Math Workshop 2022

Christina Walker

Contents

1 Basic Statistics

1.1 Notation

Common Statistical and OLS Regression Symbols and Terms – and Their Translations

Symbol	Translation
n	Sample Size
H_0	Null Hypothesis
α	P(Type I Error)
β	P(Type II Error)
$N(a, b)$	Normal (Mean = a; Var [or SD] = b)
Φ	Standard Normal CDF
χ^2	Chi-Square Distribution
$\sim$	Is Distributed As
$\overset{\rightarrow}{\sim}$	Approaches Being Distributed As
$Corr(x, y)$ or r_{xy}	Correlation of x and y
$E(x)$	Expected Value of x
RSS or SSR	Regression Sum of Squares
ANOVA	Analysis of Variance
UMVU	Uniform Minimum Variance Unbiased
GLS	Generalized Least Squares
OLS	Ordinary Least Squares
y	Dependent Variable
R^2	Coefficient of Determination
R_j^2	Auxiliary R-Square for X_j
Tol_j	Tolerance for X_j
D	Cook's Distance

Symbol	Translation
N	Population Size
H_1 or H_a	Alternative Hypothesis
$1 - \alpha$	Confidence Level of a Test
$1 - \beta$	Power of a Test
Z	Standard Normal N(0,1)
T or t	Student's t Distribution
F	F Distribution
$\dot{\sim}$	Is Approximately Distributed As
df	Degrees of Freedom
$Cov(x, y)$	Covariance of x and y
TSS or SST	Total Sum of Squares
ESS or SSE	Error Sum of Squares
MLE	Maximum Likelihood Estimation
BLUE	Best Linear Unbiased Estimator
WLS	Weighted Least Squares
x_j	The j^th Independent Variable
$\hat{y}$	"Y Hat" (Predicted Value of Y)
$\bar{R}^2$	Adjusted R-Square
VIF_j	Variance Inflation Factor for X_j
p	P-Value or Prob Value
d	Durbin-Watson Statistic

Common Population Parameters and Sample Statistics – and Their Translations

Population Parameter	Sample Statistic	Common Translation
μ	$\bar{x}$	Mean (Average)
σ^2	$\hat{\sigma}^2$ or s^2	Variance
σ	$\hat{\sigma}$ or s	Standard Deviation
ρ	$\hat{\rho}$ or p or r	(Pearson's) Correlation

Population Parameter	Sample Statistic	Common Translation
α, β_0	$\hat{\alpha}$ or $\hat{\beta}_0$ or b_0	Regression Intercept
β	$\hat{\beta}$ or b	Regression Slope Coefficient
ε	$\hat{\varepsilon}$ or e or u	Regression Error (Residual)
σ_ε	$\hat{\sigma}_\varepsilon$ or $s_{\hat{\varepsilon}}$	Standard Error of Regression

Greek Letters – and Their (Rough) Equivalences

Greek Letter			"Equivalent" to:	
Upper	Lower	Name	Upper	Lower
Α	α	Alpha	A	a
Β	β	Beta	B	b
Γ	γ	Gamma	G	g
Δ	δ	Delta	D	d
Ε	ε or ϵ	Epsilon	E	e
Ζ	ζ	Zeta	Z	z
Η	η	Eta	H	h
Θ	θ or ϑ	Theta	Q	q
Ι	ι	Iota	I	i
Κ	κ	Kappa	K	k
Λ	λ	Lambda	L	l
Μ	μ	Mu	M	m

Greek Letter			"Equivalent" to:	
Upper	Lower	Name	Upper	Lower
Ν	ν	Nu	N	n
Ξ	ξ	Xi	X	x
Ο	ο	Omicron	O	o
Π	π or ϖ	Pi	P	p
Ρ	ρ	Rho	R	r
Σ	σ or ς	Sigma	S	s
Τ	τ	Tau	T	t
Υ	υ or ϒ	Upsilon	U or Y	u or y
Φ	φ or ϕ	Phi	F	f
Χ	χ	Chi	C	c
Ψ	ψ	Psi	U	u
Ω	ω	Omega	W	w

