STA 235 - Binary Outcomes

Spring 2021

McCombs School of Business, UT Austin

Binary Outcomes

- So far, outcome has been a **continuous variable**.

What if the outcome is binary?

What can we do?

How to handle binary outcomes?

Linear Probability Model

Logistic Regression

Linear Probability Models (LPM)

- Just the same as the **multiple regression models** we've been seeing.

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \ldots + \beta_p \boldsymbol{X}_p + \varepsilon$$

Linear Probability Models (LPM)

- Just the same as the **multiple regression models** we've been seeing.

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \ldots + \beta_p \boldsymbol{X}_p + \varepsilon$$

- But now $Y \in \{0, 1\}$

How to interpret an LPM?

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

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

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

How to interpret an LPM?

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

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

- Example:

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

- $\hat{\beta}_1$ is the estimated change in probability of passing STA 235 if I study one more hour.

Let's look at an example

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

```
##   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?

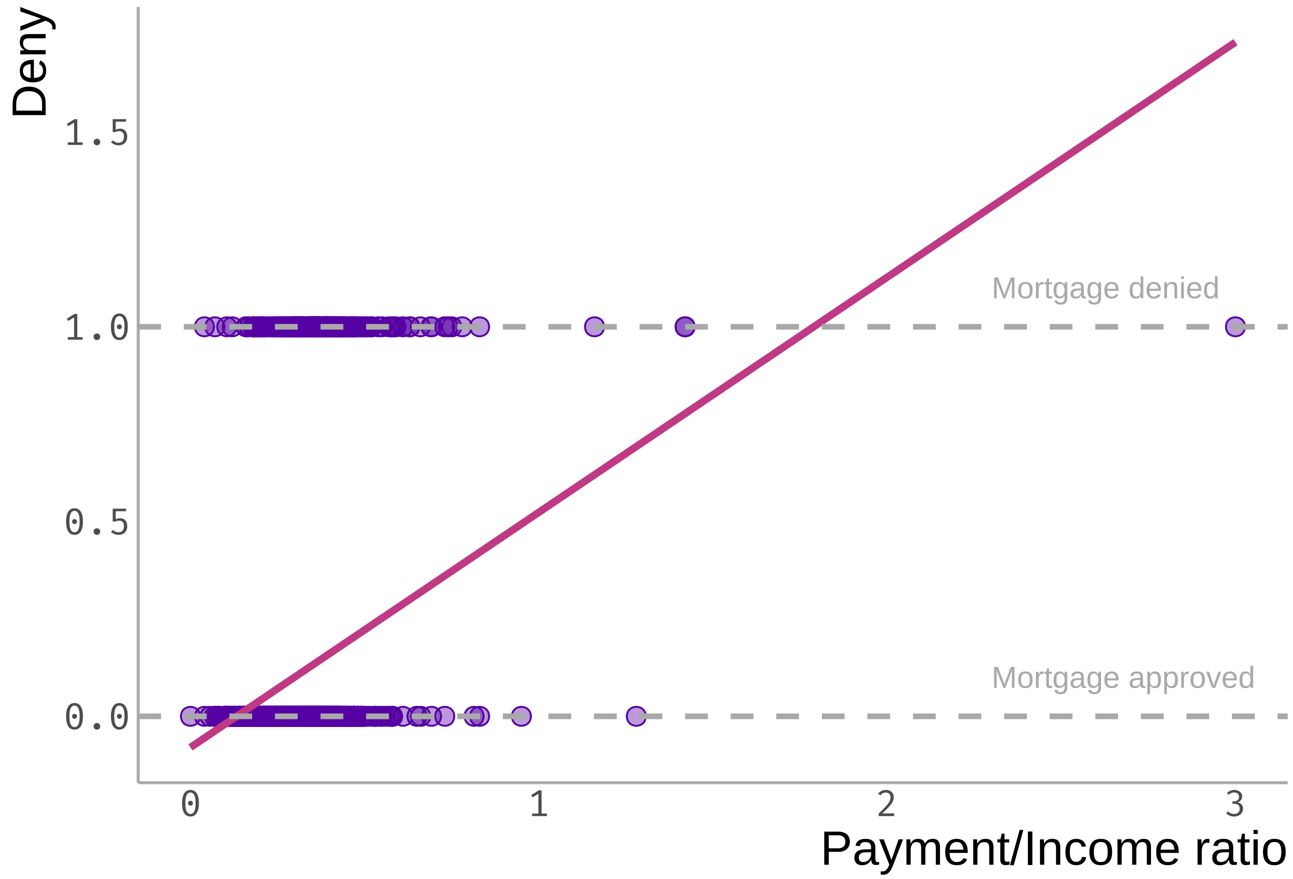
- Getting mortgage denied (1) based on payments to income ratio (`pirat`)

