STA 235H - Multiple Regression: Binary Outcomes

Fall 2023

McCombs School of Business, UT Austin

Binary Outcomes

- You have probably used binary outcomes in regressions, but do you know the issues that they may bring to the table?

What can we do about them?

How to handle binary outcomes?

Linear Probability Model

Logistic Regression

Linear Probability Models

- A Linear Probability Model is just a traditional regression with a binary outcome
- Something interesting about a binary outcome is that the expected value of Y if Y is binary is actually a probability!

$$E[Y|X_1,\dots,X_P] = Pr(Y=0|X_1,\dots,X_p)\cdot 0 + Pr(Y=1|X_1,\dots,X_p)\cdot 1$$

$$= Pr(Y=1|X_1,\dots,X_p)$$

How to interpret a LPM?

- $\hat{\beta}$'s interpreted as change in probability

$$E[Y|X_1,\ldots,X_P] = Pr(Y=0|X_1,\ldots,X_p)\cdot 0 + Pr(Y=1|X_1,\ldots,X_p)\cdot 1$$
$$= Pr(Y=1|X_1,\ldots,X_p)$$

- Example:

$$GradeA = \beta_0 + \beta_1 \cdot Study + \varepsilon$$

- $\hat{\beta}_1$ is the average change in probability of getting an A if I study one more hour.
- Studying one more hour is associated with an average increase in the probability of getting an A of $\hat{\beta}_1 \times 100$ percentage points.

Side note: Difference between percent change and change in percentage points

- Imagine that if you study 4hrs your probability of getting an A is, on average, 70% and if you study for 5hrs that probability increases to 75%.

- Then, we can say that your probability increased by 5 percentage points.

- Why is this not the same as saying that your probability increased by 5%?

- Remember percent change?

$$\frac{y_1 - y_0}{y_0} = \frac{75 - 70}{70} = 0.0714$$

- This means that, in this case, a 5 percentage point increase is equivalent to a 7% increase in probability.

Be aware of the difference in percentage points and percent!

Let's look at an example

- Home Mortgage Disclosure Act Data (HMDA) from the AER package

```
library(AER)

data("HMDA")

hmda <- data.frame(HMDA)

head(hmda)
```

```
##   deny pirat hirat     lvrat chist mhist phist unemp selfemp insurance condomin
## 1   no 0.221 0.221 0.8000000     5     2    no   3.9      no        no       no
## 2   no 0.265 0.265 0.9218750     2     2    no   3.2      no        no       no
## 3   no 0.372 0.248 0.9203980     1     2    no   3.2      no        no       no
## 4   no 0.320 0.250 0.8604651     1     2    no   4.3      no        no       no
## 5   no 0.360 0.350 0.6000000     1     1    no   3.2      no        no       no
## 6   no 0.240 0.170 0.5105263     1     1    no   3.9      no        no       no
##   afam single hschool
## 1   no     no     yes
## 2   no    yes     yes
## 3   no     no     yes
## 4   no     no     yes
## 5   no     no     yes
## 6   no     no     yes
```

Probability of someone getting a mortgage loan denied?

- Getting mortgage denied (1) based on race, conditional on payments to income ratio (`pirat`)

