A Comparative Analysis of Pandas vs. Academics

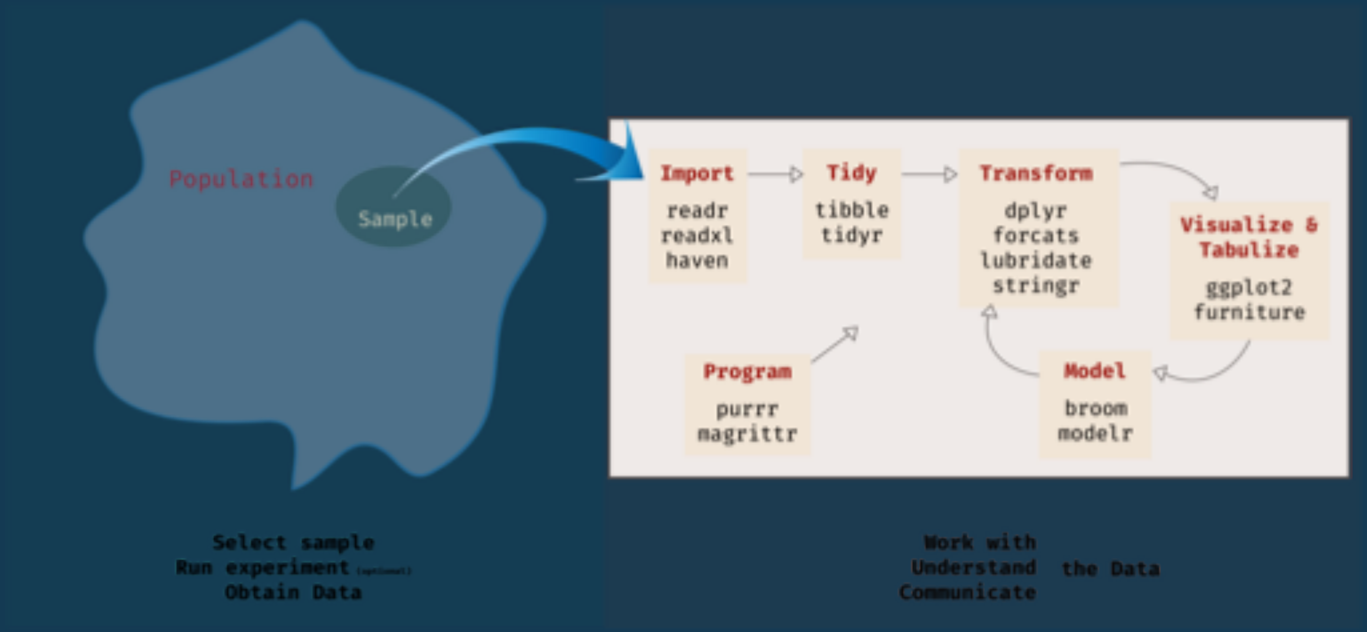
Pandas are a species of animals whose survival depends on conservation efforts by government agencies. Similarly, Academics are a sub-species with debatable value to society. Here now is an analytic comparison of the two endangered animals:

	Pandas	Academics
Moves slowly:	✓	✓
Sleeps during daytime:	✓	✓
Has permanent dark shadows under their eyes:	✓	✓
Generally avoids reproduction:	✓	✓
Society keeps them around because they're:	Very cute	Somewhat astute

