Binary Response Models

Justin Raymond S. Eloriaga

2021

Chapter Summary

The first departure from the CLRM we will discuss are the Binary Response Models. We use this when the dependent variable is some dummy. We will show that the mere usage of the OLS (i.e. LPM) is heavily biased. However, interpreting the coefficients in the appropriate models such as the Logit and Probit is not always straightforward.

Recall a Simple Bivariate CLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Recall a Simple Bivariate CLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- What we assumed so far is that y_i is a continuous quantitative variable (an interval or a ratio). But what if we want to analyze a dependent variable that is a dummy variable?

Recall a Simple Bivariate CLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- What we assumed so far is that y_i is a continuous quantitative variable (an interval or a ratio). But what if we want to analyze a dependent variable that is a dummy variable?
- For example, say we want to analyze y_i where y_i is 1 if the student belongs to a feeder school and 0 if not.

Recall a Simple Bivariate CLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- What we assumed so far is that y_i is a continuous quantitative variable (an interval or a ratio). But what if we want to analyze a dependent variable that is a dummy variable?
- For example, say we want to analyze y_i where y_i is 1 if the student belongs to a feeder school and 0 if not.
- Let us pose a research question. "Does having rich parents increase the likelihood of attending a feeder high school?"

Recall a Simple Bivariate CLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- What we assumed so far is that y_i is a continuous quantitative variable (an interval or a ratio). But what if we want to analyze a dependent variable that is a dummy variable?
- For example, say we want to analyze y_i where y_i is 1 if the student belongs to a feeder school and 0 if not.
- Let us pose a research question. "Does having rich parents increase the likelihood of attending a feeder high school?"

$$Feeder_i = \beta_0 + \beta_1 \ln(ParentWage_i) + u_i$$

On Expectations

Recall that a regression is merely trying to get the expected value of y given the x's. How does that change when y is no just 1 or 0?

On Expectations

Recall that a regression is merely trying to get the expected value of y given the x's. How does that change when y is no just 1 or 0?

$$E(y|x) = \sum_i P(y = y_i)y_i$$

On Expectations

Recall that a regression is merely trying to get the expected value of y given the x's. How does that change when y is no just 1 or 0?

$$E(y|x) = \sum_i P(y = y_i)y_i$$

This means that (since y can only be zero or one):

$$E(y|x) = P(y = 0|x) \cdot 0 + P(y = 1|x) \cdot 1$$

On Expectations

Recall that a regression is merely trying to get the expected value of y given the x's. How does that change when y is no just 1 or 0?

$$E(y|x) = \sum_i P(y = y_i)y_i$$

This means that (since y can only be zero or one):

$$E(y|x) = P(y = 0|x) \cdot 0 + P(y = 1|x) \cdot 1$$

Therefore

$$\boxed{E(y|x) = P(y = 1|x)}$$

On Marginal Effects

Using what we derived so far (and the fact that $E(u) = 0$), this basically means that

$$E(y|x) = \beta_0 + \beta_1 x = P(y = 1|x)$$

On Marginal Effects

Using what we derived so far (and the fact that $E(u) = 0$), this basically means that

$$E(y|x) = \beta_0 + \beta_1 x = P(y = 1|x)$$

How do we interpret the β_1?

On Marginal Effects

Using what we derived so far (and the fact that $E(u) = 0$), this basically means that

$$E(y|x) = \beta_0 + \beta_1 x = P(y = 1|x)$$

How do we interpret the β_1?

- β_1 is merely the change in the probability that $y = 1$ when x changes.

$$\Delta P(y = 1|x)|_{x \to x+1} = \beta_1$$

Circling back to Feeder

Relating it to our example on feeder schools

$$E(Feeder_i | \ln(ParentWage_i)) = P(Feeder_i = 1 | \ln(ParentWage_i))$$

Circling back to Feeder

Relating it to our example on feeder schools

$$E(Feeder_i|\ln(ParentWage_i)) = P(Feeder_i = 1|\ln(ParentWage_i))$$

But since $E(y|x) = \beta_0 + \beta_1 x = P(y = 1|x)$

$$E(Feeder_i|\ln(ParentWage_i)) = \beta_0 + \beta_1 \ln(ParentWage_i)$$

Circling back to Feeder

Relating it to our example on feeder schools

$$E(Feeder_i|\ln(ParentWage_i)) = P(Feeder_i = 1|\ln(ParentWage_i))$$

But since $E(y|x) = \beta_0 + \beta_1 x = P(y = 1|x)$

$$E(Feeder_i|\ln(ParentWage_i)) = \beta_0 + \beta_1 \ln(ParentWage_i)$$

Important Question

Can we merely use OLS to estimate this model given that it looks seemingly like a simple CLRM?

