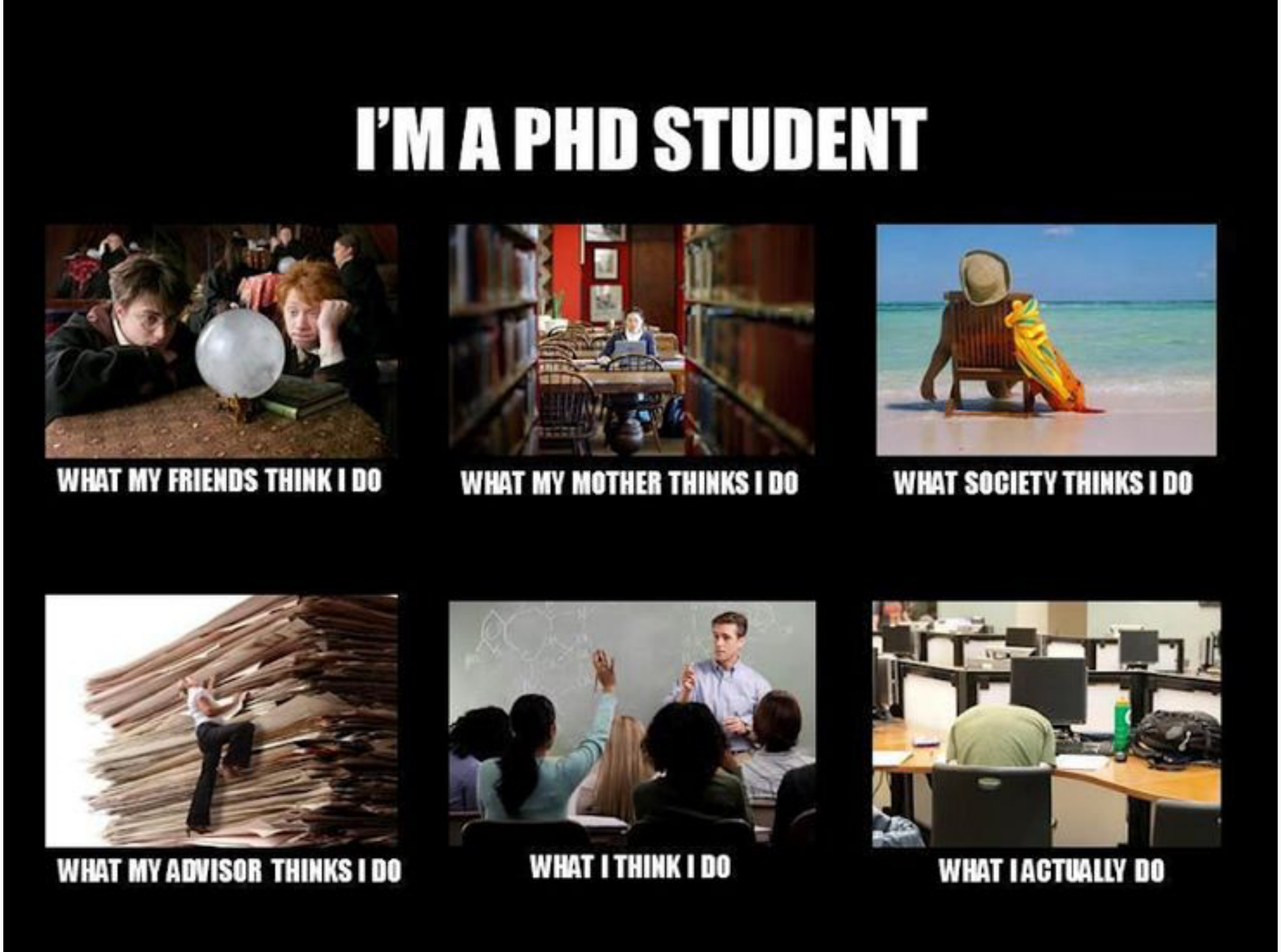
Assumptions...



EDUC 7610

Chapter 16 and 17

Diagnostics (Detecting Irregularities)

and miscellaneous stuff

Tyson S. Barrett, PhD

This is one of the most important chapters

The regression results' validity depend on whether the *assumptions* of the model hold or not

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals with mean 0*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

HOLDS?

VIOLATED?

This is one of the most important chapters

The regression results validity depend on whether the *assumptions* of the model hold or not

Violations usually occur because of ***Extreme Cases***

Leverage

The atypicalness of a case's pattern of values on the regressors in the model

A point with high leverage could be:

- ***A 55-year-old pregnant female***

  In a general population, is being 55 strange by itself? What about pregnant?