Common Mathematical Symbols and Terms – and Their Translations

Symbol	Translation	Symbol	Translation
$\sum$	Addition (Summation) Operator	$\prod$	Multiplication (Product) Operator
$\forall$	For All	$\therefore$	Therefore
st or $\ni$	Such That	$\because$	Because
$\in$	Is an Element of	$\exists$	There Exists
$\cup$	Union ("Or")	$\cap$	Intersection ("And")
$\subseteq$	Is a Subset of	$\subset$	Is a Proper Subset of
$\varnothing$ or ϕ	Null (Empty) Set	$\parallel$	Parallel
$\pm$	Plus or Minus	$\perp\!\!\!\perp$	Independent
wlog	Without Loss of Generality	$\perp$	Perpendicular
ow	Otherwise	$\angle$	Angle
$A \Rightarrow B$	A Implies B	iff or $\leftrightarrow$	If and Only If
$A \Leftrightarrow B$	"A Implies B" and "B Implies A"	Δ or δ	Change In (Delta)
lim or $\rightarrow$	Has a Limit of (Approaches)	plim or $\xrightarrow{p}$	Has a Probability Limit of
$\equiv$	Equal by Definition or Assumption	$\doteq$	Approaches (Limit) Being Equal to
$\approx$ or $\cong$	Is Approximately Equal To	$\neq$	Not Equal To
argmax	Value that Maximizes a Function	argmin	Value that Minimizes a Function
max	Maximum Value	min	Minimum Value
$\bar{A}$ or A' or A^c	The Complement of the Event "A"	$\neg$ or $\sim$	Logical Negation ("Not")
$<$	Less Than	$>$	Greater Than
$\leq$	Less Than or Equal To	$\geq$	Greater Than or Equal To
QED or ■ or □	Shows that a Proof is Complete	$(f \circ g)(x)$	Composition of Functions: f(g(x))
e	Euler's (Exponential) Constant	exp	Exponential (Power of "e")
π	Pi (3.14159...)	!	Factorial
$\log_a$	Logarithm (Base "a")	ln	Natural (Base "e") Logarithm
∂	Partial Derivative	$\int$	Integrate or Integral
∞	Infinity	$\propto$	Is Proportional To
sin	Sine	arcsin	Arc Sine
cos	Cosine	arccos	Arc Cosine
tan	Tangent	arctan	Arc Tangent
csc	Cosecant	sinh	Hyperbolic Sine
sec	Secant	cosh	Hyperbolic Cosine
cot	Cotangent	tanh	Hyperbolic Tangent
mod	Modulo Function	deg or $^\circ$	Degrees
rad	Radians	del or ∇	Gradient (Grad)
$\bullet$	Dot (Inner, Scalar) Product	$\times$	Cross (Outer) Product
$\odot$ or $*$ or $\circ$	Hadamard (Schur) Product	$\otimes$	Kronecker (Tensor) Product
$\lVert \mathbf{A} \rVert$	Norm of the **A** Vector	det(**A**) or $\lvert \mathbf{A} \rvert$	Determinant of the **A** Matrix
Tr(**A**)	Trace of the **A** Matrix	Rank(**A**)	Rank of the **A** Matrix
$\mathbf{A}^{-1}$	Inverse of the **A** Matrix	$\mathbf{A}'$ or $\mathbf{A}^T$	Transpose of the **A** Matrix
I	Identity (Unit) Matrix	**J**	Matrix Consisting of All 1's
H	Hessian Matrix	λ	Lagrange Multiplier
$\mathbb{R}$ or **R** or $\Re$	The Real Numbers	$\mathbb{Q}$ or **Q**	The Rational Numbers
$\mathbb{Z}$ or **Z**	The Integers	i	$\sqrt{-1}$ (Unit Imaginary Number)
P(A)	Probability of the Event "A"	$P(A \mid B)$	Conditional Probability ("A Given B")
pmf or PMF	Probability Mass Function	pdf or PDF	Probability Density Function
cdf or CDF	Cumulative Distribution Function	iid	Independent & Identically Distributed

1.2 Descriptive Statistics

Statistic	Definition	Formula
Mean	Average Value	$\frac{1}{n}\sum_{i=i}^{n} x_i$
Median	Exact Center when Ordered	
Mode	Most Frequently Occurring Value	
Standard Deviation	Indication of Spread	$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$
Variance	How Much Variation in RV, Relative to its Sample Mean	$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{n-1}$
Range	Spread of Data	Maximum-Minimum
IQR	Middle 50% of Data	Q3-Q1