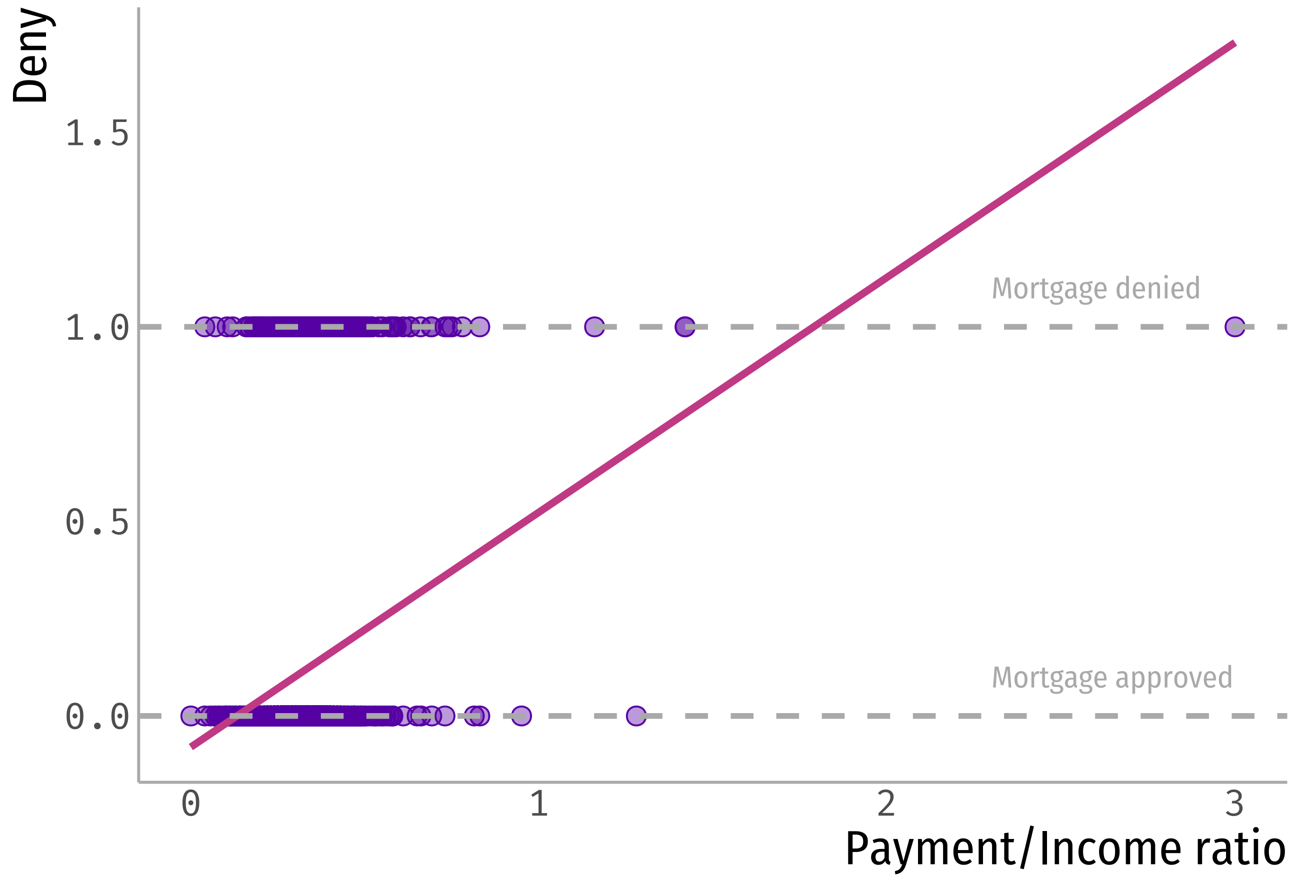
```
hmda <- hmda %>% mutate(deny = as.numeric(deny) - 1)

summary(lm(deny ~ pirat + factor(afam), data = hmda))
```

```
##
## Call:
## lm(formula = deny ~ pirat + factor(afam), data = hmda)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62526 -0.11772 -0.09293 -0.05488  1.06815
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.09051    0.02079  -4.354 1.39e-05 ***
## pirat            0.55919    0.05987   9.340  < 2e-16 ***
## factor(afam)yes  0.17743    0.01837   9.659  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3123 on 2377 degrees of freedom
## Multiple R-squared:  0.076,    Adjusted R-squared:  0.07523
## F-statistic: 97.76 on 2 and 2377 DF,  p-value: < 2.2e-16
```

- Holding payment-to-income ratio constant, an AA client has a probability of getting their loan denied that is **18 pp higher**, <u>on average, than a non AA client</u>.
- Being AA is associated to an <u>average</u> increase of **0.177 in the probability** of getting a loan denied <u>compared to a non AA</u>, holding payment-to-income ratio constant.

How does this LPM look?



Issues with a LPM?

- Main problems:
  - Non-normality of the error term
  - Heteroskedasticity (i.e. variance of the error term is not constant)
  - Predictions can be outside [0,1]
  - LPM imposes linearity assumption

Issues with a LPM?

- **Main problems**:
  - Non-normality of the error term $\rightarrow$ **Hypothesis testing**
  - Heteroskedasticity $\rightarrow$ **Validity of SE**
  - Predictions can be outside [0,1] $\rightarrow$ **Issues for prediction**
  - LPM imposes linearity assumption $\rightarrow$ **Too strict?**

Are there solutions?

- Don't use small samples: With the CLT, non-normality shouldn't matter much.
- Saturate your model: In a fully saturated model (i.e. include dummies and interactions), CEF is linear.
- Use robust standard errors: Package `estimatr` in R is great!

Run again with robust standard errors

```r
library(estimatr)

model1 <- lm(deny ~ pirat + factor(afam), data = hmda)
model2 <- lm_robust(deny ~ pirat + factor(afam), data = hmda)
```

	(1)	(2)
(Intercept)	−0.091***	−0.091**
	(0.021)	(0.031)
pirat	0.559***	0.559***
	(0.060)	(0.095)
factor(afam)yes	0.177***	0.177***
	(0.018)	(0.025)
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

- Can you interpret these parameters? Do they make sense?

Most issues are solvable, but...

What about prediction?