- ***A high-income individual receiving welfare assistance***

  In a general population, is having a high-income strange by itself? What about being on welfare assistance?

We must consider the combination of the variables to know if it has high leverage

Measured with h_i ⟶ "case *i*'s **hat value**"

Distance

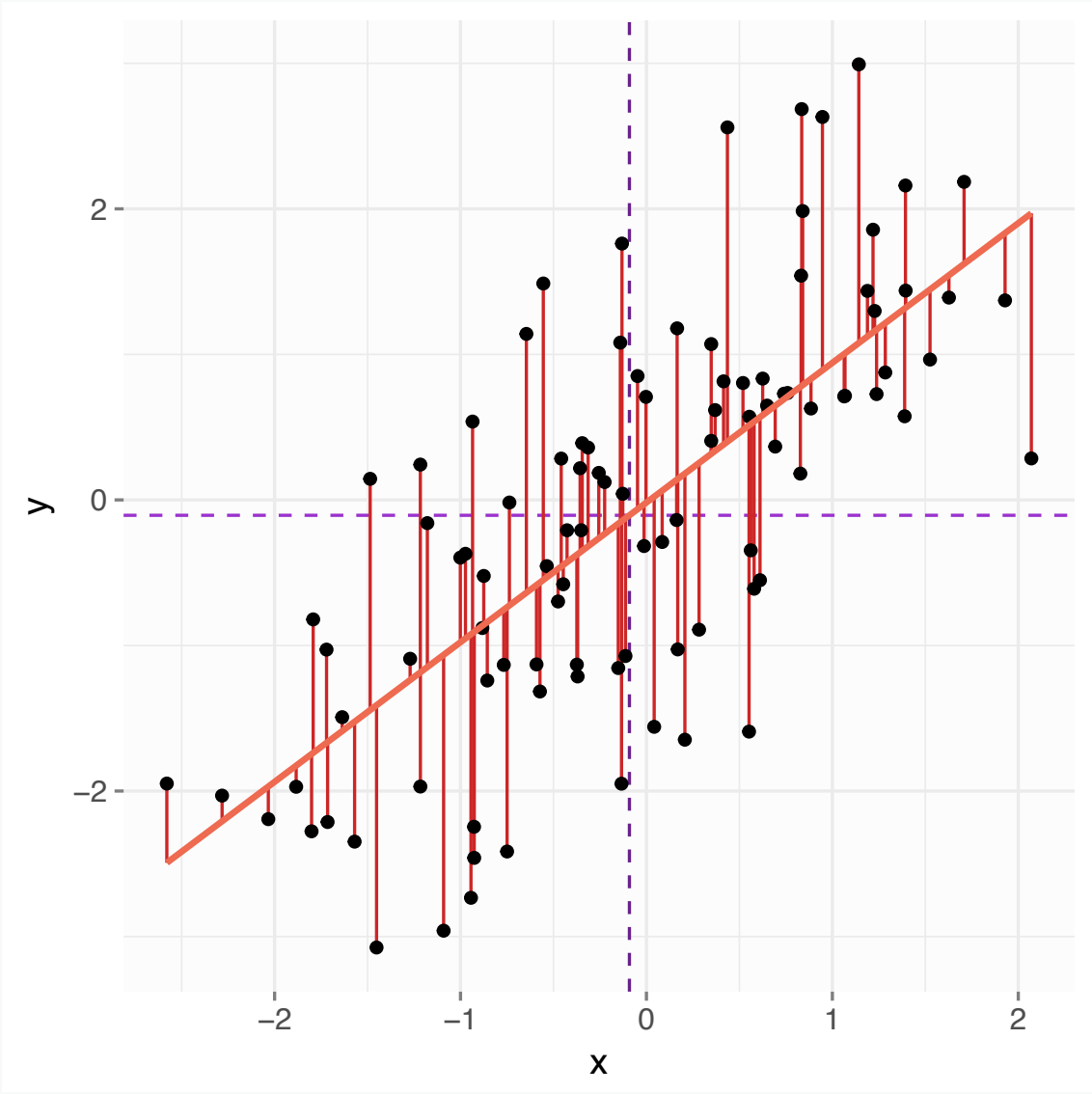
How far case *i*'s Y_i value deviates from $\hat{Y}_i$

