Limited Dependent Variables

November 8, 2022

Objectives

- Distinguish between different dependent variable structures;
- Revisit the assumptions of OLS;
- Learn (more of) the limitations of OLS;

So far you've dealt with:

1. Model a linear relationship between dependent variable y and one or more independent variables $x_1, x_2, x_3, \ldots$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \varepsilon;$$

2. Model a NON-linear relationship between dependent variable y and many independent variables x:

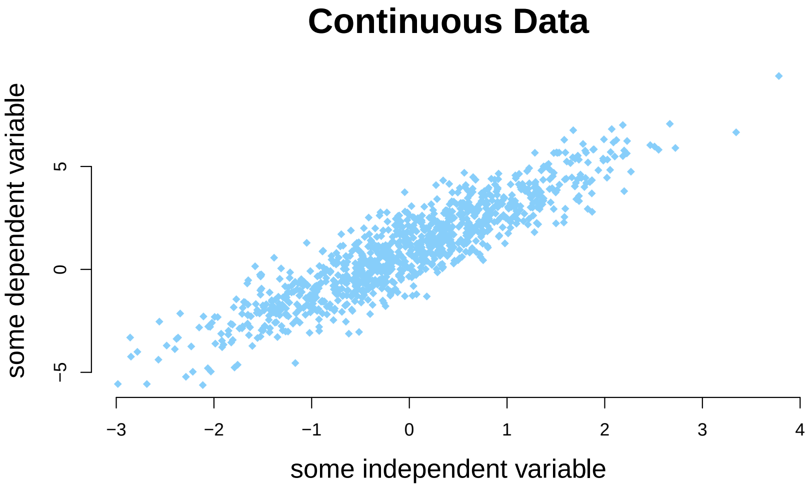
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \varepsilon;$$

3. Model relationships between independent variables:

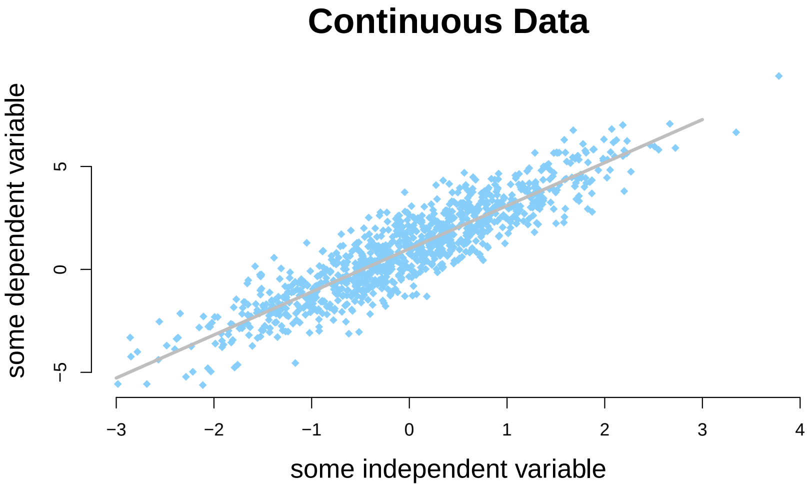
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon;$$

4. Model pseudo-experimental data using difference in differences.

So far you've dealt with:



So far you've dealt with:

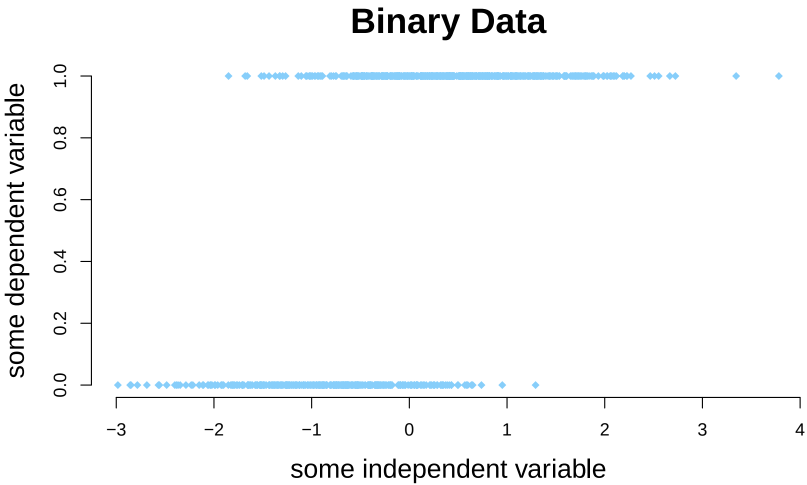


So far you've dealt with:

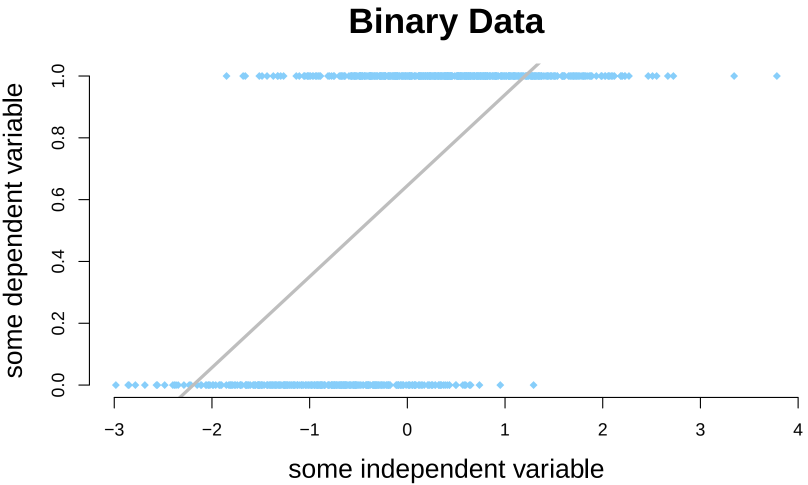
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9689	0.0305	31.76	0.0000
x	1.9734	0.0301	65.57	0.0000

actual relationship: $y = 1 + 2x + \varepsilon$

Ok, but what about...



Ok, but what about...



Ok, but what about...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6656	0.0123	54.21	0.0000
x	0.2630	0.0121	21.71	0.0000

actual relationship: $y = f(1 + 2x)$... uh oh!

And maybe even worse...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.0014	1.8726	9.61	0.0000
x	32.1421	1.8476	17.40	0.0000

actual relationship: $y = f(1 + 2x)$... yikes!

So can't we just use OLS?

Why do we choose to use OLS? Because...

So can't we just use OLS?

Why do we choose to use OLS? Because... when its assumptions are satisfied it's BLUE!

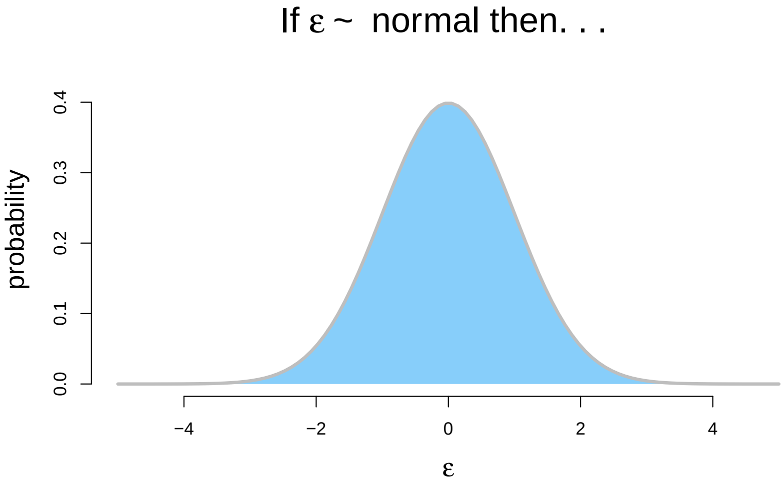
1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