Circling back to Feeder

Relating it to our example on feeder schools

$$E(Feeder_i|\ln(ParentWage_i)) = P(Feeder_i = 1|\ln(ParentWage_i))$$

But since $E(y|x) = \beta_0 + \beta_1 x = P(y = 1|x)$

$$E(Feeder_i|\ln(ParentWage_i)) = \beta_0 + \beta_1 \ln(ParentWage_i)$$

Important Question

Can we merely use OLS to estimate this model given that it looks seemingly like a simple CLRM?

Answer: No! Let's discuss three reasons why not

Prediction Outside [0,1]

Prediction Outside [0,1]

Recall that a probability should only lie between zero and one. There is no such thing as a negative probability or a probability greater than one.

Prediction Outside [0,1]

Recall that a probability should only lie between zero and one. There is no such thing as a negative probability or a probability greater than one.

- Say for example that $\beta_0 = 0.5$ and β_1 is 0.2 which is associated with $\ln(ParentWage_i)$.

Prediction Outside [0,1]

Recall that a probability should only lie between zero and one. There is no such thing as a negative probability or a probability greater than one.

- Say for example that $\beta_0 = 0.5$ and β_1 is 0.2 which is associated with $\ln(ParentWage_i)$.
- If the $\ln(ParentWage_i) = 10$, then we get

$$Feeder_i = 0.3 + 0.2 \ln(ParentWage_i) = 0.3 + 0.2 \cdot 10 = 2.3$$

Prediction Outside [0,1]

Recall that a probability should only lie between zero and one. There is no such thing as a negative probability or a probability greater than one.

- Say for example that $\beta_0 = 0.5$ and β_1 is 0.2 which is associated with $\ln(ParentWage_i)$.
- If the $\ln(ParentWage_i) = 10$, then we get

$$Feeder_i = 0.3 + 0.2 \ln(ParentWage_i) = 0.3 + 0.2 \cdot 10 = 2.3$$

- If the $\ln(ParentWage_i) = -10$, then we get

$$Feeder_i = 0.3 + 0.2 \ln(ParentWage_i) = 0.3 + 0.2 \cdot -10 = -1.7$$

Prediction Outside [0,1]

Recall that a probability should only lie between zero and one. There is no such thing as a negative probability or a probability greater than one.

- Say for example that $\beta_0 = 0.5$ and β_1 is 0.2 which is associated with $\ln(ParentWage_i)$.
- If the $\ln(ParentWage_i) = 10$, then we get

$$Feeder_i = 0.3 + 0.2 \ln(ParentWage_i) = 0.3 + 0.2 \cdot 10 = 2.3$$

- If the $\ln(ParentWage_i) = -10$, then we get

$$Feeder_i = 0.3 + 0.2 \ln(ParentWage_i) = 0.3 + 0.2 \cdot -10 = -1.7$$

- This is nonsensical! We cannot have predicted values like that.

Errors are Heteroscedastic

Errors are Heteroscedastic

When y can only take a value of zero or one, this means that the error term is not going to be homoscedastic.

Errors are Heteroscedastic

When y can only take a value of zero or one, this means that the error term is not going to be homoscedastic.

- Recall that $y_i = \beta x_i + u_i$

Errors are Heteroscedastic

When y can only take a value of zero or one, this means that the error term is not going to be homoscedastic.

- Recall that $y_i = \beta x_i + u_i$
- If $y_i = 0$, this means that $u_i = -\beta x_i$. If $y_i = 1$, this means that $u_i = 1 - \beta x_i$

Errors are Heteroscedastic

When y can only take a value of zero or one, this means that the error term is not going to be homoscedastic.

- Recall that $y_i = \beta x_i + u_i$
- If $y_i = 0$, this means that $u_i = -\beta x_i$. If $y_i = 1$, this means that $u_i = 1 - \beta x_i$
- The assumption that $E(u_i) = 0$ still holds

Errors are Heteroscedastic

When y can only take a value of zero or one, this means that the error term is not going to be homoscedastic.