Often measured with residual: $e_i = Y_i - \hat{Y}_i$

But an outlier pulls the $\hat{Y}_i$ so we can adjust it using h_i

$$\frac{e_i}{\sqrt{(1 - h_i)}}$$

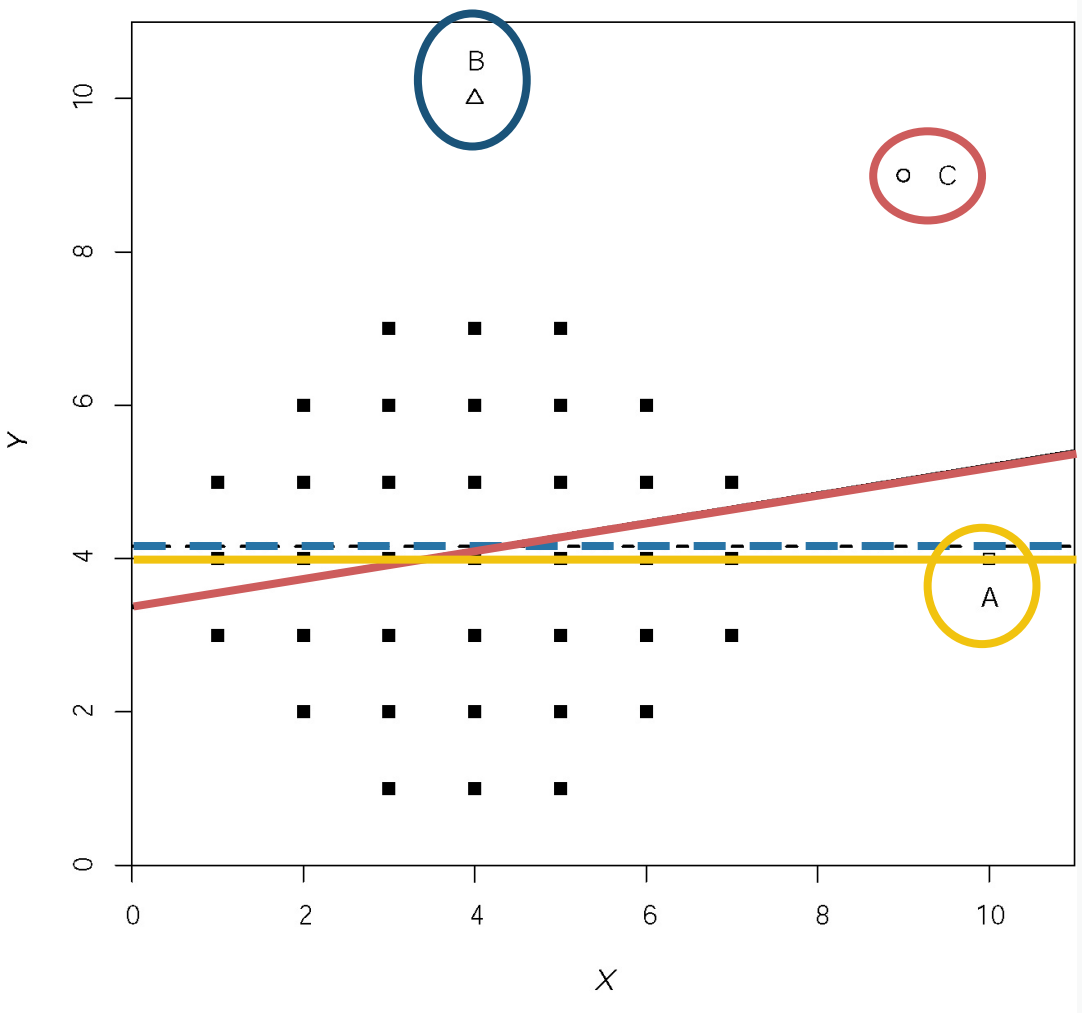
Turns out with mathemagic, we can see that h_i is equal to proportion by which case I lowers its own residual by pulling the regression surface



Influence

The extent to which its inclusion changes the regression solution or some aspect of it

Which extreme point (A, B, or C) changes the solution the most?



Influence

The extent to which its inclusion changes the regression solution or some aspect of it

Which extreme point (A, B, or C) changes the solution the most?

total influence
vs.
partial influence

Measured with *Cook's Distance*

$$Cook_i = \frac{\sum_{j=1}^{N} d_{ij}^2}{k \times MS_{residual}}$$

the change in the value of case *j*'s residual when case *i* is deleted from the model

number of regressors

The error variance from the model

Approaching Diagnostics

Diagnostic statistics are estimates of different features of the observations

i	X_1	X_2	X_3	Y
1	0	1	3	8
2	0	3	4	13
3	0	11	5	17
4	0	7	8	7
5	0	9	8	10
6	0	12	12	10
7	1	2	5	10
8	1	4	6	15
9	1	3	4	11
10	1	7	12	6
11	1	9	10	11
12	1	13	14	9

Anything look weird in this data? Any extreme values?

