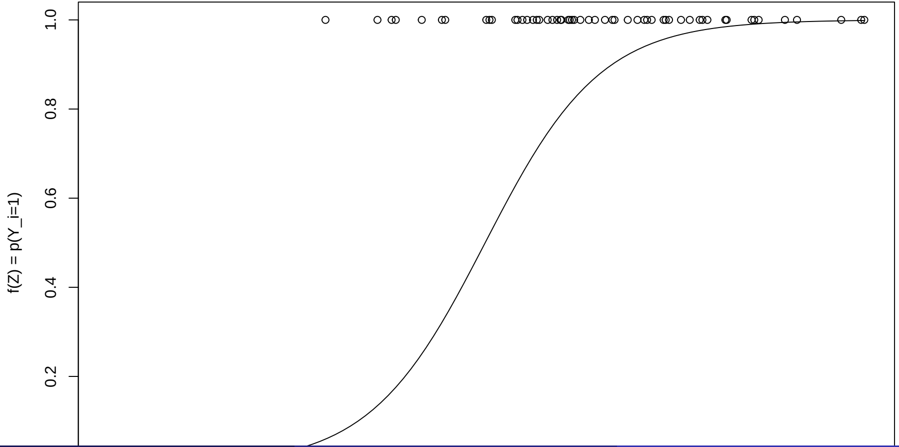
$$L_i = ln(\frac{p(Y_i = 1|X_i)}{1 - p(Y_i = 1|X_i)}) = Z_i = \beta_0 + \beta_1 X$$

Interpretation of logit in terms of odds ratio

- The coefficient for x is the log of the odds ratio between x=0 and x=1.
- More generally, this is an ugly interpretation. Use plots!

Estimation Using R

```
glm.logit <- glm(y ~ x, family = binomial(link = 'logit')
plot(x, y, type='p', xlab='Z', ylab='f(Z) = p(Y_i=1)')
curve(predict(glm.logit ,data.frame(x=x),type="resp"), ad
```



Interpretation of Logit Coefficients

VERY IMPORTANT: you can no longer interpret β_k as the marginal effect of variable x_k on y! To interpret the effect of a change of x_k by one unit, we need to calculate the derivative of $p(Y_i = 1|x_i)$ with respect to x_k:

$$\frac{\partial p(Y_i = 1|x_i)}{\partial x_{ik}} = \frac{e^{x_i\beta}}{(1 + e^{x_i\beta})^2}\beta_k$$

Note that the marginal effect of β_k depends on all the x_is. I.e., you can no longer say "an increase in x leads to an increase in y by such amount. Rather, the effect depends on the values of the other covariates. Look back at the figure of the logit curve, and note that the effect is much strong around $x = 0$ than for $x = -10$ or $x = 10$.

Interpretation of Logit Coefficients

To see this, consider first the case in which you only have one IV. Suppose you obtain the following results:

```
glm.logit <- glm(y ~ x, family = binomial(link = 'logit')
library(xtable)
print(xtable(glm.logit))
```

% latex table generated in R 3.2.4 by xtable 1.8-2 package % Tue Dec 13 12:10:08 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1989	0.2828	0.70	0.4819
x	2.9085	0.5852	4.97	0.0000

Interpretation of Logit Coefficients

What does the coefficient on x mean? Suppose for example that you estimate your logistic regression as:

$$-10 + .16x_1 + 0.02x_2$$

Then the effect on the odds of a unit increase in x_1 is $e^0.16 = 1.18$, meaning that the odds increase by 18% regardless of the value of x

$$\frac{P(event|x+1)/(1-P(event|x+1)}{P(event|x)/(1-P(event|x)}$$

Interpretation of Logit Coefficients

Now let's calculate the effect of an increase of x by 1. We will do this by calculating $E[y|x=a]$ - $E[y|x=b]$, where the expected value is given in (5). I.e.,

$$p(Y_i = 1|X_i = 0) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times 0)}}$$

```r
z0 <- coef(glm.logit)%*%c(1,0)
z1 <- coef(glm.logit)%*%c(1,1)
z2 <- coef(glm.logit)%*%c(1,2)
Eyx0 <- 1/(1+exp(-z0)); Eyx0
```

```
##            [,1]
## [1,] 0.5495617
```

```r
Eyx1 <- 1/(1+exp(-z1)); Eyx1
```

```
##            [,1]
```

Another way to interpret β:

β_1 measures the change in L for a unit change in X, that is, it tells us how the log-odds of $y = 1$ change as X increases by one unit. The intercept β_0 is the value of the log-odds of $y = 1$ if $X = 0$.

```r
glm.logit <- glm(y ~ x, family = binomial(link = 'logit')
beta <- coef(glm.logit)
newx <- 0
ez <- beta[1] + beta[2]*newx
exp(ez)/(1+exp(ez))
```