```r
hmda$deny = as.numeric(hmda$deny) - 1

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

```
##
## Call:
## lm(formula = deny ~ pirat, data = hmda)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73070 -0.13736 -0.11322 -0.07097  1.05577
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07991    0.02116  -3.777 0.000163 ***
## pirat        0.60353    0.06084   9.920  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3183 on 2378 degrees of freedom
## Multiple R-squared:  0.03974,    Adjusted R-squared:  0.03933
## F-statistic: 98.41 on 1 and 2378 DF,  p-value: < 2.2e-16
```

How does this LPM look?



Issues with an LPM?

- **Main problems**:
    - Non-normality of the error term
    - Heteroskedasticity
    - Predictions can be outside [0,1]
    - LPM imposes linearity assumption

Issues with an LPM?

- **Main problems**:
  - Non-normality of the error term $\longrightarrow$ **Hypothesis testing**
  - Heteroskedasticity
  - Predictions can be outside [0,1]
  - LPM imposes linearity assumption

Issues with an LPM?

- **Main problems**:
  - Non-normality of the error term $\longrightarrow$ **Hypothesis testing**
  - Heteroskedasticity $\longrightarrow$ **Validity of SE**
  - Predictions can be outside [0,1]
  - LPM imposes linearity assumption

Issues with an LPM?

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

Issues with an LPM?

- **Main problems**:
  - Non-normality of the error term → **Hypothesis testing**
  - Heteroskedasticity → **Validity of SE**
  - Predictions can be outside [0,1] → **Issues for prediction**
  - LPM imposes linearity assumption → **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!
- **Not appropriate for prediction**

Run again with robust standard errors

```r
library(estimatr)

model1 <- lm(deny ~ pirat, data = hmda)
model2 <- estimatr::lm_robust(deny ~ pirat, data = hmda)
```

	Model 1	**Model 2**
(Intercept)	-0.080***	-0.080**
	(0.021)	(0.035)
pirat	0.604***	0.604***
	(0.061)	(0.107)
R2	0.040	0.040
R2 Adj.	0.039	0.039
se_type		HC2
* p < 0.1, ** p < 0.05, *** p < 0.01		

- The default is the Bell-McCaffrey adjustment, a bias-reduced version of "robust" SE.

Let's include more covariates

```
model3 <- estimatr::lm_robust(deny ~ pirat + factor(afam), data = hmda)
```

	Model 1	Model 2	Model 3
(Intercept)	-0.080***	-0.080**	-0.091***
	(0.021)	(0.035)	(0.031)
pirat	0.604***	0.604***	0.559***
	(0.061)	(0.107)	(0.095)
factor(afam)yes			0.177***
			(0.025)
R2	0.040	0.040	0.076
R2 Adj.	0.039	0.039	0.075
se_type		HC2	HC2

* p < 0.1, ** p < 0.05, *** p < 0.01

Let's include more covariates

```
model3 <- estimatr::lm_robust(deny ~ pirat + factor(afam), data = hmda)
```

	Model 1	Model 2	Model 3
(Intercept)	-0.080***	-0.080**	-0.091***
	(0.021)	(0.035)	(0.031)
pirat	0.604***	0.604***	0.559***
	(0.061)	(0.107)	(0.095)
factor(afam)yes			0.177***
			(0.025)
R2	0.040	0.040	0.076
R2 Adj.	0.039	0.039	0.075
se_type		HC2	HC2
* p < 0.1, ** p < 0.05, *** p < 0.01			

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

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.

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

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

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

How does this look in a plot?

```r
logit1 <- glm(deny ~ pirat, family = binomial(link = "logit"),
              data = hmda)

prob <- predict(logit1, type = "response") # probabilities
```

How to interpret the coefficients?

```
summary(glm(deny ~ pirat + factor(afam), family = binomial(link = "logit"),
              data = hmda))
```

```
##
## Call:
## glm(formula = deny ~ pirat + factor(afam), family = binomial(link = "logit"),
##     data = hmda)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3709  -0.4732  -0.4219  -0.3556   2.8038
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.1256     0.2684 -15.370  < 2e-16 ***
## pirat             5.3704     0.7283   7.374 1.66e-13 ***
## factor(afam)yes   1.2728     0.1462   8.706  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1744.2  on 2379  degrees of freedom
## Residual deviance: 1591.4  on 2377  degrees of freedom
## AIC: 1597.4
##
## Number of Fisher Scoring iterations: 5
```

How to interpret the coefficients? (cont.)