Standard Errors: how far errors are from the estimate, the standard deviation of the sampling distribution

Why do we divide by n-1? removes degree of freedom, penalizes small sample size

Degrees of freedom: number of observations that are free to vary

5 number summary:

1. Minimum
2. Q1
3. Median (Q2)
4. Q4
5. Maximum

1.3 Correlation

Correlation: Measures the strength and direction of a linear relationship (R)

-1	0	1
X increases, Y decreases	No predictable pattern	X Increases, Y Increases

Visualization:

1.4 Causality

Causality is where one action causes outcome of another (i.e., A causes B)

Causal inference: Comparison between factual and counterfactual where the key causal variable of interest is the treatment variable

The fundamental problem of causal inference: the counterfactual cannot be observed, we must use causal identification strategies

The goal of political science work is often evaluating causal theories but there are four causal hurdles:

1. Is there a credible causal mechanism that connects x to y?
2. Can we rule out the possibly that y causes x?
3. Is there covariation between x and y?
4. Have we controlled for confounding variables, z, that might make the relationship between x and y spurious [not valid]?

There are two broad approaches to designing research:

1. Experimental Design: The researcher controls and randomly assigns values of the I.V. to subjects
2. Observational Studies: The researcher does not have control over the I.V. values which occur naturally. Decreased internal validity (selection bias), pretreatment variables may differ between groups. Increased external validity: can examine treatments that are implemented in the relevant population (i.e., diff in diff, within, before/after, cross section). Uses statistical control, the researcher tries to adjust for confounders.

Experiments:

Tim McDaniel

Sampling Methods

Remember the "big picture" goal here: Select a sample in an efficient (e.g., cost, time) manner that is sufficiently representative of the population.

Sample Type	Description	Example(s)	Pros ("Good") and Cons ("Bad")
Random (also known as **"Simple Random"**)	The population elements always have equal chances of being selected into the sample. There is nothing systematic about the selection process; instead, it is totally random (i.e., pattern-free). The probability of an element being selected does not depend upon whether some other specific element was selected (i.e., selections are independent across elements).	Generate random numbers and use them to select some stocks from the NY Stock Exchange. Program a machine to randomly dial telephone numbers for a national survey.	Pros: Usually representative of the population. No selection bias (i.e., nothing in the selection process makes the sample unrepresentative). Cons: Can be resource expensive, or impossible, to implement (especially with populations that are very large, or hard to define or locate).
Stratified Random	Based upon some relevant criteria, the population is divided into subgroups called "strata." (Note: The more similar the elements within a stratum are, the better.) Then random samples are selected from within each of these strata.	Divide the population into socioeconomic cohorts (e.g., Low, Medium, High) then randomly select some elements from each of these cohorts.	Pros: The sample is representative of the population relative to the grouping criteria. Cons: The sample may not be representative of the population relative to other factors.
Cluster	The population is divided (sometimes intentionally, other times naturally) into groups called "clusters." (Note: The more representative of the population each cluster is, the better.) Then a random sample of clusters is selected, with all of the elements in each selected cluster becoming part of the sample.	Randomly select six homerooms at a high school, with every student in each homeroom being part of the sample. Survey all of the houses in six randomly selected city blocks.	Pros: Can be cheap, fast, and easy – especially if the population is already divided into clusters (e.g., homerooms, city blocks). Cons: If each selected cluster is not representative of the population then the sample will not be representative either.
Systematic	Elements are selected from the population in a systematic (non-random; i.e., orderly, methodical, and with an intentional pattern) manner.	Select the person listed first on each page in a telephone book. Select every third fly on a slide.	Pros: Can be cheap, fast, and "almost random." Cons: If the population is not randomly distributed then the sample will not be representative.
Convenience	Pretty much just what you would think: Elements are selected from the population according to whatever is deemed the most convenient (easiest, quickest, most simple) method and means.	When conducting a survey in a mall, select people who seem to be the most potentially amenable as they walk near you.	Pros: Can be extremely cheap, fast, and easy. Cons: An extremely high risk of selection bias, so the sample would not be representative.
Judgment	Elements are carefully, thoughtfully, and deliberately selected from the population in order to construct a sample that is, in the judgment of the experienced and knowledgeable person conducting the study, highly representative of that population.	A political scientist selects five counties she deems collectively representative of the state across many different criteria (urban vs. rural, rich vs. poor, race, etc.)	Pros: Can be cheap, fast, and easy. If done well it can result in a highly representative sample. Cons: If done poorly it can result in a highly biased and unrepresentative sample.