Eye-balling a data set (especially larger ones) is not very productive

Approaching Diagnostics

Diagnostic statistics are estimates of different features of the observations

regular residual

i	X_1	X_2	X_3	Y	$\hat{Y}$	e	$_{d}e$
1	0	1	3	8	9.523	−1.523	−2.497
2	0	3	4	13	10.204	2.796	3.800
3	0	11	5	17	16.994	0.006	0.022
4	0	7	8	7	8.857	−1.857	−2.377
5	0	9	8	10	10.893	−0.893	−1.096
6	0	12	12	10	8.528	1.472	2.349
7	1	2	5	10	10.664	−0.664	−0.911
8	1	4	6	15	11.345	3.655	4.637
9	1	3	4	11	13.037	−2.037	−3.048
10	1	7	12	6	6.270	−0.270	−0.444
11	1	9	10	11	11.016	−0.016	−0.020
12	1	13	14	9	9.669	−0.669	−1.149

residual with this case removed from Y_i

Instead, let's use the diagnostics

Which has a high distance?

Approaching Diagnostics

Diagnostic statistics are estimates of different features of the observations

Measures leverage

i	X_1	X_2	X_3	Y	$\hat{Y}$	e	${}_de$	MD	h
1	0	1	3	8	9.523	−1.523	−2.497	3.374	0.390
2	0	3	4	13	10.204	2.796	3.800	1.991	0.264
3	0	11	5	17	16.994	0.006	0.022	7.157	0.734
4	0	7	8	7	8.857	−1.857	−2.377	1.488	0.219
5	0	9	8	10	10.893	−0.893	−1.096	1.117	0.185
6	0	12	12	10	8.528	1.472	2.349	3.192	0.374
7	1	2	5	10	10.664	−0.664	−0.911	2.069	0.271
8	1	4	6	15	11.345	3.655	4.637	1.413	0.212
9	1	3	4	11	13.037	−2.037	−3.048	2.733	0.332
10	1	7	12	6	6.270	−0.270	−0.444	3.389	0.391
11	1	9	10	11	11.016	−0.016	−0.020	1.395	0.210
12	1	13	14	9	9.669	−0.669	−1.149	3.681	0.418

Which has a high leverage?

Measures leverage and ranges from 1/N to 1

Approaching Diagnostics

Diagnostic statistics are estimates of different features of the observations

Measures influence

i	X_1	X_2	X_3	Y	$\hat{Y}$	e	$_de$	MD	h	str	tr	$Cook$
1	0	1	3	8	9.523	−1.523	−2.497	3.374	0.390	−0.932	−0.923	0.139
2	0	3	4	13	10.204	2.796	3.800	1.991	0.264	1.558	1.746	0.218
3	0	11	5	17	16.994	0.006	0.022	7.157	0.734	0.005	0.005	0.000
4	0	7	8	7	8.857	−1.857	−2.377	1.488	0.219	−1.004	−1.005	0.071
5	0	9	8	10	10.893	−0.893	−1.096	1.117	0.185	−0.473	−0.449	0.013
6	0	12	12	10	8.528	1.472	2.349	3.192	0.374	0.889	0.876	0.118
7	1	2	5	10	10.664	−0.664	−0.911	2.069	0.271	−0.372	−0.351	0.013
8	1	4	6	15	11.345	3.655	4.637	1.413	0.212	1.968	2.563	0.260
9	1	3	4	11	13.037	−2.037	−3.048	2.733	0.332	−1.191	−1.228	0.176
10	1	7	12	6	6.270	−0.270	−0.444	3.389	0.391	−0.165	−0.155	0.004
11	1	9	10	11	11.016	−0.016	−0.020	1.395	0.210	−0.009	−0.008	0.000
12	1	13	14	9	9.669	−0.669	−1.149	3.681	0.418	−0.419	−0.396	0.032

Which has a high influence?

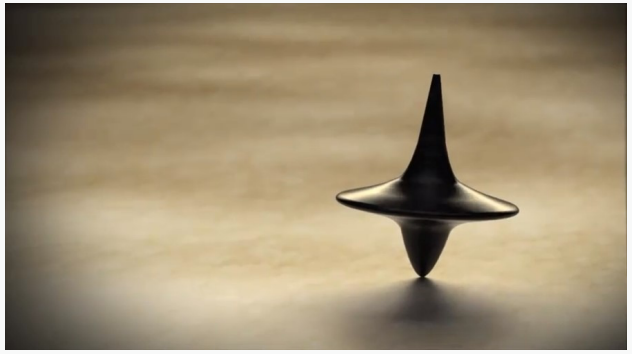
Assumptions

The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals with mean 0*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

Note: there are a lot of “tests” for these assumptions but they usually have their own assumptions so we won’t discuss them here



Assumptions

The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals with mean 0*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

Omitted variables is largely theoretically based
Can show up as weird residuals (maybe we are missing an effect)
We'll see this in the next slide

This is easy to test: are there at least two values in X?