2. Independent variables are not related to each other – **no multicollinearity**;
3. Independent variables have **no measurement error**;
4. Noise term is a random variable following the **normal distribution**;

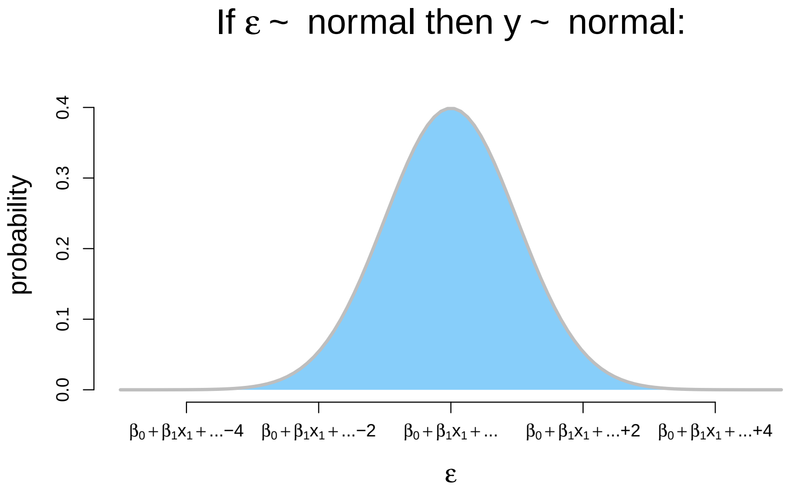
So can't we just use OLS?

Why do we choose to use OLS? Because... when its assumptions are satisfied it's BLUE!



So can't we just use OLS?

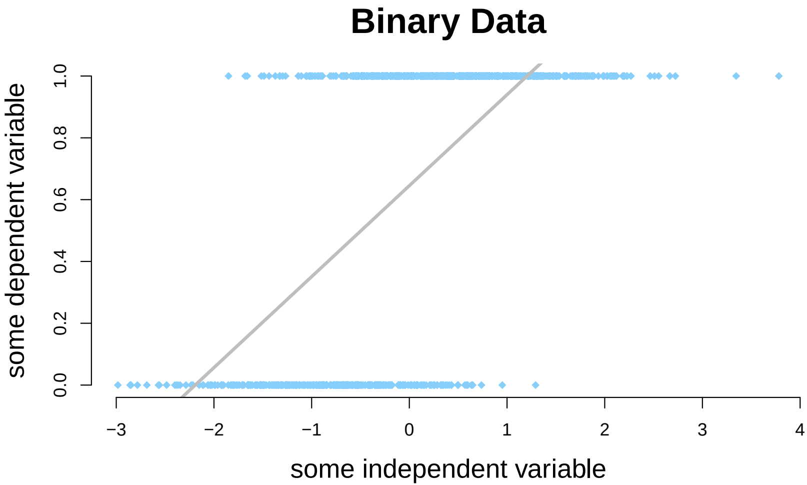
Why do we choose to use OLS? Because... when its assumptions are satisfied it's BLUE!



What does this mean? On the theory side:

- The **linear function** and **normal errors** assumptions require that y be able to take on any value!
- If the dependent variable is binary, i.e. always either 0 or 1 then...
  1. either **linear function** or **normal errors** are **wrong**, or...
  2. something exceedingly unlikely happened.

What does this mean? On the practical side:



What does this mean? On the practical side:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6656	0.0123	54.21	0.0000
x	0.2630	0.0121	21.71	0.0000

Let's use this to make predictions...

If $x =$	$y = 0.6656 + 0.2630x$	then $y = \ldots$
0	$y = 0.6656 + 0.2630 * 0$	0.6656

What does this mean? On the practical side:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6656	0.0123	54.21	0.0000
x	0.2630	0.0121	21.71	0.0000

Let's use this to make predictions...