Bias	Definition	Example
Sampling	Some people have higher chance of selection	Phonebook
Non-response	People not responding & responding differ	Survey QR code
Omitted Variable	Leaving out variables that should be in model	Leaves out party
Voluntary	People with strong feelings are more likely to respond	People love/hate candidate
Social Desirability	Respond incorrectly due to social acceptability	Taliban
Framing	Question is worded in a way people don't answer based on facts	Leading question

Hawthorne effect: people behave differently if they know they are being studied

Placebo Effect: beneficial effect produced by placebo that cannot be attributed to treatment effect

SATE Formula: $\frac{1}{n}Y_i(1) - Yi(0)$ (treatment-control). The average treatment effect compared to control

Difference in Difference: ($\bar{y}$ after treatment - $\bar{y}$ before treatment)-($\bar{y}$ after control - $\bar{y}$ before control)

Visualization:

Before and After/Within Unit: compare the same unit before and after treatment → has time bias. Compare NJ before and after, able to adjust for unit specific issues.

Cross Section: unit and time bias, compare PA and NJ after treatment, compare treated unit with control unit after treatment, unit specific and time confounding

2 Research Process

Good Research:

1. Theorize Effect
2. Collect Data
3. Test Only That Effect
4. If $p < .05$, Conclude Evidence for Effect

Bad Research:

1. Collect Data
2. Test Many Effects
3. Find Where $p < .05$
4. Conclude Effect

Theory: A statement of the possible causal relationship between 2+ concepts

To come to a conclusion about whether our theory is likely to be correct, we make empirical observations and compare abstract, theoretical ideas with reality.

2.1 Hypothesis Testing

Hypothesis: a statement about the world that could be tested to be true or false

Suppose we want to test a hypothesis dealing with the mean of a population, so testing whether $\mu =$ (some number). We decide on the value of this number before we compute any hypothesis testing statistics. This hypothesis is the null hypothesis, H_0.

If we say $H_0 : \mu = k$ where k is some fixed, pre-determined constant, then our alternative hypothesis must be everything else, $H_a : \mu \neq k$

Using hypothesis testing, we can see how sure we are that the mean of a population = 0 (for example). Since we are seldom, if ever, able to take the mean of an entire population, we generally draw a random sample and estimate μ using $\bar{x}$.

The alternative hypothesis ($H_1 or H_A$) is what we are seeking evidence for, as we are trying to find if the sample is extreme enough (from null) to suggest the alternative is true.

If our hypothesis is that $H_0 : \mu = 0$ (no effect), if we draw a sample and find that $\bar{x}$ is close to 0, we can be confident that $\mu = 0$ is true.

The closer $\bar{x}$ is to 0, the more confident we are that the hypothesis is true.

We can never be 100% sure of what our population parameter's true value is. Instead, we can only guess at (estimate) it based on our sample and its corresponding sample statistics.

Parameters are unknowable fixed values we try to estimate. They themselves do not have uncertainty, they are "godly" figures we estimate by taking a sample and estimating sample statistics.

Common parameters:
μ = mu = mean of numerical variable = $\bar{x}$
σ = sigma = standard deviation = s
π = pi = proportion of categorical variable = p
ρ = rho = correlation between 2 variables = r
β= beta = gradient between variables = b

To test the null, we temporarily assume that it is true. Then if

- Nothing too strange/unexpected is observed, we have no reason to think the null is untrue
- If something strange/unexpected is observed, then this indicates the assumption of the null was most likely wrong, so we have reason to think the null is not true and that the alternative may be true.

We are confident but not sure. We never prove anything in a hypothesis test, we can only infer.