EDUC 7610

Chapter 4

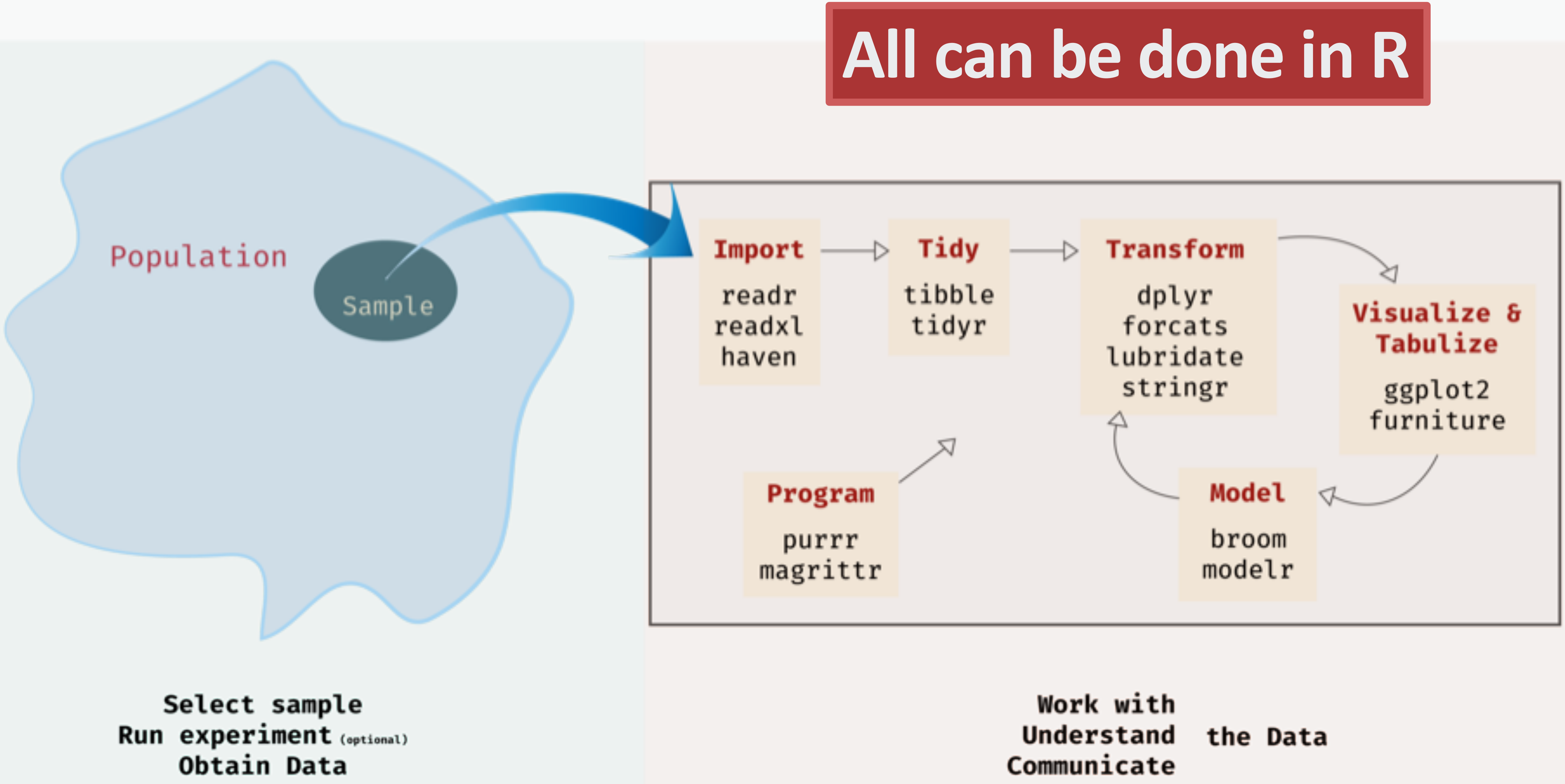
Statistical Inference



Fall 2018

Tyson S. Barrett, PhD

The Whole Idea



Why Statistical Inference?

So far, we've used regression just to describe our sample

But our goal is to understand the population, not just our sample

There is a "true" value out there in the population

- *But we don't have access to it (unless we use a census)*

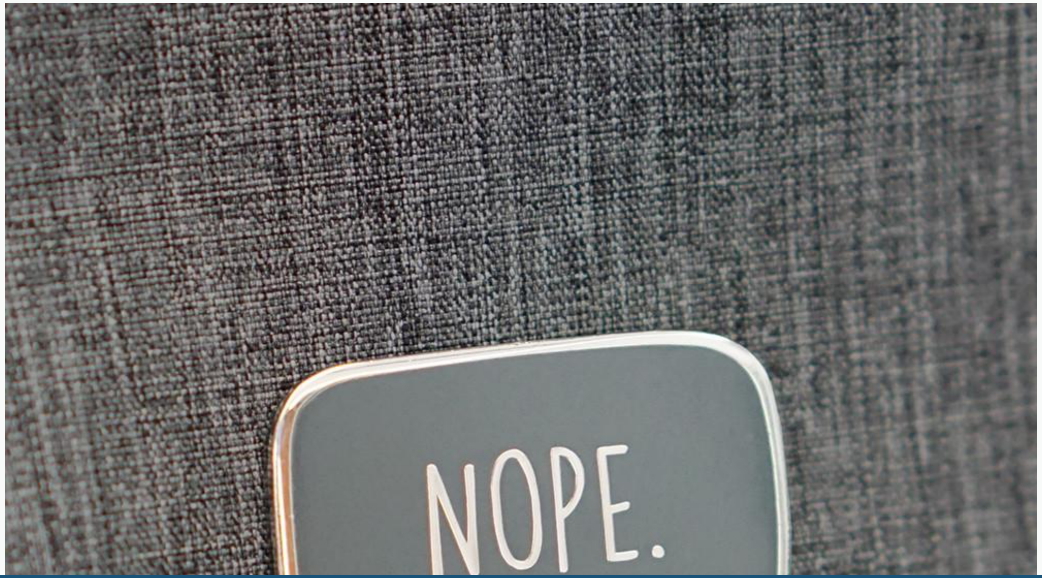
⇒ So we estimate it using our sample

Why Statistical Inference?

