

# STA 235H - Binary Outcomes

Fall 2021

McCombs School of Business, UT Austin

# Last Week

- Discussed **Regression Discontinuity Designs**:
  - Strong internal validity vs. limited external validity
  - Assumptions behind RD designs
  - Robustness checks
- Finished with **causal inference** chapter.



# Today



- Talking about **models with binary outcomes**:
  - Linear probability models vs. logistic regressions
- Start our **prediction chapter**:
  - Bias vs. Variance trade-off, importance of cross-validation, model selection.

# Binary Outcomes

- We have been using **binary outcomes** in regressions, but haven't fully discussed the issues they might bring.

**What can we do about them?**



# How to handle binary outcomes?

Linear Probability Model

Logistic Regression

# How to interpret a LPM?

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

$$\begin{aligned} E[Y|X_1, \dots, X_p] &= Pr(Y = 0|X_1, \dots, X_p) \cdot 0 + Pr(Y = 1|X_1, \dots, X_p) \cdot 1 \\ &= Pr(Y = 1|X_1, \dots, X_p) \end{aligned}$$

# How to interpret a LPM?

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

$$\begin{aligned} E[Y|X_1, \dots, X_p] &= Pr(Y = 0|X_1, \dots, X_p) \cdot 0 + Pr(Y = 1|X_1, \dots, X_p) \cdot 1 \\ &= Pr(Y = 1|X_1, \dots, X_p) \end{aligned}$$

- Example:

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

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

# 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 condominium
## 1    no 0.221 0.221 0.80000000      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
```

# 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 → **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

```
library(estimatr)

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

	Model 1	Model 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)
R2	0.076	0.076
R2 Adj.	0.075	0.075
se_type		HC2
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

**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.

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

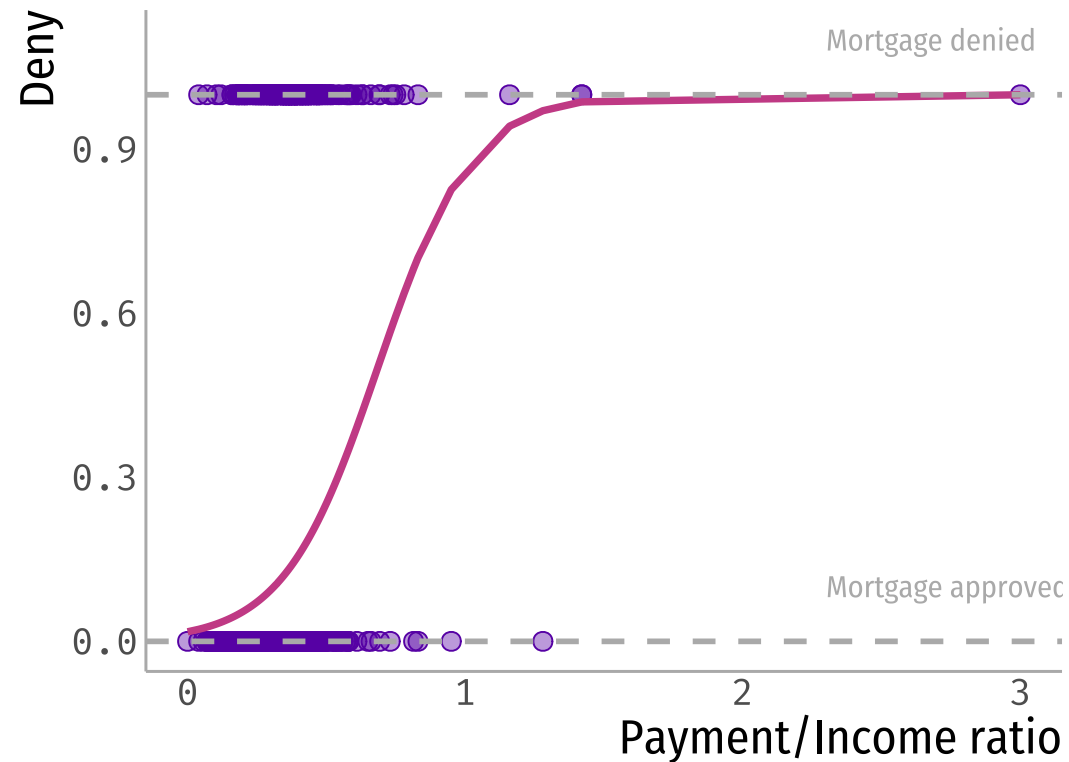
Where in a **logistic regression**:  $F(x) = \frac{1}{1+\exp(-x)}$