We never accept the null, we only fail to reject the null or say there is not enough evidence to suggest the null is incorrect. So we either (1) fail to reject the null or (b) reject the null and accept the alternative hypothesis.

2.2 Standard Deviation

What is considered "close"?

We measure closeness in terms of standard deviation units.

Since we hardly (if ever) know the population standard deviation, we use our sample to compute s, an estimate of σ.

If we assume our population is normally distributed, we estimate the variance using S^2.

Since our null hypothesis is $\mu = 0$, we assume this is true and see if anything contradictory happens. If anything does happen (like $\bar{x}$ is many standard deviation units from 0) then we will not be very comfortable with out assumption that $\mu = 0$. Remember x_{stand} tells us how many standard deviation units x-bar is from mu, where we assume mu = 0.

2.3 P-Value

What does it mean to be extreme enough?

The Universal Decision Rule: A general rule for determining whether to reject H_0 or not. At the $1 - \alpha$ level of confidence, reject H_0 if the p-value is less than α. Fail to reject H_0 at the $1 - \alpha$ confidence level, if the p-value is greater than α.

We construct a rejection region by finding a point on the x-axis which we consider too extreme. You can customize it based on how strict you want to be, but most common is 5%.

P-Value: Measure of how extreme our sample is, it is how likely we would be to get this statistic if null was true. The probability of observing another $\bar{x}$ which is even more standard deviation units from 0. The probability that some value for our test statistic is at least as extreme as the one we have observed.

The smaller the p-value (i.e., the less are shaded in the tails) the farther X_{stand} is from 0 and the more likely we are to **reject** $H_0 : \mu = 0$. The further $\bar{x}$ is from 0, the large $|x_{stand}|$ is, so the smaller the p-value is, the less confident we are that $H_0 : \mu = 0$ is really true.

Visualization:

What 3 things impact statistical significance:

- Size of the coefficient
- Size of standard error
- The number of observations

Type I: Mistakenly reject the null hypothesis, finding an effect when there is not one, the probability of Type I error is significance level, so lower values of alpha reduce the probability of a Type I error

Type II: Higher values of alpha reduce probability of Type II error, mistakenly fail to reject the null hypothesis

Discussion about the Validity of P-Values: Most journals expect the p-values by reported and utilized in statistical analysis. However, there are lots of people who argue that the p-value is given way more weight than it should be.

For example, Wasserstain, Schirm, and Lazar (2019) introduction to The American Statistician Journal about the use and abuse of p-values. Some of their points are: (1) There is a tiny difference between being 95.1% confident and 94.9%, but we treat them as all or nothing, (2) there is nothing special about 95%, they are arbitrary, (3) there are factors besides chance and randomness that determine the size of the probability of observing even more evidence for the alternative hypothesis than you already have, (4) an association/relationship still may exist even if p-value is not significant or vice versa, (5) a p-value does not consider the substantive important of the magnitude of the relationship [magnitude: absolute values, or distance from 0].

HOW DO WE INTERPRET, EXPLAIN, AND USE A P-VALUE (“PROB-VALUE”)?

Mathematical statisticians figured out the specifics of the Critical Value, the Test Statistic value, and the p-value using calculus involving the mathematical formula for the relevant probability distribution (e.g., Normal; Student’s t; Fisher’s F; Chi Square; etc.). This was all done under the (temporary, for the purposes of the hypothesis test itself) assumption that “the Null Hypothesis is true.”

Boiling it down and omitting the gory math details:

> **The p-value is the probability−if you did all this again, using a different sample−of observing even** ***more evidence for the Alternative*** **(i.e., against the Null)** ***Hypothesis*** **than you have now.**

p-value < α	p-value > α
↓	↓
The p-value is small.	The p-value is large.
↓	↓
There is a just a small probability of observing even more evidence for the Alternative Hypothesis than you have now.	There is a large probability of observing even more evidence for the Alternative Hypothesis than you have now.
↓	↓
So you must have a lot of evidence for the Alternative Hypothesis now!	So you must have just a little bit of evidence for the Alternative Hypothesis now!
↓	↓
The Alternative Hypothesis probably is true.	The Alternative Hypothesis probably is not true.
↓	↓
We are $(1- \alpha)\%$ confident that H_1 is true.	We are not $(1- \alpha)\%$ confident that H_1 is true.
↓	↓
At the $(1- \alpha)\%$ confidence level, we **Accept H_1**.	At the $(1- \alpha)\%$ confidence level, we **Reject H_1**.