Is our sample going to be exactly identical to the population we pulled it from?

Why Statistical Inference?

Is our sample going to be exactly identical to the population we pulled it from?



Sampling Variance (Error)

Causes uncertainty in our estimates

To infer about the population, we need to make some assumptions

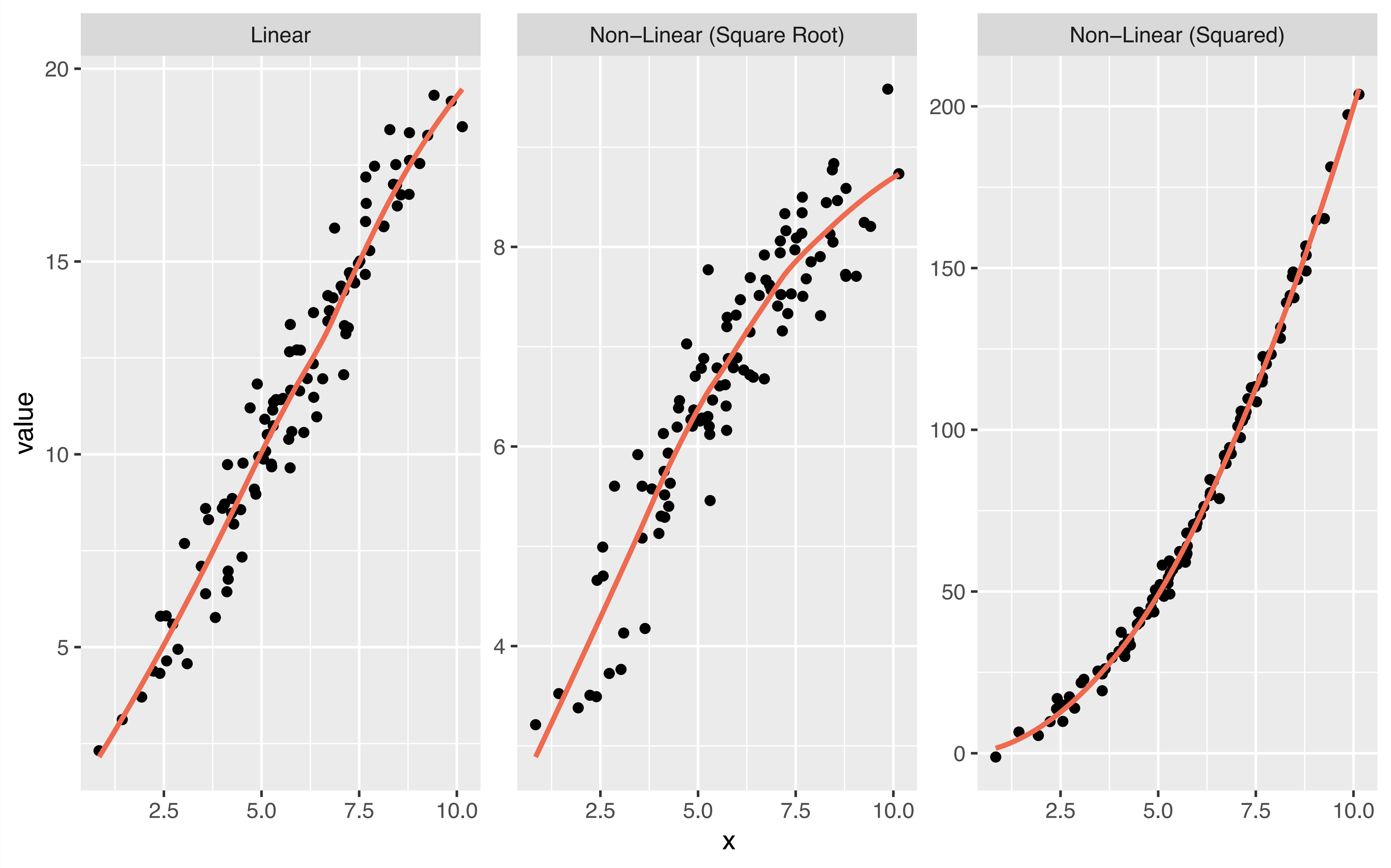
1 **Linearity** – the relationship between outcome and predictors is approx. linear