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



# How does this look in a plot?

```
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!

- An **odd** is the probability of success over probability of failure:  $\frac{p}{1-p}$

# How to interpret the coefficients? (cont.)

No easy way!

- An **odd** is the probability of success over probability of failure:  $\frac{p}{1-p}$ 
  - e.g. "Your odds of getting into grad school are 2:1" (meaning, your probability of getting in is twice as much as your probability of not getting in)

# How to interpret the coefficients? (cont.)

## No easy way!

- An **odd** is the probability of success over probability of failure:  $\frac{p}{1-p}$ 
  - e.g. "Your odds of getting into grad school are 2:1" (meaning, your probability of getting in is twice as much as your probability of not getting in)
- An **odds ratio** is the odds for scenario 1 over the odds for scenario 2:  $\frac{p_1}{1-p_1} \cdot \frac{1-p_2}{p_2}$ 
  - e.g. "Your odds of getting into grad school if you are male are 1.5 times higher than if you are a female"

# How to interpret the coefficients? (cont.)

No easy way!

- Coefficients in the output are **log odds ratio**:

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

- $(\exp(\beta_1) - 1) \cdot 100\%$  is the expected average increase in the odds of  $Y = 1$  for a one unit increase of  $X_1$ , holding other variables constant.
- The odds of  $Y = 1$  is  $\exp(\beta_1)$  times higher/lower for a one unit increase of  $X_1$ , holding other variables constant.

# How to interpret the coefficients? (cont.)

- Let's go back to our example:

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

```
##  
## Call:  glm(formula = deny ~ pirat + factor(afam), family = binomial(link = "logit"),  
##       data = hmda)  
##  
## Coefficients:  
##      (Intercept)          pirat  factor(afam)yes  
##          -4.126           5.370           1.273  
##  
## Degrees of Freedom: 2379 Total (i.e. Null);  2377 Residual  
## Null Deviance:      1744  
## Residual Deviance: 1591    AIC: 1597
```

- $(\exp(1.27) - 1) \cdot 100\% = 256\% \rightarrow$  Your odds of being denied a mortgage loan are 257% greater if you are African American vs not African American, holding payments to income ratio constant.
- $(\exp(1.27)) = 3.56 \rightarrow$  Your odds of being denied a mortgage loan are 3.6 times greater if you are African American vs not African American, holding payments to income ratio constant.

# How to interpret the coefficients? (cont.)

- Let's look at **probabilities**
- 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 = hmدا)

# Calculate the mean of `pirat` in the HDMA dataset
mean_pirat <- hmدا %>% select(pirat) %>% summarize_all(mean) %>% pull()

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

predictions_afam
```

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



# How to interpret the coefficients? (cont.)

- Let's look at **probabilities**
- 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)

mean_pirat <- hmda %>% select(pirat) %>% summarize_all(mean) %>% pull()

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

predictions_afam
```

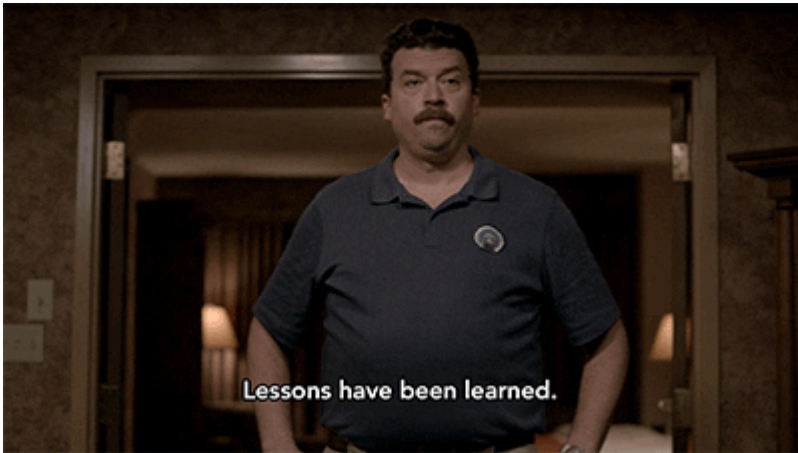
```
##           1           2
## 0.08714775 0.25422824
```

```
diff(predictions_afam)
```

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

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

# Main takeaway points



- LPM and Logistic Regression can **both be useful** depending on the context.
  - LPM for explanation (causal inference) and Logistic Regression for prediction.
- **Be careful** with the interpretation!

# 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)"