- Recall that $y_i = \beta x_i + u_i$
- If $y_i = 0$, this means that $u_i = -\beta x_i$. If $y_i = 1$, this means that $u_i = 1 - \beta x_i$
- The assumption that $E(u_i) = 0$ still holds

$$u_i = \begin{cases} -\beta x_i & \text{if } y_i = 0 \\ 1 - \beta x_i & \text{if } y_i = 1 \end{cases}$$

Proof for Heteroscedastic Errors

$$Var(u_i|x_i) = E(u_i^2|x_i)$$

Proof for Heteroscedastic Errors

$$Var(u_i|x_i) = E(u_i^2|x_i)$$

- Recall that $Var(u_i) = E[u_i - E(u_i)]^2 = E[u_i]^2$. In a similar vein, we can expand this as follows like what we had in the proof earlier.

$$Var(u_i) = \sum_j P(y_i = y_j) \cdot u_i^2$$

Proof for Heteroscedastic Errors

$$Var(u_i|x_i) = E(u_i^2|x_i)$$

- Recall that $Var(u_i) = E[u_i - E(u_i)]^2 = E[u_i]^2$. In a similar vein, we can expand this as follows like what we had in the proof earlier.

$$Var(u_i) = \sum_j P(y_i = y_j) \cdot u_i^2$$

- Therefore, we can calculate the variance as just

$$Var(u_i|x_i) = P(y_i = 0|x_i) \cdot (-\beta x_i)^2 + P(y_i = 1|x_i) \cdot (1 - \beta x_i)^2$$

Proof for Heteroscedastic Errors

$$Var(u_i|x_i) = E(u_i^2|x_i)$$

- Recall that $Var(u_i) = E[u_i - E(u_i)]^2 = E[u_i]^2$. In a similar vein, we can expand this as follows like what we had in the proof earlier.

$$Var(u_i) = \sum_j P(y_i = y_j) \cdot u_i^2$$

- Therefore, we can calculate the variance as just

$$Var(u_i|x_i) = P(y_i = 0|x_i) \cdot (-\beta x_i)^2 + P(y_i = 1|x_i) \cdot (1 - \beta x_i)^2$$

- Let $P_i = \beta x_i = P(y_i = 1|x_i)$. Also note that $P(y_i = 0|x_i) + P(y_i = 1|x_i) = 1$. Therefore $P(y_i = 0|x_i) = 1 - P(y_i = 1|x_i)$