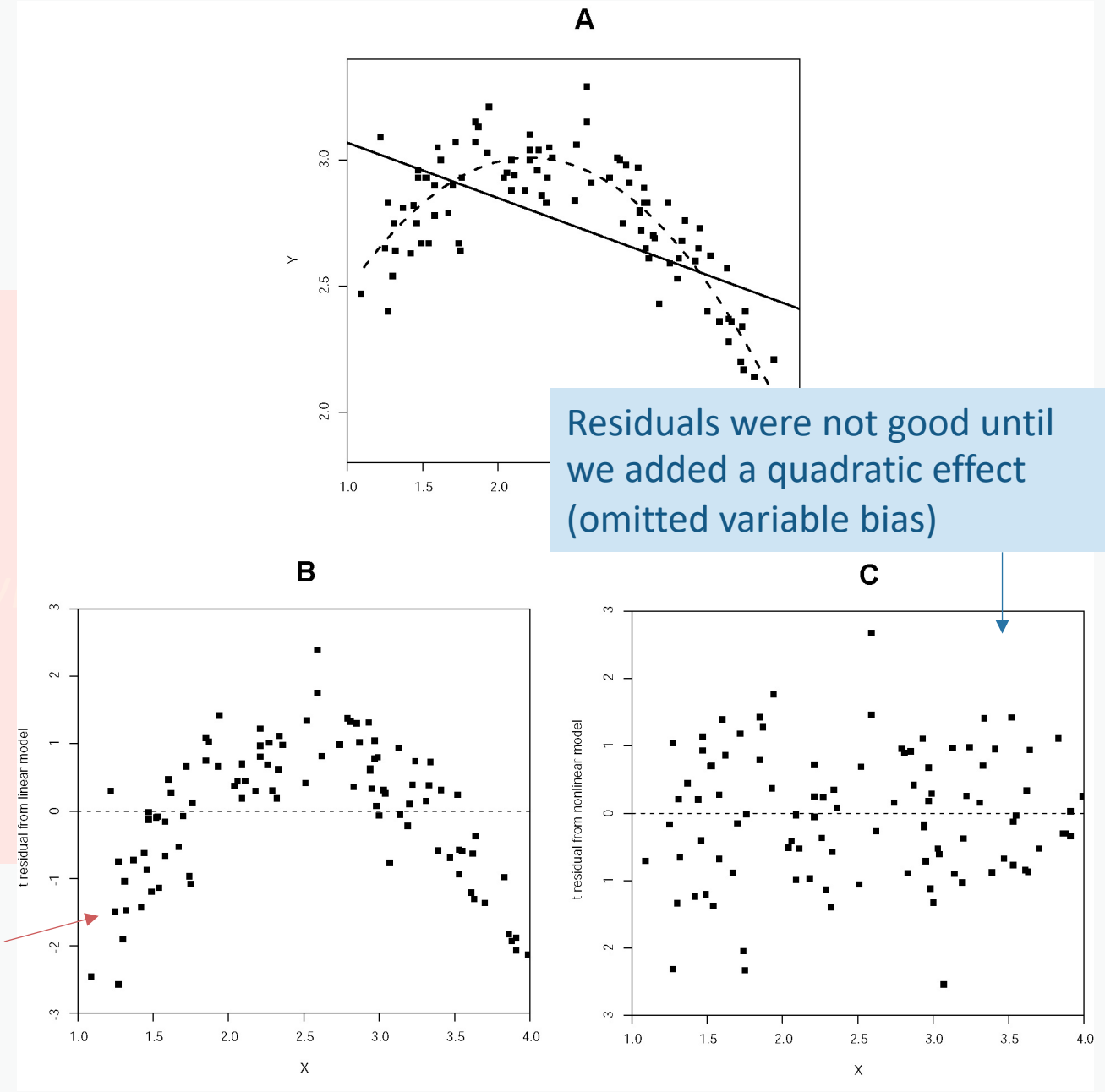
Assumptions

The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals w*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

Use a scatterplot of t-residuals against X



Residuals were not good until we added a quadratic effect (omitted variable bias)

