EC203: Applied Econometrics
Dummy variables

Dr. Tom Martin

University of Warwick

Illustrative reading:

- Wooldridge: Chapter 7
- Dougherty: Chapter 5
- Gujarati: Chapter 3

Quantitative and qualitative variables

In the majority of the models so far, we have explicitly considered variables with quantitative meaning: wage, education, experience, food security, household income and IQ.

What about qualitative variables? We have considered a few, including: region, gender and training programs. In this lecture we see how to deal with them explicitly.

Quantitative and qualitative variables

id	wage	education	male	married	north
1	12	18	1	1	1
2	7	15	0	0	0
3	38	18	0	1	1
⋮	⋮	⋮	⋮	⋮	
487	14	16	0	1	0
488	13	12	1	0	0
489	29	21	1	1	1

A single additive dummy variable

Suppose we are interested in estimating the following wage equation:

$$W_i = \alpha + \beta_1 E_i + \beta_2 M_i + \epsilon_i$$

Where W_i is hourly wage, E_i is years of education and $M_i = 1$ if individual i is male, and $M_i = 0$ if individual i is female.

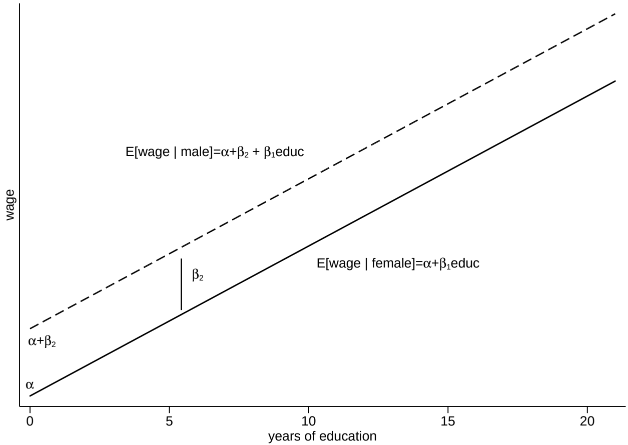
A single additive dummy variable

The parameter β_2 represents the difference in mean earnings of males relative to females, holding education constant. To see this:

- If $M = 0 : E[W|M = 0, E] = \alpha + \beta_1 E$
- If $M = 1 : E[W|M = 1, E] = \alpha + \beta_1 E + \beta_2$
- Taking differences:
- $\beta_2 = E[W|M = 1, E] - E[W|M = 0, E]$

A single additive dummy variable

At each year of education, β_2 is the expected difference in male wages relative to female wages.



A single additive dummy variable

Suppose we have the estimated regression:

$$\hat{W}_i = \underset{(0.652)}{-1.651} + \underset{(0.279)}{2.273} M_i + \underset{(0.050)}{0.506} E_i$$

Regressions are commonly displayed in table form:

Table: Wage regression results

Variable	**Coefficient**	**(Std. Err.)**
educ	0.506	(0.050)
male	2.273	(0.279)
Intercept	-1.651	(0.652)

Males earn 2.3 US dollars per hour more than females, holding education constant.[1]

[1] Females with no years of education earn -1.651 US dollars per hour. The intercept is not particularly useful.

More than one additive dummy variable

Suppose our categorical variable has more than two choices. Consider a categorical variable for region, which has four (m) outcomes:

- northeast (NE): $NE_i = 1$ if in NE, 0 otherwise
- northwest (NW): $NW_i = 1$ if in NW, 0 otherwise
- southeast (SE): $SE_i = 1$ if in SE, 0 otherwise
- southwest (SW): $SW_i = 1$ if in SW, 0 otherwise

In the model include (m-1) outcomes as dummy variables:

$$W_i = \alpha + \beta_1 E_i + \beta_2 NE_i + \beta_3 NW_i + \beta_4 SW_i + \epsilon_i$$

More than one additive dummy variable

Turning each dummy variable on individually and taking expectations:

1. $E[W|NE = NW = SW = 0, educ] = \alpha + \beta_1 E_i$
2. $E[W|NE = 1, NW = SW = 0, educ] = \alpha + \beta_2 + \beta_1 E_i$
3. $E[W|NE = 0, NW = 1, SW = 0, educ] = \alpha + \beta_3 + \beta_1 E_i$
4. $E[W|NE = NW = 0, SW = 1, educ] = \alpha + \beta_4 + \beta_1 E_i$

More than one additive dummy variable

Then taking differences we get:

- $\beta_2 = E[W|NE = 1, NW = SW = 0, E] - E[W|NE = NW = SW = 0, E]$.
  - The expected difference between wages between the northeast and southeast, holding education constant.