Proof for Heteroscedastic Errors

Proof for Heteroscedastic Errors

- Rewriting, we get

$$Var(u_i|x_i) = (1 - P_i) \cdot (-\beta x_i)^2 + P_i \cdot (1 - \beta x_i)^2$$

Proof for Heteroscedastic Errors

- Rewriting, we get

$$Var(u_i|x_i) = (1 - P_i) \cdot (-\beta x_i)^2 + P_i \cdot (1 - \beta x_i)^2$$

- Since we know that $P_i = \beta x_i = P(y_i = 1|x_i)$

$$(1 - \beta x_i) \cdot (-\beta x_i)^2 + (\beta x_i) \cdot (1 - \beta x_i)^2$$

Proof for Heteroscedastic Errors

- Rewriting, we get

$$Var(u_i|x_i) = (1 - P_i) \cdot (-\beta x_i)^2 + P_i \cdot (1 - \beta x_i)^2$$

- Since we know that $P_i = \beta x_i = P(y_i = 1|x_i)$

$$(1 - \beta x_i) \cdot (-\beta x_i)^2 + (\beta x_i) \cdot (1 - \beta x_i)^2$$

- Factoring things out

$$(1 - \beta x_i)\beta x_i[\beta x_i + 1 - \beta x_i]$$

Proof for Heteroscedastic Errors

- Rewriting, we get

$$Var(u_i|x_i) = (1 - P_i) \cdot (-\beta x_i)^2 + P_i \cdot (1 - \beta x_i)^2$$

- Since we know that $P_i = \beta x_i = P(y_i = 1|x_i)$

$$(1 - \beta x_i) \cdot (-\beta x_i)^2 + (\beta x_i) \cdot (1 - \beta x_i)^2$$

- Factoring things out

$$(1 - \beta x_i)\beta x_i[\beta x_i + 1 - \beta x_i]$$

- Simplifying, we get

$$\boxed{Var(u_i|x_i) = (1 - \beta x_i)\beta x_i}$$

Proof for Heteroscedastic Errors

- Rewriting, we get

$$Var(u_i|x_i) = (1 - P_i) \cdot (-\beta x_i)^2 + P_i \cdot (1 - \beta x_i)^2$$

- Since we know that $P_i = \beta x_i = P(y_i = 1|x_i)$

$$(1 - \beta x_i) \cdot (-\beta x_i)^2 + (\beta x_i) \cdot (1 - \beta x_i)^2$$

- Factoring things out

$$(1 - \beta x_i)\beta x_i[\beta x_i + 1 - \beta x_i]$$

- Simplifying, we get

$$\boxed{Var(u_i|x_i) = (1 - \beta x_i)\beta x_i}$$

- The variance is clearly dependent on i and is not a constant. Therefore, the variance is heteroscedastic (not constant).

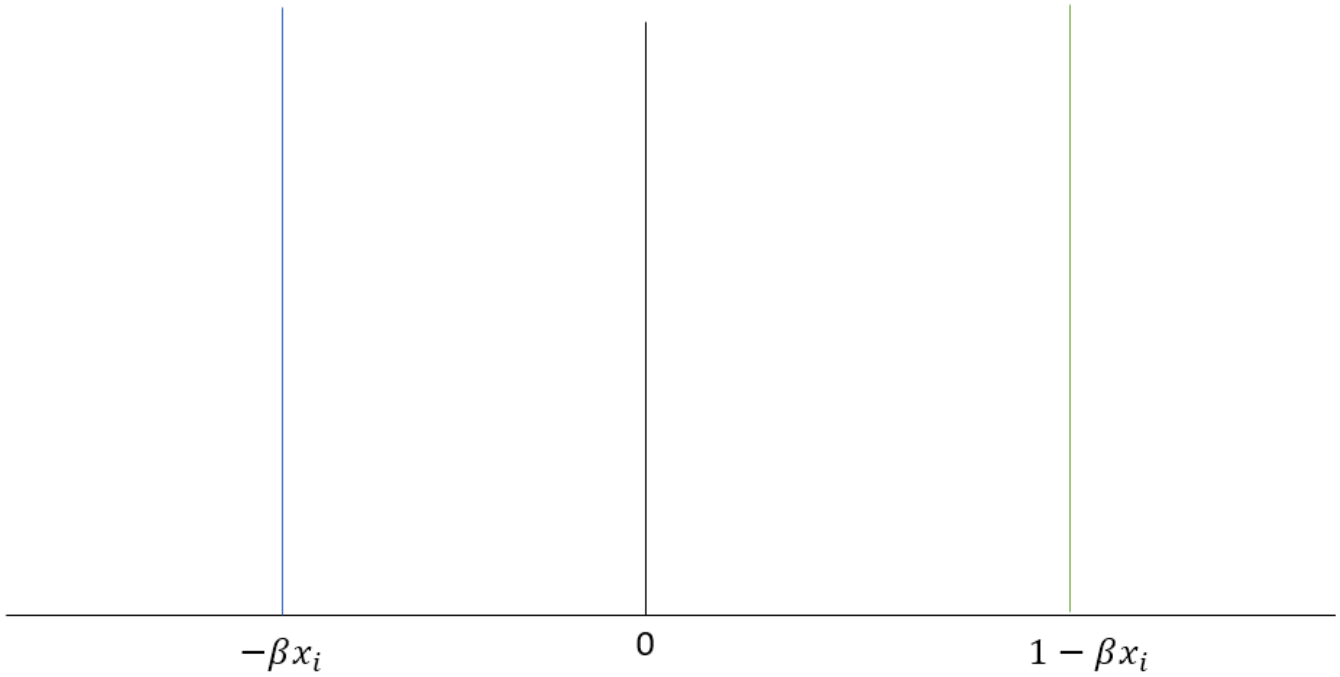
Errors are not Normally Distributed

Errors are not Normally Distributed

Before, the assumption was $u_i \sim iidN(0, \sigma^2)$. However, when the dependent variable is a dummy, the residuals are *Bernoulli Distributed*, hence, can only take two values.