- **No easy way!**
    - Coefficients in the output are **log odds ratio**

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

How to interpret the coefficients? (cont.)

- **No easy way!**
  - Coefficients in the output are **log odds ratio**

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

- With a little bit of algebra, you can solve for p:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

How to interpret the coefficients? (cont.)

- **No easy way!**
  - Coefficients in the output are **log odds ratio**

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

- With a little bit of algebra, you can solve for p:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

- **Differences are not constant for across values of X_j**

How to interpret the coefficients? (cont.)

- E.g. Choose coefficient of interest, and fix the other variables to their mean or mode:

```
logit1 <- glm(deny ~ pirat + factor(afam), family = binomial(link = "logit"),
              data = hmda)

predictions_afam <- predict(logit1, newdata = data.frame("afam" = c("no", "yes"),
                                                         "pirat" = c(mean(hmda$pirat), mean(hmda$pirat))),
                            type = "response")
predictions_afam
```

```
##          1          2
## 0.08714775 0.25422824
```

How to interpret the coefficients? (cont.)

- E.g. Choose coefficient of interest, and fix the other variables to their mean or mode:

```
logit1 <- glm(deny ~ pirat + factor(afam), family = binomial(link = "logit"),
              data = hmda)

predictions_afam <- predict(logit1, newdata = data.frame("afam" = c("no", "yes"),
                                                         "pirat" = c(mean(hmda$pirat), mean(hmda$pirat))),
                            type = "response")
predictions_afam
```

```
##          1          2 
## 0.08714775 0.25422824
```

```
diff(predictions_afam)
```

```
##         2 
## 0.1670805
```

- Remember that for the LPM model, $\hat{\beta}_{afam} = 0.177$

Wrapping things up: Which one do we choose?

- Both logit and LPM have **pros and cons**.
- A lot of the time, **depends on what you want to do**.

Wrapping things up: Which one did you choose?

	LPM for prediction	
LPM for explanation	**no**	**yes**
no	11	4
yes	11	3

Wrapping things up: Which one do we choose? (cont.)

LPM

Pros:

- Simplicity
- Interpretability

Cons:

- Cannot be used for prediction
- Robust SE

Logit

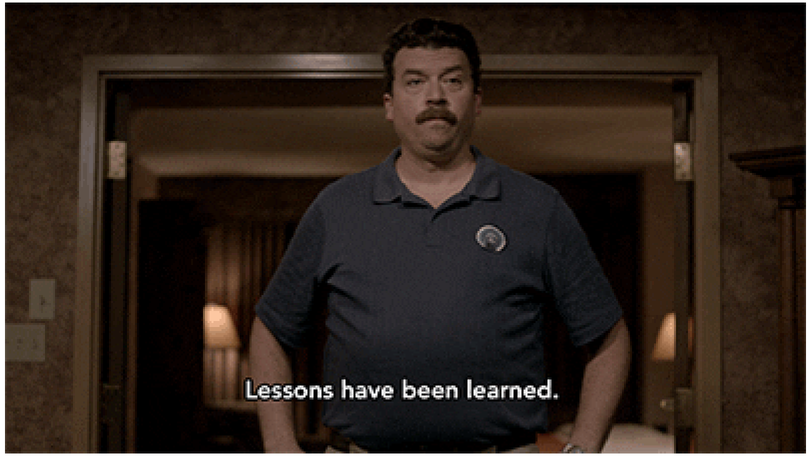
Pros:

- Bounded probabilities
- Flexibility

Cons:

- Log odds ratio
- Doesn't play well with FE

Main takeaway points



- LMP and Logistic Regression can **both be useful** depending on the context.
- **Be careful** with the interpretation!
- Remember to always **plot** your data.

Next week

- We start with:

  Causal Inference

- Homework 1 will be **posted today**.
  - Start early!
- **Readings for next week** are also posted on the website.

References

- Hanck, C. et al. (2020). "Econometrics with R". *Regression with a Binary Dependent Variable*
- James, G. et al. (2017). "Introduction to Statistical Learning with Applications in R". *Chapter 4.3*
- Grace-Martin, K. (2018). “Why logistic regression for binary responses?"
- Bellemare, M. (2013) “A Rant on Estimation with Binary Dependent Variables (Technical)"