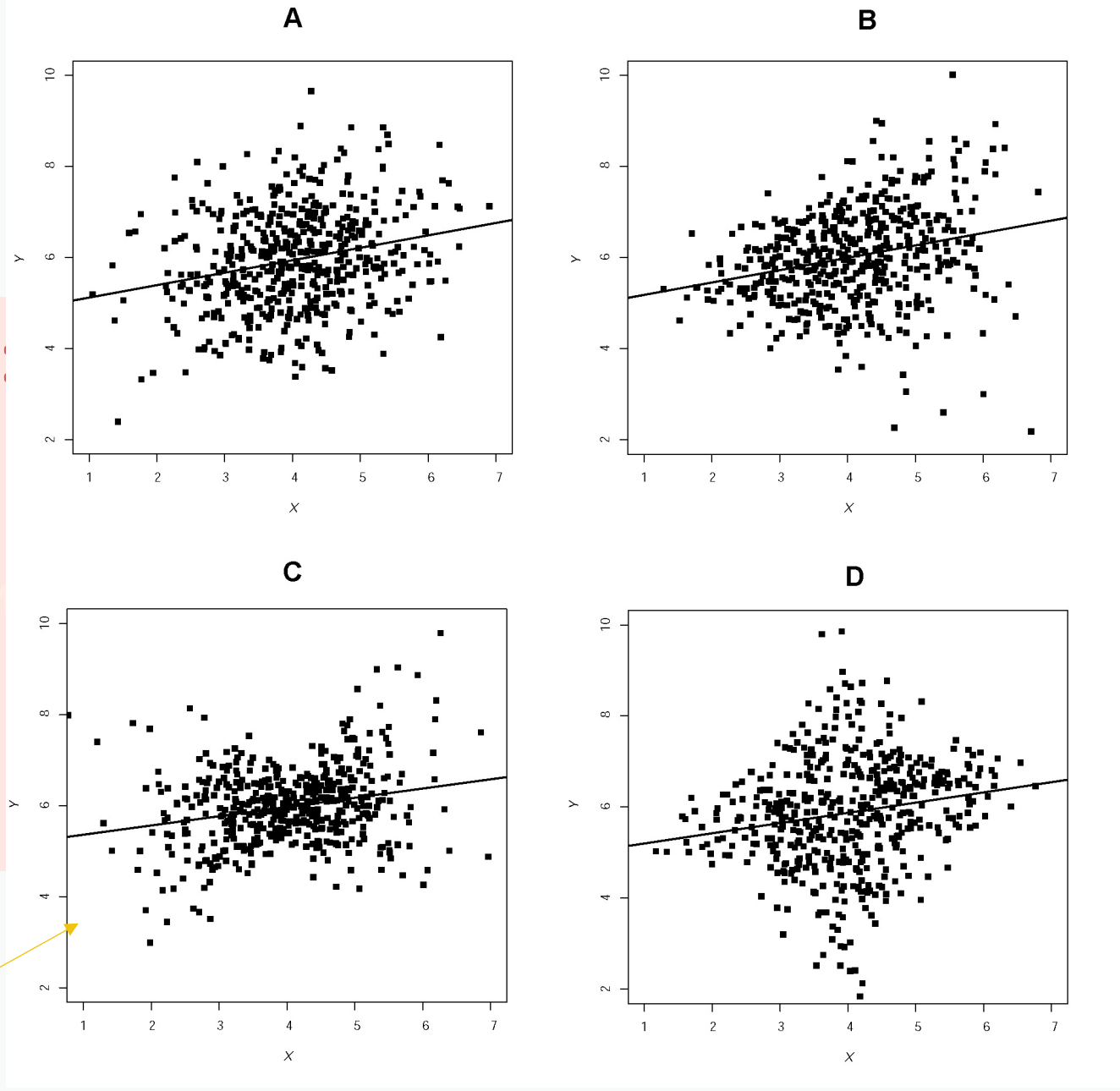
Assumptions

The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals w*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

Scatterplots of residuals on X or Y on X



Assumptions

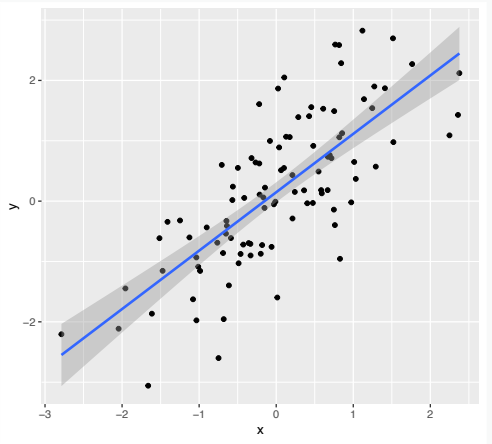
The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals with mean 0*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

This one is somewhat tricky but important:

- The residuals at each point of x are normally distributed is the assumption
- Are more points closer to the line than far away?