2 **Homoscedasticity** – the conditional distributions of Y have equal variances

3 **Conditional Distribution of Y** – is normally distributed

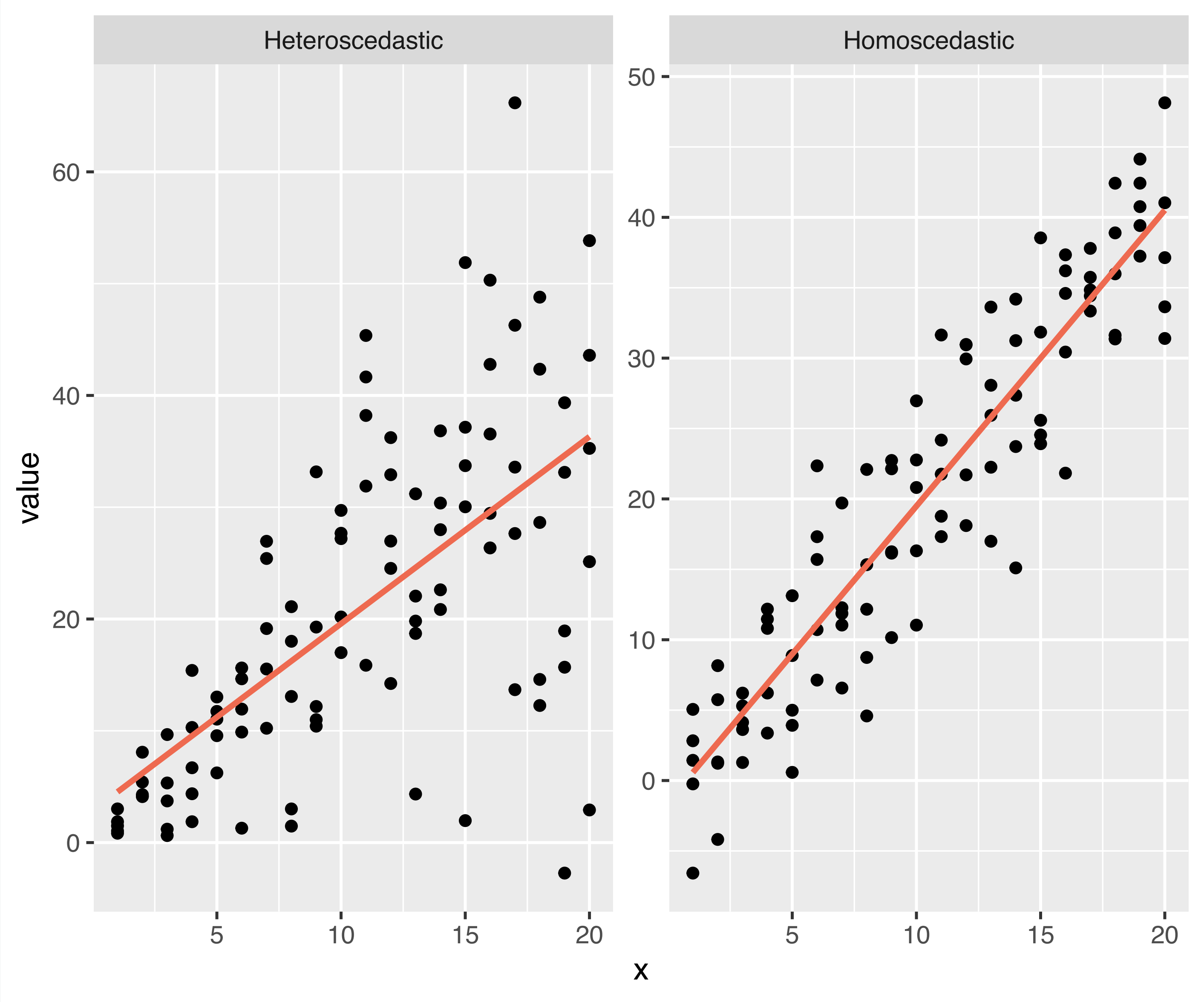
4 **Independent Sampling** – each member of our sample is independent of the other members

Linearity



Linear
Non−Linear (Square Root)
Non−Linear (Squared)
value
x
20
15
10
5
8
6
4
200
150
100
50
0
2.5
5.0
7.5
10.0