Logistic Regression

- Typically used in the context of binary outcomes (*Probit is another popular one*)

- Nonlinear function to model the conditional probability function of a binary outcome.

$$Pr(Y=1|X_1,...,X_p) = F(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)$$

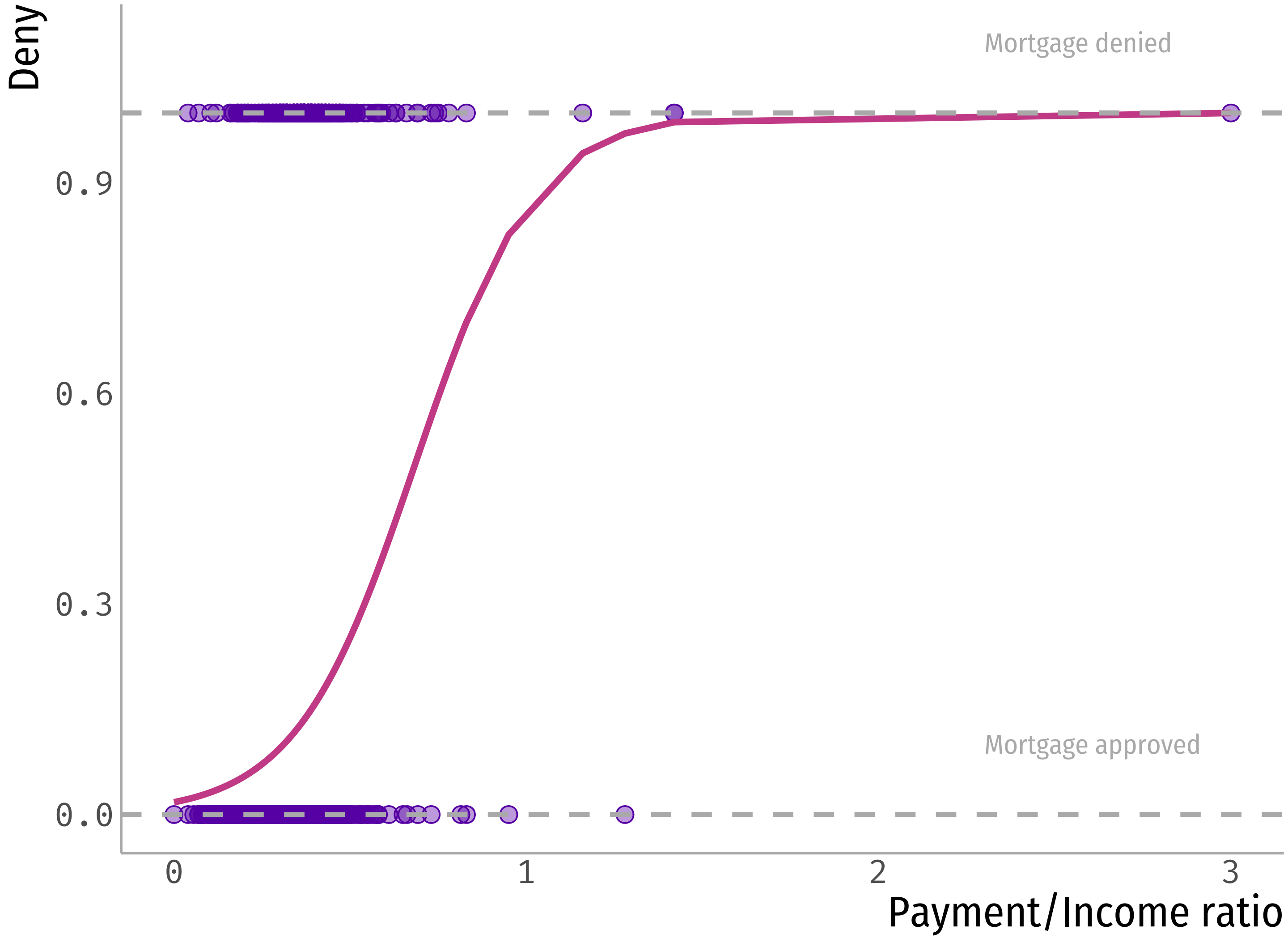
Where in a logistic regression: $F(x) = \frac{1}{1+exp(-x)}$

- *In the LPM,* $F(x) = x$

- A logistic regression doesn't look pretty:

$$Pr(Y|X_1,...,X_p) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}$$

How does this look in a plot?



When will we use logistic regression?

- As you discovered in the readings, logit is great for prediction (**much better** than LPM).
- For explanation, however, **LPM simplifies interpretation**.

Use LPM for explanation and logit for prediction

(but remember robust SE!)

Takeaway points

- Always make sure to check your data:
    - What are analyzing? Does the data behave as I would expect? Should I exclude observations?
- For LPM, always include robust standard errors!

References

- Ismay, C. & A. Kim. (2021). "Statistical Inference via Data Science". Chapter 6 & 10.