- $\beta_3 = E[W|NE = 0, NW = 1, SW = 0, E] - E[W|NE = NW = SW = 0, E]$.
  - The expected difference between wages between the northwest and southeast, holding education constant.

- $\beta_4 = E[W|NE = NW = 0, SW = 1, E] - E[W|NE = NW = SW = 0, E]$.
  - The expected difference between wages between the southwest and southeast, holding education constant.

The southeast is our reference category: the category omitted from the regression.

Multiplicative dummy variables

In the specifications considered so far we have only allowed the intercept to change.

What if we hypothesise that the slope also changes?

For instance, not only do males earn more than females, but males also see a greater return to an additional year of education.

To test this hypothesis we could specify the model:

$$W_i = \alpha + \beta_1 E_i + \beta_2 M_i + \beta_3 (M_i * E_i) + \epsilon_i$$

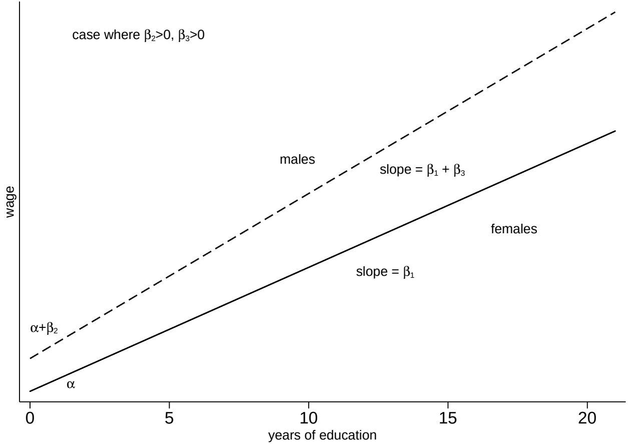
Where $M_i * E_i$ is the male dummy (M_i) multiplied by the years of education variable (E_i).

Multiplicative dummy variables

Turning on the dummy variable and taking expectations:

- If $M = 0 : E[W|M = 0, E] = \alpha + \beta_1 E_i$
- If $M = 1 : E[W|M = 1, E] = \alpha + \beta_1 E_i + \beta_2 + \beta_3 E_i$
- $E[W|M = 1, educ] = (\alpha + \beta_2) + (\beta_1 + \beta_3)E_i$.
- The intercepts are: α and $\alpha + \beta_2$ for females and males, respectively.
- While the slope coefficients are: β_1 and $\beta_1 + \beta_3$ for females and males, respectively.

Multiplicative dummy variables



Multiplicative dummy variables

Thus, given the model:

$$W_i = \alpha + \beta_1 E_i + \beta_2 M_i + \beta_3 (M_i * E) + \epsilon_i$$

The interpretations are:

- α is expected wage of females, when education is zero.
- β_1 is the increase in wages associated with an additional year of education for females.
- β_2 is the increase in the expected wages of males relative to females, when education is zero.
- β_3 is the increase in wages associated with an additional year of education for males relative to females.

Multiplicative dummy variables

This multiplicative case can be extended to the more general categorical case. For a categorical variable with m outcomes (NE, NW, SW, SE), omit one variable which is then your base category. For example,

$$\begin{aligned} W_i = &\alpha + \beta_1 E_i + \beta_2 NE_i + \beta_3 NW_i + \beta_4 SW_i + \\ &\beta_5 (NE_i * E_i) + \beta_6 (NW_i * E_i) + \beta_7 (SW_i * E_i) + \epsilon_i \end{aligned}$$

Multiplicative dummy variables

Turning on the dummy variables, taking expectations and differences:

- $E[W|NE = NW = SW = 0, E] = \alpha + \beta_1 E_i$
- α is the mean wage in southeast among individuals with zero years of education.
- β_1 is the mean wage return to an additional year of education in the southeast.

