

# STA 235H - Multiple Regression: Outliers, Multicollinearity, and Heteroskedasticity

Fall 2022

McCombs School of Business, UT Austin

**Why should we inspect our data before doing anything else?**

# Identifying outliers

- How do we **identify outliers**?
  - Visual inspection (e.g. plots, tables)
  - Creating thresholds (e.g. z-scores, IQ)
- There is **no definite way to identify outliers**
  - Like the characterization of pornography, "I know it when I see it" (P. Stewart, 1964)

**Let's go to R**

# What to do with outliers?

## 1. Check them!

- Make sure there's no coding error; try to understand what's happening there.

## 2a. If they are wrongly coded:

- You can remove them, always adding a note of why you did so
- Be aware of sample selection!

## 2b. If they are correctly coded:

- Run analysis both with and without outliers (don't just drop them!).
- Robust results: Do not depend exclusively on a few observations.

What about multicollinearity?

# What's the problem with multicollinearity?

- What do you think would happen if I ran this regression?

$$Cancer = \beta_0 + \beta_1 PackCigarettes + \beta_2 NCigarettes + \varepsilon$$

- If two covariates are **perfectly collinear**, we can't run the regression
  - R will drop one of them!
- But what happens when they are highly correlated but not *perfectly collinear*?

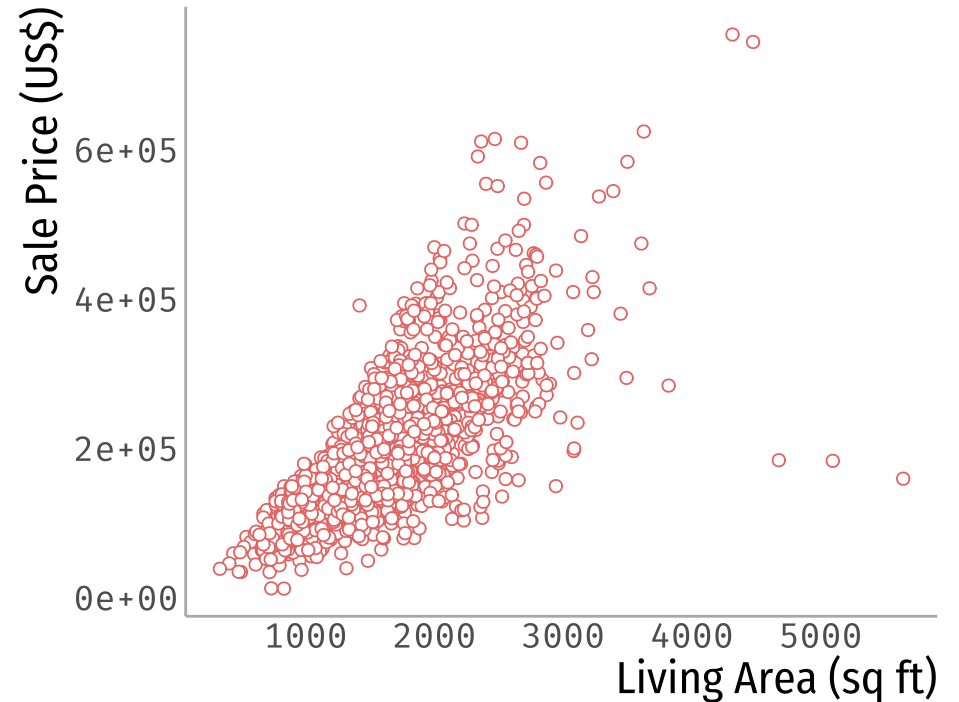
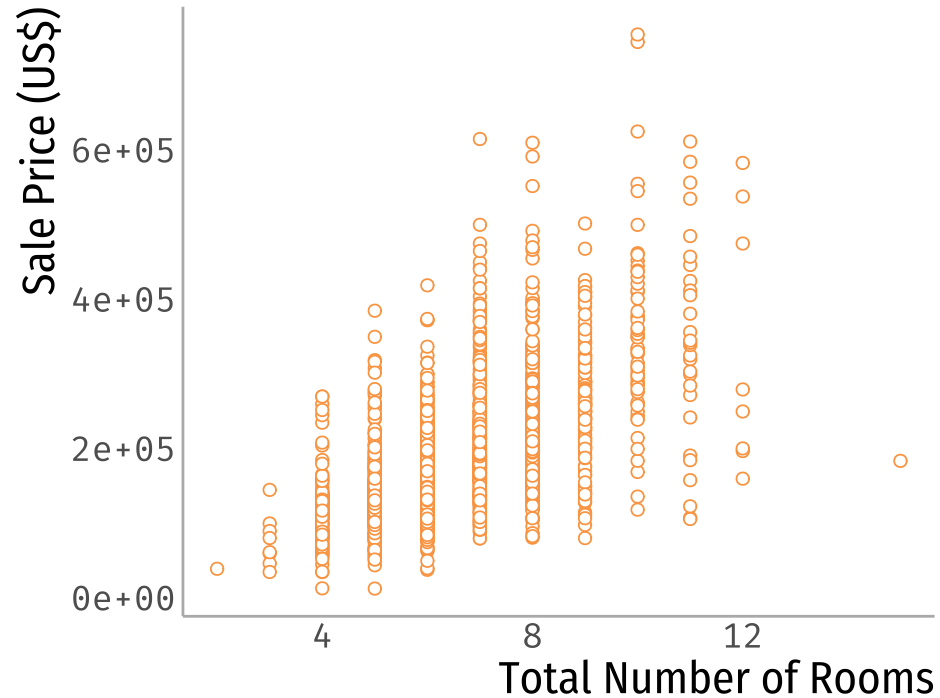
# Let's look at some housing data