Errors are not Normally Distributed

Before, the assumption was $u_i \sim iidN(0, \sigma^2)$. However, when the dependent variable is a dummy, the residuals are *Bernoulli Distributed*, hence, can only take two values.



How do we overcome these limitations

How do we overcome these limitations

Recall that $P(Feeder_i = 1|\ln(ParentWage_i) = \beta_0 + \beta_1 \ln(ParentWage_i)$.

How do we overcome these limitations

Recall that $P(Feeder_i = 1|\ln(ParentWage_i) = \beta_0 + \beta_1 \ln(ParentWage_i)$.

- When we use OLS, we know that it is problematic since it may give a predicted value that is outside the closed interval [0,1].

$$-\infty < \beta_0 + \beta_1 \ln(ParentWage_i) < +\infty$$

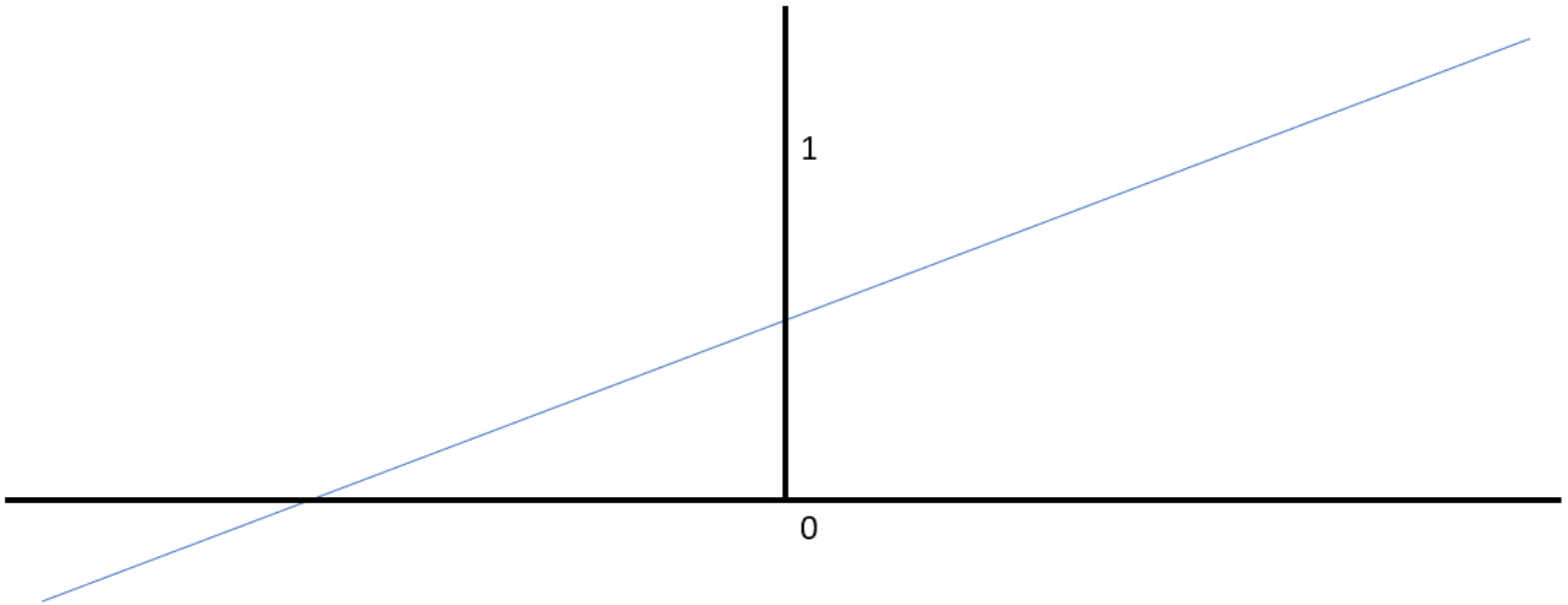
How do we overcome these limitations

Recall that $P(Feeder_i = 1|\ln(ParentWage_i) = \beta_0 + \beta_1 \ln(ParentWage_i)$.