Multiplicative dummy variables

Turning on the dummy variables, taking expectations and differences:

- $E[W|NE = 1, NW = SW = 0, E] = (\alpha + \beta_2) + (\beta_1 + \beta_5)E_i$
- $\beta_2 + \beta_5 E_i = E[W|NE = 1, NW = SW = 0, E] - E[W|NE = NW = SW = 0, E]$
- β_2 is the mean wage in northeast relative to the southeast, among individuals with zero years of education.
- β_5 is the mean wage return to an additional year of education in northeast relative to the southeast.

Multiplicative dummy variables

Turning on the dummy variables, taking expectations and differences:

- $E[W|NE=0,NW=1,SW=0,E]=(\alpha+\beta_3)+(\beta_1+\beta_6)E_i$
- $\beta_3+\beta_6 E_i = E[W|NE=0,NW=1,SW=0,E]-E[W|NE=NW=SW=0,E]$
- β_3 is the mean wage in northwest relative to the southeast, among individuals with zero years of education.
- β_6 is the mean wage return to an additional year of education in northwest relative to the southeast.

Multiplicative dummy variables

Turning on the dummy variables, taking expectations and differences:

- $E[W|NE = NW = 0, SW = 1, E] = (\alpha + \beta_4) + (\beta_1 + \beta_7)E_i$
- $\beta_4 + \beta_7 E_i = E[W|NE = NW = 0, SW = 1, E] - E[W|NE = NW = SW = 0, E]$
- β_4 is the mean wage in southwest relative to the southeast, among individuals with zero years of education.
- β_7 is the mean wage return to an additional year of education in southwest relative to the southeast.

Interactive dummy variables

What happens if we interact dummy variables with dummy variables? Suppose we have the model ($male = 1$ if male and 0 if female, while $north = 1$ if employed in the north, 0 in the south):

$$W_i = \alpha + \beta_1 M_i + \beta_2 N_i + \beta_3 (M_i * N_i) + \epsilon_i$$

What effect does each coefficient pick up?

Interactive dummy variables

Turning dummy variables on and taking expectation:

- $E[W|M = 0, N = 0] = \alpha$ (1)
- $E[W|M = 1, N = 0] = \alpha + \beta_1$ (2)
- $E[W|M = 0, N = 1] = \alpha + \beta_2$ (3)
- $E[W|M = 1, N = 1] = \alpha + \beta_1 + \beta_2 + \beta_3$ (4)

What effect does each coefficient pick up?

Interactive dummy variables

Taking differences:

- $(2) - (1) : E[W|M = 1, N = 0] - E[W|M = 0, N = 0] = \beta_1$ (A)
- β_1: difference in expected wages of males and females, in the south. The gender pay gap in south.
- $(3) - (1) : E[W|M = 0, N = 1] - E[W|M = 0, N = 0] = \beta_2$
- β_2: difference in expected wages in the north and south, for females. The north south pay gap for females.

Interactive dummy variables

Taking differences:

- $(4)-(3): E[W|M=1,N=1]-E[W|M=0,N=1]=$ $\beta_1+\beta_3$ (B)
- $\beta_1+\beta_3$: difference in expected wages of males and females, in the north. The gender pay gap in north.
- $(B)-(A): \{E[W|M=1,N=1]-E[W|M=0,N=1]\}-\{E[W|M=1,N=0]-E[W|M=0,N=0]\}=\beta_3$
- β_3: difference in gender pay gaps between north and south.

Chow test I: differences across groups

The population regression of interest is the short wage equation:

$$W_i = \beta_0 + \beta_1 E_i + \beta_2 exp_i + \beta_3 exp_i^2 + \epsilon_i$$

- Our population of interest includes both males and females.
- It is often hypothesised that male and female labour markets are substantively different, such that, in fact, male and female wage equations should be estimated separately.
- How should one statistically test for such a possibility?

Chow test I: differences across groups

The hypothesis is suggesting, that instead of a joint wage equation:

$$W_i = \beta_0 + \beta_1 E_i + \beta_2 exp_i + \beta_3 exp_i^2 + \epsilon_i \quad (1)$$

We have a separate wage equation for each gender:

Male wage equation:

$$W_i = \beta_0^M + \beta_1^M E_i + \beta_2^M exp_i + \beta_3^M exp_i^2 + \epsilon_i^M \quad (2a)$$

Female wage equation:

$$W_i = \beta_0^F + \beta_1^F E_i + \beta_2^F exp_i + \beta_3^F exp_i^2 + \epsilon_i^F \quad (2b)$$

Chow test I: differences across groups

The null hypothesis we need to test is thus:

$$H_0 : \beta_0^M = \beta_0^F, \beta_1^M = \beta_1^F, \beta_2^M = \beta_2^F, \beta_3^M = \beta_3^F$$

Against the alternative that at least one of the coefficients is different.