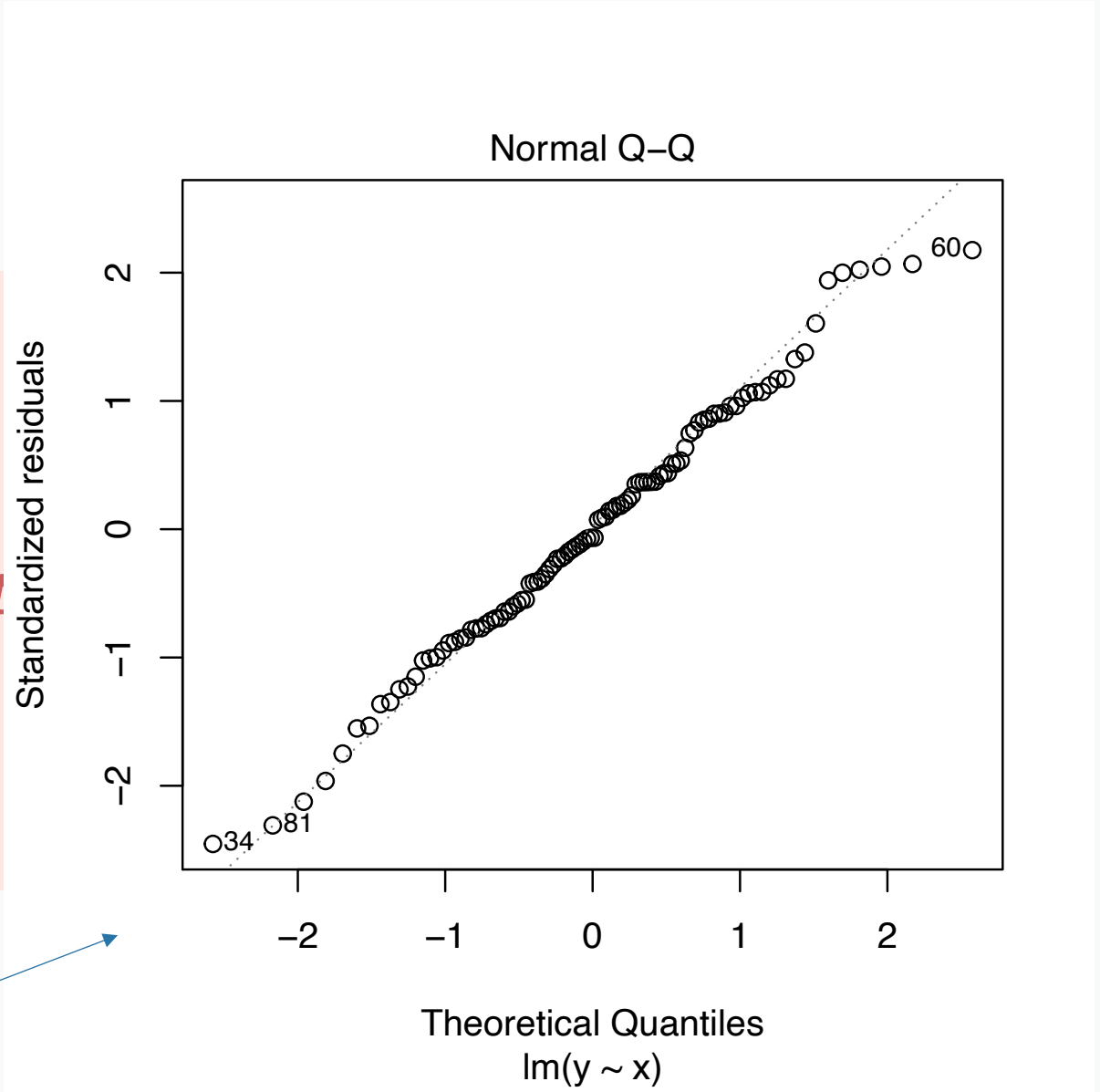
Assumptions

The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals wit*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

Usually tested with a Q-Q plot



Assumptions

The book provides four basic assumptions of regression, we make explicit two implicit ones below

Assumptions of the model:

1. *Linear relationship*
2. *Homoscedasticity of residuals*
3. *Normally-distributed residuals with mean 0*
4. *No omitted variables*
5. *Independence of residuals*
6. *Variance of X > 0*

Generally theoretical in nature

- Are the observations (participants) independent?
- Are the observations connected in some way?
- Time-series almost always violate this assumption because previous time points are correlated with current ones
- In many cases, we can use *Multilevel Modeling* here

Dealing with Irregularities

The book provides four basic ways to deal with irregularities, we add a fifth

Correction

Correct the error that lead to the extreme value

Transformation

Transform the outcome or predictors using a monotonic transformation (log, square root, etc.)