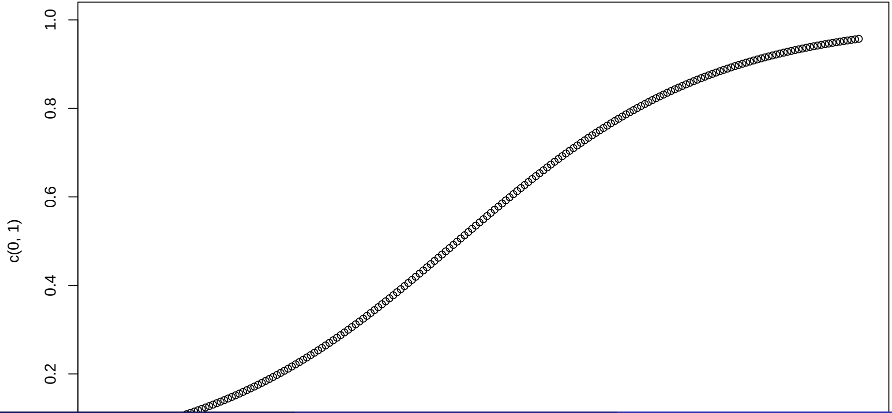
```
## (Intercept)
##   0.5495617
```

```r
# note that this is equivalent to R's canned function:
predict(glm.logit, newdata=data.frame(x = newx), type='re
```

```
##          1
```

Plot the effect of X on Y:

```
plot(c(-1,1), c(0,1), type='n')
for(i in seq(-1,1, 0.01)){
     points(i, predict(glm.logit, newdata=data.frame(x =
}
```



Plot the effect of X on Y:

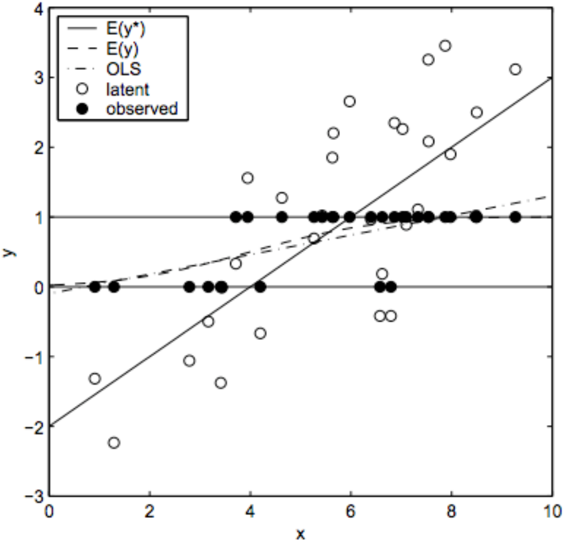
The best way to present your results is to plot or report predicted probabilities. However, you always need to be aware that you need to hold all other variables at a given level. The level you choose will affect the effect!! Typically what is done is to hold variables at the median, vary x, and report predicted probability. Or hold variables at the median and set x to its 25 percentile, 50 percentiles, and 75 percentile.

Latent Variable Interpretation

There is another way to understand the logit. The idea is that while we are observing y as 0 or 1, there is an underlying variable that is continuous. For example, we may observe only whether a voter votes for candidate A or B. But maybe one citizen has a very strong utility for candidate A, whereas another has only a slightly higher utility for A than for B. Both will vote for A, and that is all we observe, but their underlying y is different. We call this underlying variable a "latent" variable and denote it $y*$. In short,

$$y_i = \begin{cases} 1 & if \;\; y_i* > 0 \\ 0 & if \;\; y_i* \leq 0 \end{cases} \tag{6}$$

Latent Variable Interpretation



Latent variable model

Maximum Likelihood Estimation

In the case of a binary DV, there are two possible outcomes, and we can calculate the probability of each:

$$\begin{aligned} P(Y = 1|X) &= F(X\beta) \\ P(Y = 0|X) &= 1 - F(X\beta), \end{aligned} \tag{7}$$

where $F(\cdot)$ is the logistic function, i.e., $F(X\beta) = \frac{e^{X\beta}}{1+e^{X\beta}}$. Now we just need to realize that our dependent variable is the same as bernouilli variable, since it can take on 2 values, 1 or 0. So we can write:[1]

$$L_i = P(Y = y_i|X) = P(Y = 1|X)^{y_i} \times P(Y = 0|X)^{1-y_i}$$

so our likelihood function is simply

$$L = \prod_{i=1}^{n} P(Y = y_i|X) = \prod_{i=1}^{n} P(Y = 1|X)^{y_i} \times P(Y = 0|X)^{1-y_i}$$

[1] Remember that the PDF (technically, the pmf) of the bernoulli distribution is $p^x(1 - p^x)$