- This is a multiple restriction test, thus, we can use an F-test.
- Imposing the null we get the restricted model: model (1).
- The unrestricted model is (2a) and (2b) taken together.
- So $F = \frac{[RSS_R - RSS_U]/q}{RSS_U/dof}$.
- Where $RSS_R = RSS^1, RSS_U = RSS^{2a} + RSS^{2b}, q = k + 1$ and finally $dof = n - 2(k + 1)$, number of observations minus the number of parameters in the unrestricted model.

Chow test I: differences across groups

- Although the above procedure is known as a Chow test, it is simply a specific F-test.
- Recall, even if all the interactions are statistically insignificant (t-tests) they can be highly significant together.
- We can carry out this type of test for any hypothesised structural break: before after crisis, public/private, married/unmarried, urban/rural, England/Scotland, communist/capitalist,....
- Any situation, where you can put forward a strong argument for a structural break.

Chow test I: differences across groups

Generalising the above situation, suppose we have the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i, i = 1, 2, ..., n \quad (1)$$

We hypothesise there is a structural break. Let $D_i = 1$ for $i = 1, ..., n_1$ and $D_i = 0$ for $i = n_1 + 1, n_1 + 2, ..., n$, such that:

$$Y_i = \beta_0^1 + \beta_1^1 X_{1i} + \beta_2^1 X_{2i} + ... + \beta_k^1 X_{ki} + \epsilon_i^1, i = 1, 2, ..., n_1 \quad (2a)$$
$$Y_i = \beta_0^2 + \beta_1^2 X_{1i} + \beta_2^2 X_{2i} + ... + \beta_k^2 X_{ki} + \epsilon_i^2, i = n_1 + 1, ..., n \quad (2b)$$

Chow test I: differences across groups

Thus, we want to run the following test:

$$H_0 : \beta_0^1 = \beta_0^2, \beta_1^1 = \beta_1^2, ..., \beta_k^1 = \beta_k^2$$

- Under the null we have model (1), our restricted model. The alternative is (2a) and (2b), our unrestricted model (together).
- So $F = \frac{[RSS_R - RSS_U]/q}{RSS_U/dof}$, where $RSS_R = RSS^1, RSS_U = RSS^{2a} + RSS^{2b}, q = k + 1$ and finally $dof = n - 2(k + 1)$, number of observations minus the number of parameters in the unrestricted model.

Chow test II: differences across groups

Consider again the short wage equation:

$$W_i = \beta_0 + \beta_1 E_i + \beta_2 exp_i + \beta_3 exp_i^2 + \epsilon_i \quad (1)$$

An alternative way to test if there is a structural break between male and female wage functions is the following. Define a male dummy. Then specify the following regression:

$$\begin{aligned} W_i = &\beta_0 + \beta_1 E_i + \beta_2 exp_i + \beta_3 exp_i^2 + \\ &\delta_0 M_i + \delta_1 (M_i * E_i) + \delta_2 (M_i * exp_i) + \\ &\delta_3 (M_i * exp_i^2) + \epsilon_i \quad (2) \end{aligned}$$

Chow test II: differences across groups

The hypothesis we are then required to test is:

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$$

Where in this case:

- We have a multiple restriction test, so we can use an F-test.
- Under the null, our restricted model is model (1).
- The restricted model is model (2).
- So $F = \frac{[RSS_R - RSS_U]/q}{RSS_U/dof}$, where $RSS_R = RSS^1, RSS_U = RSS^2, q = k+1$ and finally $dof = n - 2(k+1)$, number of observations minus the number of parameters in the unrestricted model.

Chow test II: differences across groups

More generally, suppose we have a dummy variable $D = 1, 0$.
To test for a structural break, run the regression:

$$
\begin{aligned}
Y =& \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + ... + \delta_k X_{ki} \\
& + \gamma_0 D_i + \gamma_1 D_i * X_{1i} + \gamma_2 D_i * X_{2i} + ... + \gamma_k D_i * X_{ki} + \epsilon_i
\end{aligned}
$$

- Then test the null $H_0 : \gamma_0 = \gamma_1 = \gamma_2 = ... = \gamma_k = 0$ using an F-test.
- The unrestricted is model above, while restricted is $Y_i = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + ... + \delta_k X_{ki} + \epsilon_i$
- If we reject the null, then we have a structural break.