Elimination

Remove the extreme value

- *I recommend if you do this to report the results both with and without the extreme value in any publication*

Robustification

Use an alternative approach that is less sensitive to the extreme value

Generalized Linear Models

A family of approaches that can assess categorical, ordinal, and otherwise strange outcomes

- *Chapter 18 is about these and we'll discuss them much more then*

Robustification

Use an alternative approach that is less sensitive to the extreme value

Two main ways of robustifying

1. Use alternative way of estimating coefficients
2. Use alternative way of estimating the standard error (or, more generally speaking, the uncertainty)

We'll talk about this one

Robustification

Use an alternative approach that is less sensitive to the extreme value

Heteroscedasticity-Consistent Standard Errors

Adjusts the SE's to be less sensitive to extreme values using *sandwich estimators*

They are *consistent* (gets closer and closer to the right value as sample size increases)

Many versions, HC3 and HC4 are best

Bootstrapping

Resamples from the sample with replacement to come up with an empirical distribution of the estimate

Can obtain SE's, CI's, and p-values

Works with any statistic

Permutation

Randomly shuffles the data and re-runs the model many times

The proportion of these shuffled models that are as big or bigger than the original model gives us info on p-values

Works with any statistic

Robustification

Use an alternative approach that is less sensitive to the extreme value

Heteroscedasticity-Consistent Standard Errors

Adjusts the SE's to be less sen
san

The
and
sam

Many versions, HC3 and HC4 are best

Bootstrapping

Resamples from the sample with

Permutation

Randomly shuffles the data and

ed
er
us

Works with any statistic

Can use whether or not homoscedasticity exists

Remember:

No model is “correct” but some models are useful

Some Miscellaneous Stuff

Measurement Error

Power

Specification Error

Non-interval Outcomes

Missing Data

Some Miscellaneous Stuff

Measurement Error

A weakness of regression but not something to worry about excessively

Non-int
Outco

Reliability

The proportion of a variable's variability that is attributable to variability in the true scores

All measures have less than perfect reliability

Some Miscellaneous Stuff

Measurement Error

A weakness of regression but not something to worry about excessively

So should we leave a variable out if it has poor reliability?

Random Measurement Error

Measurement Error in Y

Increases SE

No bias in coefficients (R is tho)

Measurement Error in X

Usually attenuates coefficients

Increases SE

Some Miscellaneous Stuff

Measurement Error

Power

Specification Error

The probability of obtaining a statistically significant effect if in fact an effect actually exists

Anything that affects the SE affects the power

$$SE(b_j) = \sqrt{\frac{MS_{residual}}{N \times Var(X_j) \times Tol_j}}$$

Some Miscellaneous Stuff

The most difficult aspect of regression may be specifying it correctly

Specification Error

Many issues discussed in this chapter can result from model miss-specification (e.g., leaving out a quadratic effect)

Undercontrol vs. Overcontrol

The Fork

Open unless you condition on Z

The Pipe

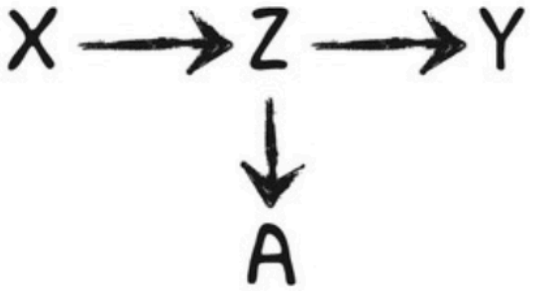
Open unless you condition on Z

The Collider

Closed until you condition on Z

Overcontrol

The Descendant



Conditioning on A is like conditioning on Z

Some Miscellaneous Stuff

How should we handle this?

- Treat it as continuous?
- Use a transformation?
- Depends on distribution, but likely a GLM is better (See Chapter 18)

Non-interval Outcomes

Likert scales and similar outcomes are common

Some Miscellaneous Stuff

Three main ways to handle missing data:

1. Pairwise Deletion
2. Listwise Deletion
3. Imputation

Multiple Imputation is one of the best

A cousin (Full Information Maximum Likelihood) is also a top choice

Non-interval Outcomes

Missing Data

Data are often missing throughout the sample that weren't planned for