If $x =$	$y = 0.6656 + 0.2630x$	then $y = \ldots$
0	$y = 0.6656 + 0.2630 * 0$	0.6656
1	$y = 0.6656 + 0.2630 * 1$	0.9286
2	$y = 0.6656 + 0.2630 * 2$	1.1916
-3	$y = 0.6656 + 0.2630 * (-3)$	-0.1234

What does this mean? On the practical side:

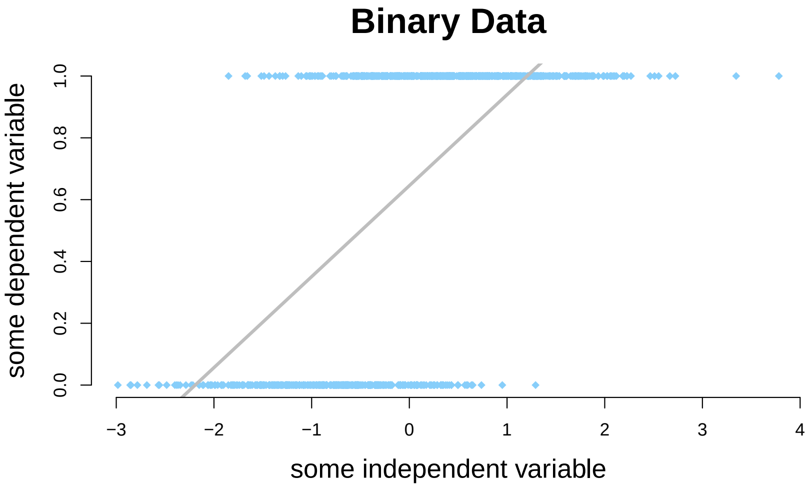
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6656	0.0123	54.21	0.0000
x	0.2630	0.0121	21.71	0.0000

Let's use this to make predictions...

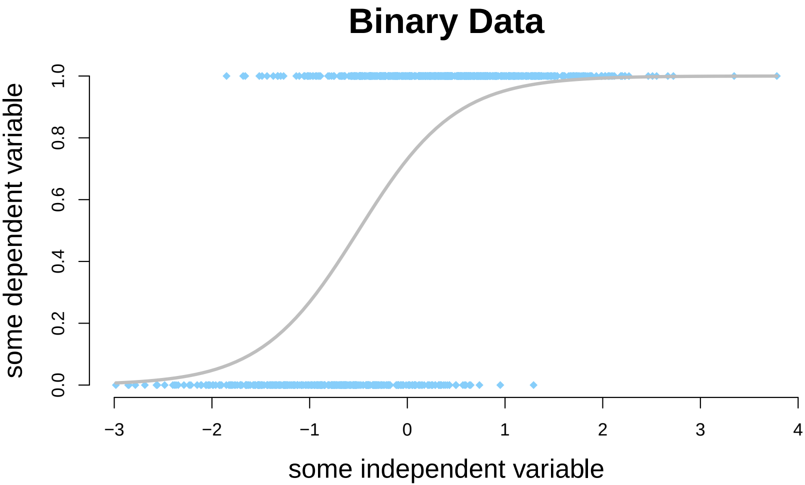
If $x =$	$y = 0.6656 + 0.2630x$	then $y = \ldots$
0	$y = 0.6656 + 0.2630 * 0$	0.6656
1	$y = 0.6656 + 0.2630 * 1$	0.9286
2	$y = 0.6656 + 0.2630 * 2$	1.1916
-3	$y = 0.6656 + 0.2630 * (-3)$	-0.1234

Leads to nonsense!

What does this mean? OLS is not the best model.



What does this mean? OLS is not the best model.



So how do we deal with this?

A very general way of addressing this type of problem (weird dependent variable) is to use a **Generalized Linear Model** (GLM).

So how do we deal with this?

A very general way of addressing this type of problem (weird dependent variable) is to use a **Generalized Linear Model** (GLM).

GLMs have three required components:

1. A probability distribution that describes the dependent variable;
2. A linear model $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$;
3. A link function that relates the linear model to the dependent variable distribution.

Binary data: GLM = Logistic regression;

Why should we care?

Limited dependent variables require different modeling strategies – we'll explore one of them next week.