Chow test II: differences across groups

Note this second form of the Chow test is more flexible, as we can also use this to test for subsets of variables. For instance, to test for a change in intercept and slope coefficient on just X_1, we would do the following:

- Run the unrestricted model:
  $Y_i = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + ... + \delta_k X_{ki} + \gamma_0 D_i + \gamma_1 D_i * X_{1i} + \epsilon_i$
- Then test the hypothesis $H_0 : \gamma_0 = \gamma_1 = 0$.
- Where our restricted model would be
  $Y_i = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + ... + \delta_k X_{ki} + \epsilon_i$

Linear Probability Model: LPM

So far we have only considered the possibility of independent dummy variables. What if our dependent variable is a dummy variable?[2]

- To go to university or not (1-yes, 0-no)?
- To enter a micro-credit program or not (1-yes,0-no)?
- To be a teenage mother or not (1-yes, 0-no)?
- To migrate or not (1-yes, 0-no)?

[2] These are also commonly known as binary choice models, limited dependent variable models, dummy dependent variable models, or qualitative response models.

Linear Probability Model: LPM

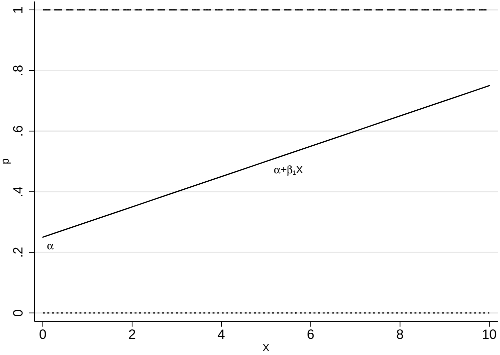
Suppose the probability of an event Y occurring for individual i is p_i. Further, assume that we can observed each p_i in our sample.

If this was the case, then assuming p_i can be modelled as a linear function of X's, we get:

$$p_i = p(Y_i = 1) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

Linear Probability Model: LPM

For a single covariate we can graph the model. For example, the probability of going to university (p) against household income (X).



Linear Probability Model: LPM

If we could observe this (latent) variable p_i - people's specific likelihood of going to university, for example - then we could simple run an OLS regression against X's as usual.

However, we cannot observe p_i, we can only observe either:

$$Y_i = 1 \; or$$
$$Y_i = 0$$

That is, we can only observe the actual decision made.

Linear Probability Model: LPM

For example: Y_i equals one if an individual goes to university and 0 otherwise.

- Therefore, β does not capture the change in Y for a given change in X.
- Since Y changes discretely by -1 or 1, or does not change.

So what does β capture?

Linear Probability Model: LPM

Given a population model:

$$Y_i = \alpha + \beta_1 X_1 + ... + \beta_k X_k + \epsilon_i$$

It can be split into two components:

- 1. $E[Y_i|\mathbf{X}] = \alpha + \beta_1 X_1 + ... + \beta_k X_k$.
- This is the deterministic component: The expected value of Y.
- 2. $Y_i - E[Y_i|\mathbf{X}] = \epsilon_i$.
- This is the unexplained component: The error term.

Linear Probability Model: LPM

The crucial point to recognise is that, when Y_i is a binary variable taking on values 0 and 1, it is always true that:

- The probability of success $(p = 1)$ is exactly the same as the expected value of Y.
- That is, $E[Y|\mathbf{X}] = 1 * (p_i|\mathbf{X}) + 0 * (1 - (p_i|\mathbf{X})) = (p_i|\mathbf{X}) = P(Y = 1|\mathbf{X})$.
- So the deterministic part is an estimate of the underlying latent variable: the probability you go to university of not.
- Typically written: $\hat{P}(Y = 1|X) = p(\hat{X}) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$

Linear Probability Model: LPM

Given the above discussion the interpretation of the LPM is straight forward:

- $p(\hat{X}) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$
- $\frac{\partial p(X)}{\partial X_j} = \beta_j$
- A one unit change in X_j is associated with a β_j change in the probability that an individual undertakes event Y.