- Housing data from Ames, Iowa:

```
library(tidyverse)
housing <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/Week3/2_OLS_Issu
housing <- housing %>% filter(Bldg.Type=="1Fam")
```



# How do number of rooms and living area relate to sale price?



# Let's run a model

```
lm_full <- lm(SalePrice ~ TotRms.AbvGrd + Gr.Liv.Area + Garage.Cars + Fireplaces + Total.Bsmt.SF, data = housing)
summary(lm_full)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	-41874.0627	4464.9400	-9.3784	0.0000
## TotRms.AbvGrd	1314.4670	1079.3420	1.2178	0.2234
## Gr.Liv.Area	62.9742	3.4525	18.2403	0.0000
## Garage.Cars	33846.9897	1452.9358	23.2956	0.0000
## Fireplaces	8029.2652	1558.3305	5.1525	0.0000
## Total.Bsmt.SF	53.8685	2.4248	22.2154	0.0000

- What can you say about **total rooms**?

# Multicollinearity in Regressors

- If two covariates are **highly correlated**, it is difficult to separate the contribution of each!
  - E.g. They move together
- **Be careful** with interpretations
  - The same with extrapolation zones!

# Can we do something about it?

- **Exclude** one variable
  - Depends on what you are analyzing.
- Create a **new variable** that captures clean information
  - E.g. Residuals
- **Aggregate** both variables into one:
  - Linear combination (e.g. add them)
  - Create an index

What about binary responses?

# 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

# How to interpret a LPM?

- A Linear Probability Model is just a **traditional regression with a binary outcome**
- $\hat{\beta}$ '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:

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

- $\hat{\beta}_1$  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**.

# 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
```

- Holding payment-to-income ratio constant, an AA client has a probability of getting their loan denied that is **18 pp higher**, on average, than a non AA client.
- Being AA is associated to an average increase of **0.177 in the probability** of getting a loan denied compared to a non AA, 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 → **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 `estimat` in R is great!

# 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)
Std.Errors	HC2	
+ 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.

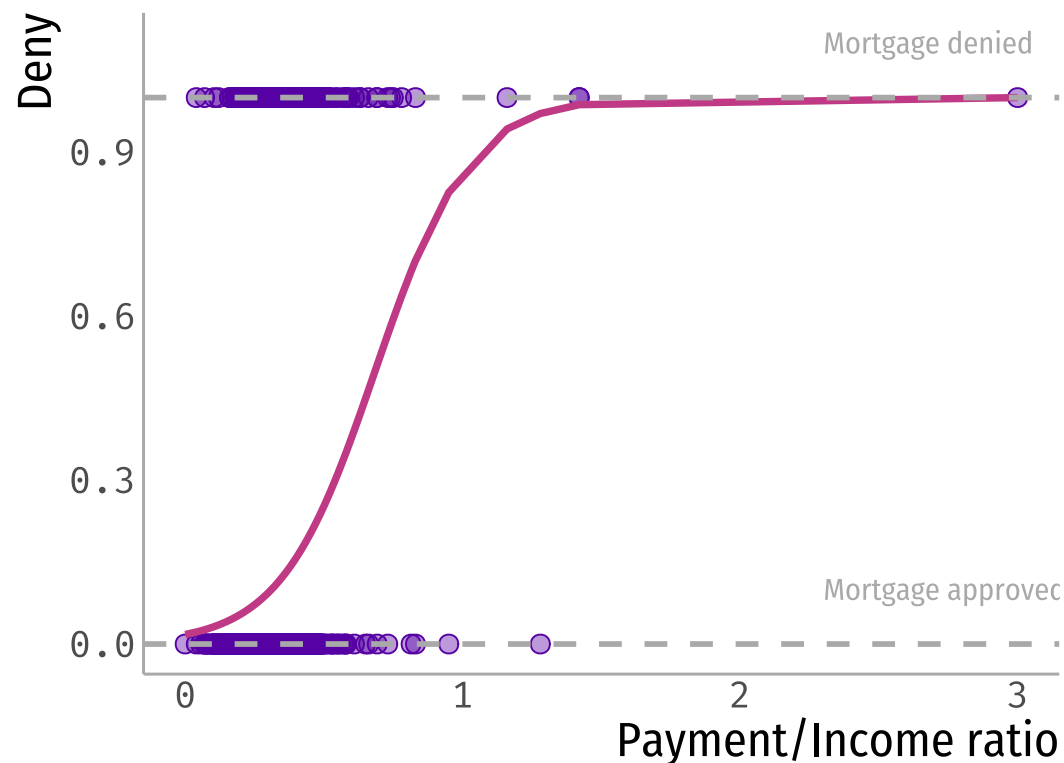
$$Pr(Y = 1|X_1, \dots, X_p) = F(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)$$

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
```



# 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?
- Is it a problem if my covariates are **highly correlated**?
  - Sometimes yes, sometimes no.
- For LPM, **always include robust standar errors**!



# References

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