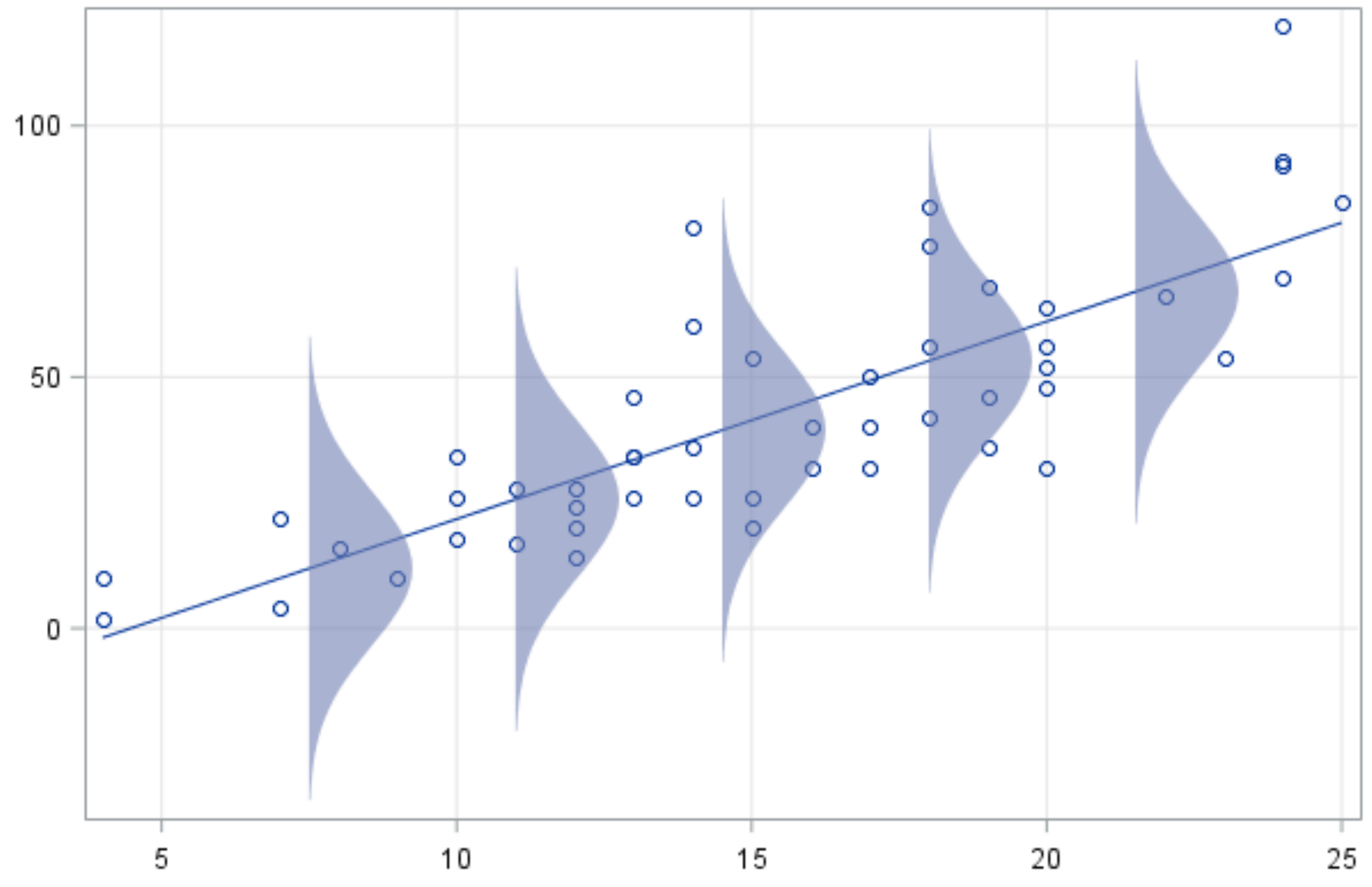
Homoscedasticity



Conditional Distribution of Y

At each point of x, there is an assumed normal distribution around the line

Central Limit Theorem helps us here (samples above 30 don't rely on this assumption as much)



Independent Sampling

Each member of our sample (e.g., person, class, animal) must be independent of the others

- No influence from one member to another

Name some situations where this would be violated

When this is violated, we can use multilevel modeling techniques

What about violations of these assumptions?

1 **Linearity** – if this is violated we can try different specifications (e.g., square or square root of a predictor); otherwise, violating this is disastrous

2 **Homoscedasticity** – can mess with your standard errors; can use special estimators (sandwich estimator, robust SEs)