Linear Probability Model: LPM

For instance, suppose are interested in estimating the following model:

$$LF_i = \alpha + \beta_1 nw_i + \beta_2 e_i + \beta_3 age_i + \beta_4 K6_i + \beta_5 K618_i + \epsilon_i$$

Where:

- Our population of interest is women of working age.
- $LF = 1$ if in the labour force and zero if not.
- nw is other sources of non-labour income (including male wage income).
- K6 is the number of children in the household aged 6 or under.
- K618 is the number of children in the household aged between 6 and 18.

Linear Probability Model: LPM

Estimating the equation we get:

$$\hat{LF}_i = 0.586 - 0.003nw_i + 0.038e_i + 0.16age_i - 0.262K6_i + .130K618_i$$

To interpret, remember that a change in the independent variable changes the probability $LF = 1$. For example:

- Another year of education increases the probability of labour force participation by 0.038, on average, holding the other covariates constant. (3.8%)
- Having one additional child under the age of 6 reduces the probability of participation by -0.262, on average, holding all else constant. (26.2%).

Linear Probability Model: LPM

The LPM is often used in empirical research because it is straight forward to interpret. However, there are three main problems with the LPM:

1. Non-normality:
   - $Y_i = 1 : \epsilon_i = 1 - \alpha - \beta_1 X_{1i} - ... - \beta_k X_{ki} = 1 - p_i$
   - $Y_i = 0 : \epsilon_i = 0 - \alpha - \beta_1 X_{1i} - ... - \beta_k X_{ki} = -p_i$
   - Hence, ϵ can take only two values and therefore doesn't follow a normal distribution: it will have a bimodal distribution.
   - Thus, normality fails: $\epsilon | X_1, X_2, ..., X_k \sim N(0, \sigma^2)$.

Linear Probability Model: LPM

2. Non-constant conditional variance in the errors (heteroskedasticity):
   - Since the $E[\epsilon_i] = 0$, $Var(\epsilon_i) = E[\epsilon^2]$
   - $Var(\epsilon) = P[Y = 1] * (1 - p_i)^2 + P[Y = 0] * (-p_i)^2$
   - $Var(\epsilon) = p_i * (1 - p_i)^2 + (1 - p_i) * (p_i)^2$
   - $Var(\epsilon) = p_i(1 - p)(1 - p_i + p_i)$
   - $Var(\epsilon) = p_i(1 - p)$
   - Given $p_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki}$, the variance of ϵ varies with X.
   - Thus, homoskedasticity fails: $Var(\epsilon | X_1, X_2, ..., X_k) = \sigma^2$

Linear Probability Model: LPM

3. predicting probabilities outside of the $[0, 1]$ interval.
   - $\hat{LF}_i =$ $0.586 - 0.003nw_i + 0.038e_i + 0.16age_i - 0.262K6_i + .130K618_i$
   - Estimating the predicted probabilities for all individuals in the sample we get $\hat{p} \in [-0.345, 1.127]$
   - This doesn't make much sense.

Linear Probability Model: LPM

Predicted probabilities outside of the unit interval are troubling if we want to make predictions. However, there are ways round this that allow us to predict a zero-one outcome:

- Let $\hat{Y}_i$ denote the fitted values: which may lie outside the 0,1 interval.
- Define $\tilde{Y}_i = 1$ if $\hat{Y}_i \geq 0.5$ and $\tilde{Y}_i = 0$ if $\hat{Y}_i < 0.5$.
- Then we have $\tilde{Y}_i$ for $i = 1, ..., n$, which take on the value 0 or 1.
- One can then use $\tilde{Y}_i$ and Y_i to obtain the proportion of correct predictions.
- This is a widely used goodness-of-fit measure.

Linear Probability Model: LPM

Furthermore, the above problems with the LPM (3 in particular) motivate non-linear models, such as:

- Probit and logit models are a very common type of model in binary choice.
- Multinominal probit and logit models are useful for multiple choices (which university to go to?).
- Other limited dependent variable models include: tobit models (how much do people give to charity - lots of zeros, plus some positive values) and count data models (how many GCSE's did you achieve: 0, 1, 2, 3,....)
- Essentially, different types of dependent variable can sometimes require different estimation methods.
- They typically utilise maximum likelihood estimation as an estimation strategy, as opposed to OLS.

However, the LPM is still widely used and is important to understand.