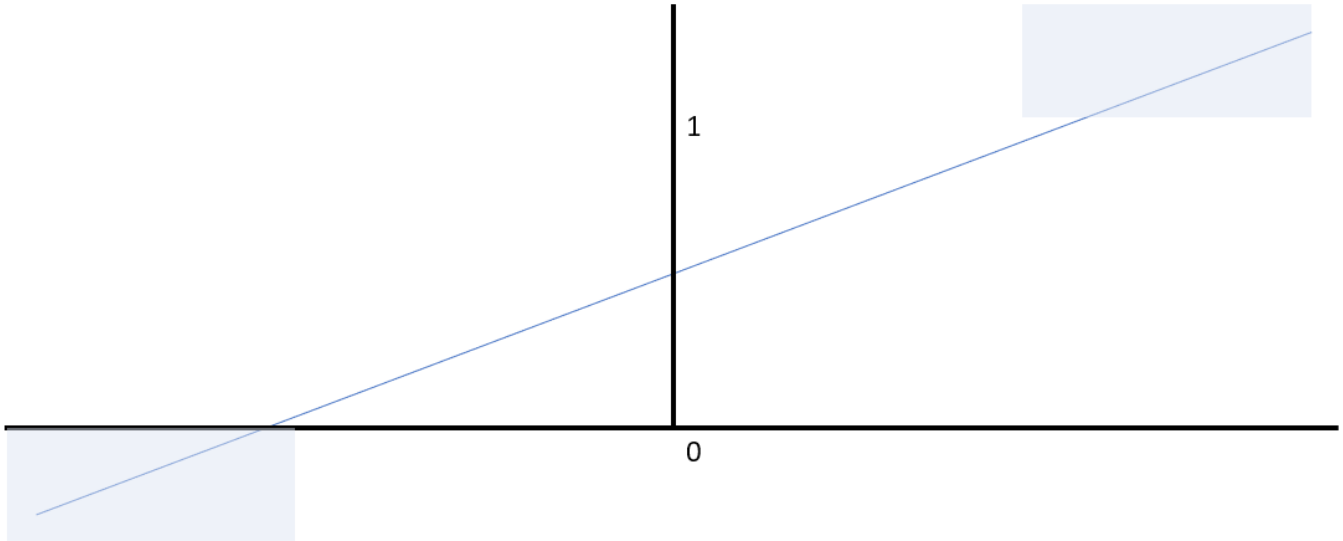
- When we use OLS, we know that it is problematic since it may give a predicted value that is outside the closed interval [0,1].

$$-\infty < \beta_0 + \beta_1 \ln(ParentWage_i) < +\infty$$

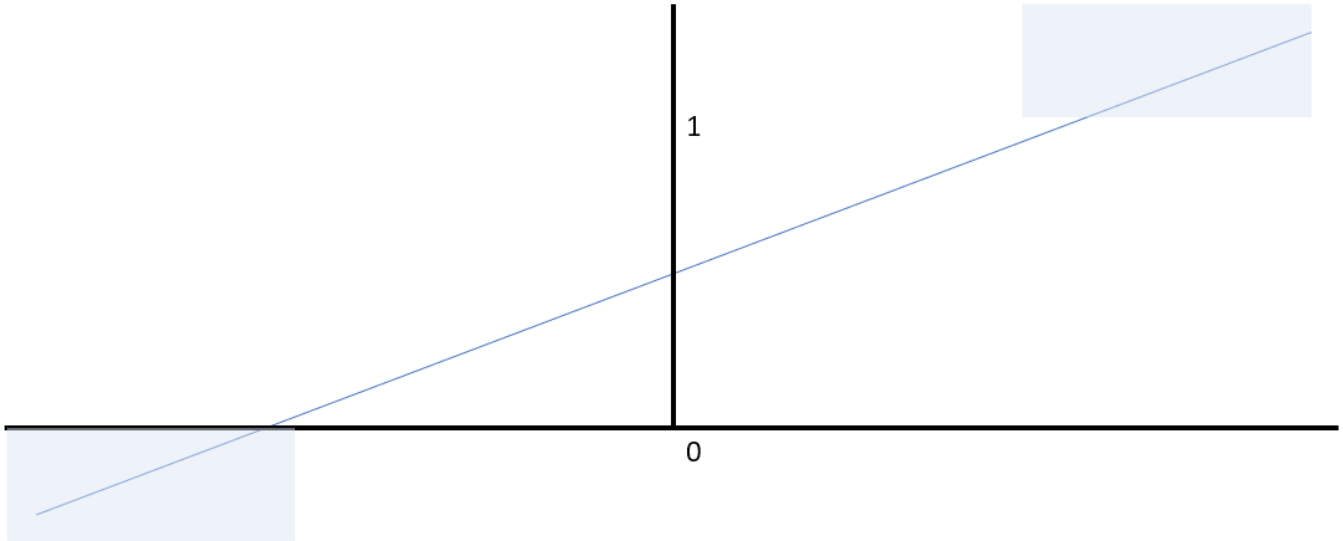
- Graphically...