3 **Conditional Distribution of Y** – often not too bad in larger samples

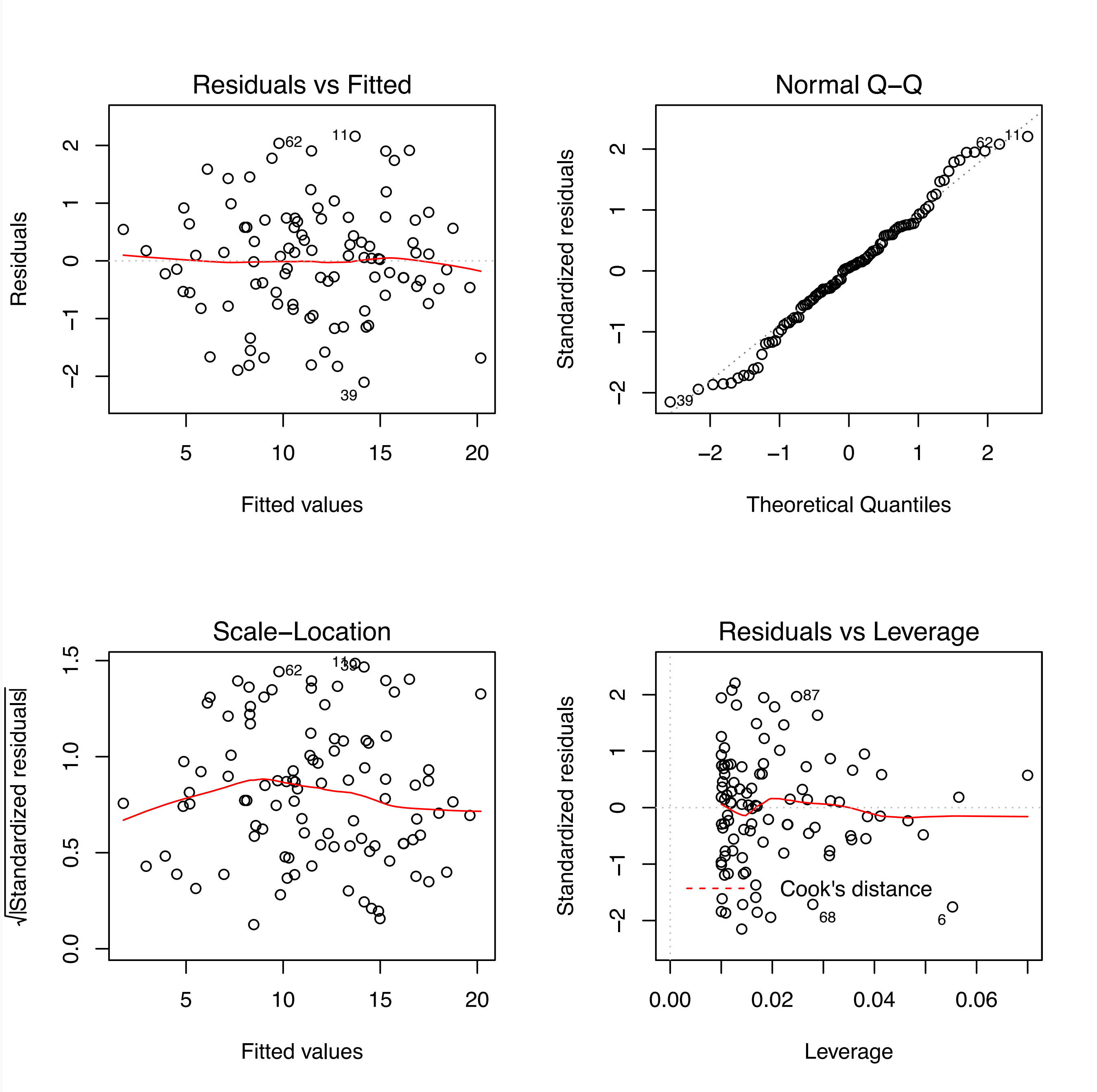
4 **Independent Sampling** – can sometimes really mess up your results (simpson's paradox); use multilevel modeling to fix

Assumptions and Residuals

All of the assumptions can be framed in terms of the residuals

Residuals are normal, homoscedastic, have a mean of zero at all points of x, and are uncorrelated

= i.i.d. (independently and identically distributed)



Quick Aside about Vocab and Notation

Expected Value

If we did something a thousand times, what value do we expect?

$E(Y)$ $E(b_j)$ $E(Y)$

Unbiased Estimation

An estimate that arrives at the expected value

$$E(Y_i) = \frac{1}{n} \sum Y_i$$

Quick Aside about Vocab and Notation

Is the following unbiased?

$$E(Y_i) = \frac{1}{n}\sum Y_i + 1$$

No. If we did this many, many times, on average we'd be off by 1

Regression is an UNBIASED estimator of the population value

We could show this mathematically

Ordinary Least Squares Regression is B.L.U.E.

It is the most precise (the smallest accurate standard errors)

Best

It is unbiased (it estimates the population value)

Linear

It is a linear model

Unbiased

Everything we are doing with regression is an estimate

Estimator

Note: Maximum likelihood regression is very similar

So what does all this mean?

Regression provides us with the "best" linear, accurate way to understand a population using a sample

Regression Results in ANOVA form

Regression results often are lead by an ANOVA table or information from an ANOVA table

Remember that ANOVA is just a special case of regression?

Source	SS	df	MS	F
Regression	$\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$	k	$SS_{regression}/k$	$MS_{regression}/MS_{residual}$
Residual	$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$	$N - k - 1$	$SS_{residual}/(N - k - 1)$	
Total	$\sum_{i=1}^{N}(Y_i - \bar{Y})^2$	$N - 1$		

What do we want to be able to infer?