Problem with the OLS Line

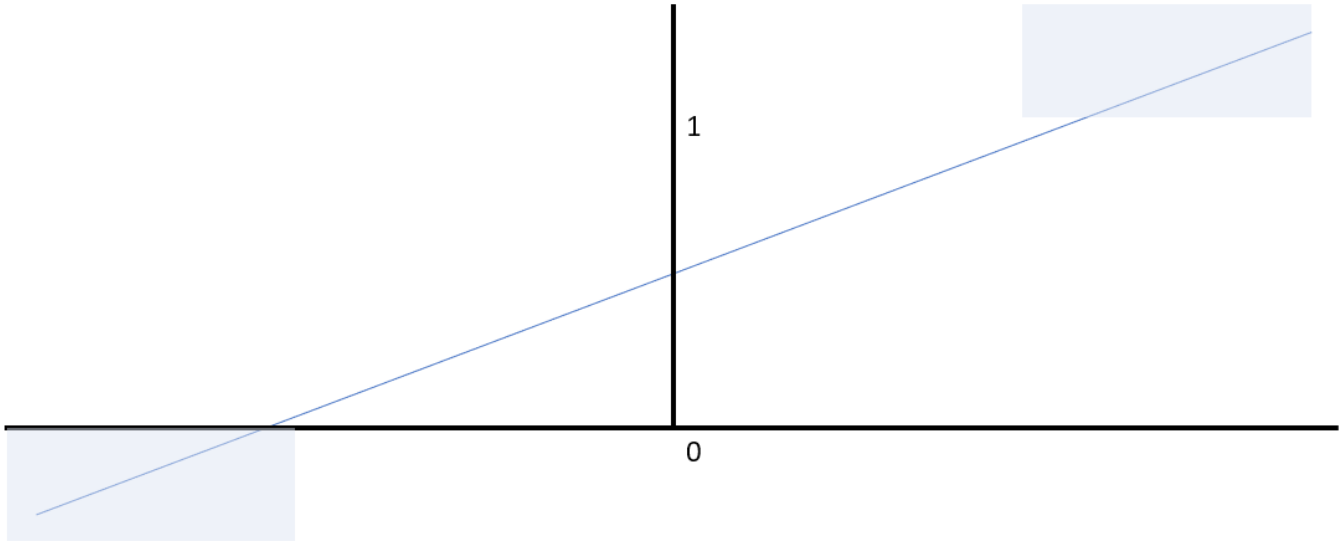


Problem with the OLS Line



The problem are the shaded regions which indicate a probability value that is nonsensical (i.e. negative or greater than one).

Problem with the OLS Line



The problem are the shaded regions which indicate a probability value that is nonsensical (i.e. negative or greater than one).

Idea

We need some way to shrink and force the line to fit between zero and one. That shrinkage will solve the issue and that is precisely what Logit and Probit are.