1. Multiple R (or R^2)
2. Regression Coefficients
3. (Partial) Correlation

Inference: Multiple R

This tests the entire model

- Do the predictor(s) together have a relationship with the outcome?
- *Common to discuss the model as a whole before discussing the individual predictors*

Statistic of Interest	Test Statistic	Significance	Example
R^2 (or adjusted R^2)	F-statistic $F = \frac{MS_{reg}}{MS_{res}}$	P < .05 suggests there is a relationship among the predictor(s) and outcome	The model that included SES explained 30% more of the variance in the outcome and was significantly better (p < .001)

Inference: Multiple R

The Null Hypothesis: Model is no better than comparison model
(either a null model or another "nested" model)

The Alternative: Model is better than comparison

Statistic of Interest	Test Statistic	Significance	Another Example
R^2 (or adjusted R^2)	F-statistic $F = \frac{MS_{reg}}{MS_{res}}$	P < .05 suggests there is a relationship among the predictor(s) and outcome	The model explained 45% of the variation in the outcome and is significantly better than the null model (p = .002).

Inference: Regression Coefficients

This testing each individual predictor

- Do each predictor have a relationship with the outcome?
- *Most common way of interpreting regression*

Statistic of Interest	Test Statistic	Significance	Example
b_j or β_j	T-statistic	$P < .05$ suggests there is a relationship among this predictor and the outcome	Controlling for the covariates, for a one unit increase in SES, there is an associated decrease of b_1 in the outcome ($p = .03$).

Inference: Regression Coefficients

This testing each individual predictor

- Do each predictor have a relationship with the outcome?
- *Most common way of interpreting regression*

We do the same tests for the standardized coefficients as well (just with standardized variables instead of the raw ones)

Inference: Regression Coefficients

This testing each individual predictor

- Do each predictor have a relationship with the outcome?
- *Most common way of interpreting regression*

Statistic of Interest	Test Statistic	Significance	Example
b_j or β_j	T-statistic	P < .05 suggests there is a relationship among this predictor and the outcome	Controlling for the covariates, for a one SD increase in SES, there is an associated decrease of b_1 SDs in the outcome (p = .03).

Inference: Regression Coefficients

Important Pieces of the Coefficients

- *The Estimate*
- *The Standard Error of the Estimate*
  - *Testing the null hypothesis*
  - *Confidence Intervals*

Inference: Regression Coefficients

The Estimate

Simple

$$b_j = \frac{Cov(X,Y)}{Var(X)}$$

Multiple

$$\text{all } b_j s = (X'X)^{-1} X'Y$$

Inference: Regression Coefficients

The Standard Error

$$SE(b_j) = \sqrt{\frac{MS_{residual}}{(N)\,Var(X_j)\,(1 - R_j^2)}}$$

estimate of variance of the residuals

Sample size used in analysis

Variance of that predictor

R_j^2 here is the R^2 from the model with all variables but j

this is called the *tolerance*

Inference: Regression Coefficients

The Standard Error

$$SE(b_j) = \sqrt{\frac{MS_{residual}}{(N)\,Var(X_j)\,(1 - R_j^2)}}$$

What increases the SE?

$\uparrow MS_{residual}$

$\downarrow N$

$\downarrow Var(X_j)$

$\downarrow (1 - R_j^2)$

Inference: Regression Coefficients

$$(1 - R_j^2)$$ = The Tolerance of X_j

A measure of the *independence* of X_j from the other predictors (i.e., measures the *collinearity*)

- When Tol = 0, there is perfect collinearity
- When 1 > Tol > 0, there is some correlation between predictors
- When Tol = 1, there is no correlation at all between predictors

Inference: Regression Coefficients

$$(1 - R_j^2)$$ = The Tolerance of X_j

A measure of the *independence* of X_j from the other predictors (i.e., measures the *collinearity*)

$$\text{Variance Inflation Factor}_j = \frac{1}{1 - R_j^2}$$

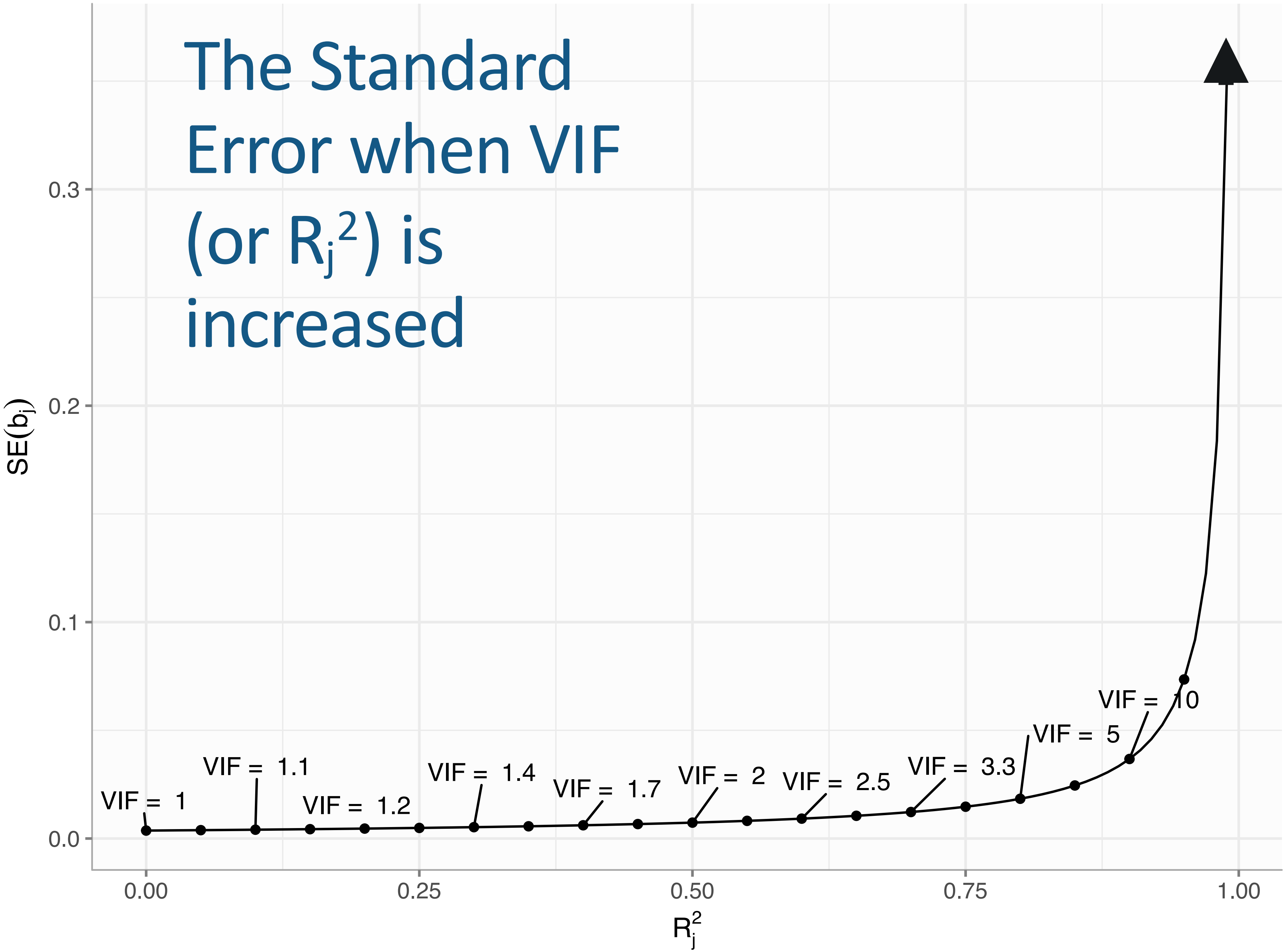
Inference: Regression Coefficients

The Standard Error

$$SE(b_j) = \sqrt{\frac{MS_{residual}}{(N)\,Var(X_j)\,(1 - R_j^2)}}$$

$$SE(b_j) = \sqrt{VIF_j} * \sqrt{\frac{MS_{residual}}{(N)\,Var(X_j)}}$$

The Standard Error when VIF (or R_j^2) is increased



Inference: Regression Coefficients

Using the Standard Error we can now do two important things

Null Hypothesis Test

$$t = \frac{b_j - \text{null value of } b_j}{SE(b_j)}$$

Confidence Interval

$$CI = b_j \pm t_{\alpha/2} * SE(b_j)$$

Using either we can test the null hypothesis and make inferences about the population

Inference: Partial Correlation

This testing each individual predictor

- Do each predictor have a relationship with the outcome?
- *Less common but still used*
- Directly tied to the t for b_j
  - *Just in different units (or in this case, no units)*
- Less **robust** if not testing if $H_0 = 0$ (requires bivariate normality)

The ability for a method to give accurate results even when assumptions don't hold

Inference: Partial Correlation

This testing each individual predictor

- Do each predictor have a relationship with the outcome?
- *Less common but still used*

Statistic of Interest	Test Statistic	Significance	Example
$r_{partial}$	T-statistic	P < .05 suggests there is a correlation among this predictor and the outcome	Controlling for the covariates, the correlation between SES and the outcome is $r_{partial}$.

Inference: Partial Correlation

Confidence intervals are tougher here

- *Since there are bounds (i.e., can't be below 0 or above 1)*

See Page 115 for the steps to obtain this

Inference: Conditional Means

First thing, let’s talk about **centering**

Centering a variable means subtracting a centering-value from it

- We can *mean* center
- We can *median* center
- We can center on *any value* we choose

When we do this, it changes the interpretation of the **intercept**

Inference: Conditional Means

To obtain $SE(\hat{Y}_G)$ where G is a specific set of points

→ Center each variable at that specific set of points

For example, we may want to know the language ability of a child and obtain the confidence interval of that estimate for a someone that is 8 years old and whose mother has 15 years of schooling (some college)

Some Miscellaneous Issues

1. Collinearity – how bad is it?
2. Contradicting Inferences – is regression lying?
3. Sample size and non-significance – should we remove non-significant predictors?