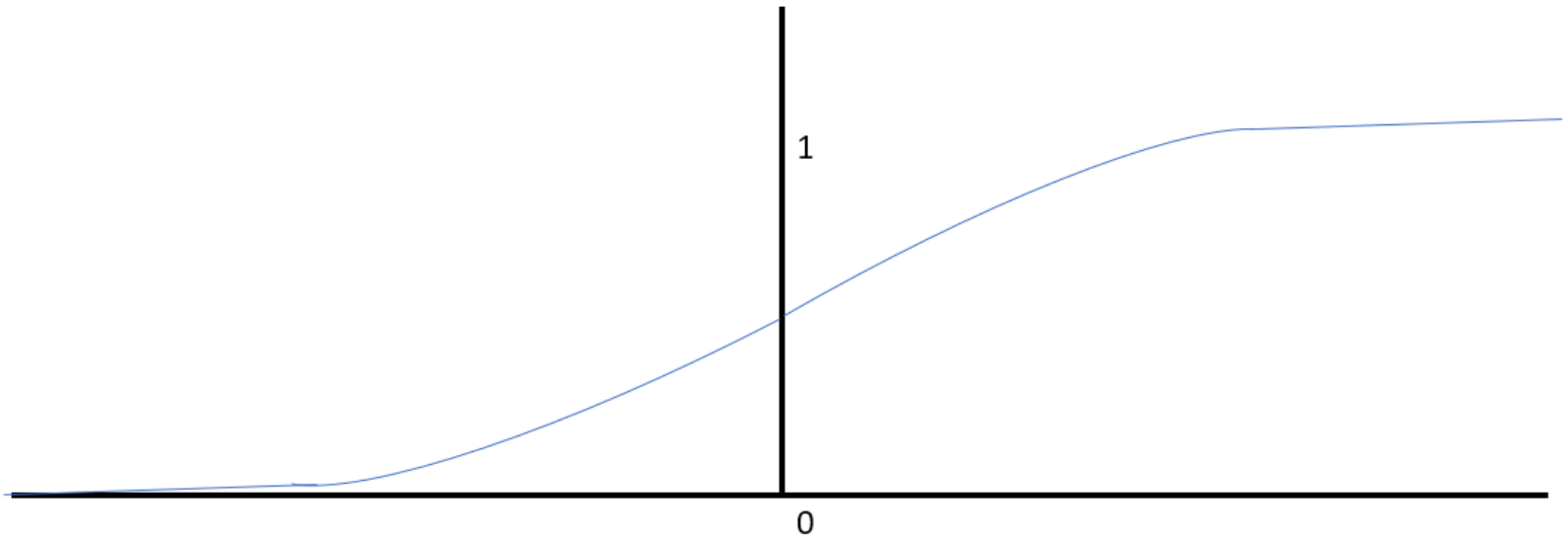
Logit and Probit at a Glance

Logit and Probit at a Glance

Essentially, Logit and Probit are non linear transformations of the regression equation such that it creates some non linear function $F(z)$ where $F(-\infty) = 0$ and $F(+\infty) = 1$.

Logit and Probit at a Glance

Essentially, Logit and Probit are non linear transformations of the regression equation such that it creates some non linear function $F(z)$ where $F(-\infty) = 0$ and $F(+\infty) = 1$.



On Logit

On Logit

The Logit model uses the *Logistic Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \frac{exp(z)}{1 + exp(z)} = L(z)$$

On Logit

The Logit model uses the *Logistic Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \frac{exp(z)}{1 + exp(z)} = L(z)$$

- z is essentially your regression equation βx. Since we know that it is equal to the predicted y, we observe the following.

On Logit

The Logit model uses the *Logistic Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \frac{exp(z)}{1 + exp(z)} = L(z)$$

- z is essentially your regression equation βx. Since we know that it is equal to the predicted y, we observe the following.
- First, as $z \to -\infty$, the $F(z) \to 0$.

On Logit

The Logit model uses the *Logistic Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \frac{exp(z)}{1 + exp(z)} = L(z)$$

- z is essentially your regression equation βx. Since we know that it is equal to the predicted y, we observe the following.
- First, as $z \to -\infty$, the $F(z) \to 0$.
- Second, as the $z \to +\infty$, the $F(z) \to 1$.

On Probit

On Probit

The Probit model uses the *Standard Normal Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \int_{-\infty}^{z} \phi(u)du$$

On Probit

The Probit model uses the *Standard Normal Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \int_{-\infty}^{z} \phi(u)du$$

- z is essentially your regression equation βx. Since we know that it is equal to the predicted y, we observe the following.

On Probit

The Probit model uses the *Standard Normal Cumulative Distribution Function* as the shrinkage function by which it can perform the model. The function, say $F(z)$, is given as

$$F(z) = \int_{-\infty}^{z} \phi(u)du$$

- z is essentially your regression equation βx. Since we know that it is equal to the predicted y, we observe the following.
- Like Logit, we can predict some probability for a predicted y (until z). But since this is a cdf, we know that if we evaluate $F(-\infty)$, this will tend to zero. Conversely, evaluating $F(+\infty) = 1$.