Way back in the day (perhaps in your “Introduction to Statistics” course!) you had to draw some pictures, read a probability table, and do some calculating to determine the value of a p-value. But of course now that grunt work is all done, and the p-value simply reported, by statistical software.

I want you to have an intuitive feel for and conceptual understanding of (as opposed to merely a superficial plug-and-chug cookbook know-how about) what a p-value is and how to interpret, explain, and use it. Hopefully these last few pages have put you on that path. But do not be discouraged if you pretty much, but perhaps not 100% fully, understand all of this; we just covered (and, hopefully for you, reviewed) a good-sized portion of an “Introduction to Statistics” course relatively quickly!

TO "P" OR NOT TO "P": THAT IS THE QUESTION...

A current discussion and debate in statistics involves the appropriateness of using, or even reporting, p-values (sometimes called "Prob-values") in statistical work – including work involving linear regression, the primary subject considered in this course.

The use of p-values has been common and widespread in statistics, most specifically in work involving hypothesis testing. There is a rich and extensive legacy of published statistical research that uses p-values to provide evidence regarding research conclusions. Most (though this number might be declining...) journals expect that p-values be reported and utilized in many types of statistical analysis.

Therefore, **even if using p-values magically went away tomorrow, it would still be important that we understand what they are all about – i.e., how to *properly* interpret, explain, and use them.**

One of the readings (Wasserstein, Schirm, and Lazar) listed on the course syllabus is an editorial introduction to a special 2019 edition of *The American Statistician* journal devoted to the use, abuse, and misuse of p-values. Here is a summary of that summary of some of the points presented in the 43 papers in that issue. (Clearly, I have omitted or glossed over a LOT of details here [you're welcome!].)

- Suppose you are working at the 95% ($\alpha = 0.05$) confidence level. There is just a tiny difference between being 95.1% ($p = 0.049 < 0.05$) versus 94.9% ($p = 0.051 > 0.05$) confident about differing results of a hypothesis test. So why do we treat and report these tiny differing p-value results in such a stark dichotomous all-or-nothing "significant or not significant" manner?
- What is so special about a 95% confidence level? Or 90%? Or 99%? Even if we accept the current use of p-values and significance testing, the specific threshold value involved (90% or 95% or 99% or 90-whatever%) – even if traditionally used – is still, to a large degree, arbitrary.
- There are other factors besides chance and randomness that determine the size of "the probability of observing even more evidence for the Alternative Hypothesis than you have now."
- An association or relationship (or, with some methodologies, an effect) still might not really-truly exist even if "p-value < α" or might really-truly exist even if "p-value > α".
- Phrases like "statistically significant" and the "* or ** or ***" stars notation should not be used.
- When addressing the broad question of "significance," a p-value does not consider the contextual and practical importance of the *size* of a relationship. The p-value only considers the "effect size" relative to a single specific hypothetical Null Hypothesis parameter value (e.g., zero).
- Best case, p-values are merely one of many types of evidence that should be reported and considered, as opposed to a be-all-and-end-all "proof," regarding a relationship or association.

Here is a synopsis of my (evolving...) views on this matter:

> I think that p-values are worthy of reporting. They can be meaningful and add descriptive insight if properly used. But the misuse of p-values induces false and inappropriate certainty into questions and methods that are all about measuring, reporting, and managing things are inherently uncertain. We should accept and embrace that uncertainty instead of trying to make it arbitrarily disappear.

See the readings (and, if you want, me) for much more on this evolving topic.

2.4 Confidence Intervals

We want to test our estimate of parameter and see how “confident” we are about this assumed value.

We never reject or fail to reject in absolute terms, instead we can be at most 100(1-α)% confident when rejecting H_0. Just as it was important to decide on a value for k before conducting this procedure, it is important to decide on a minimum acceptable confidence level for rejecting H_0 before computing the p-value.

If we want to be at least 95% confident that H_0 is true, our desired confidence level is set to .95.
3 ways to tell Statistical Significance:

- is $p \leq .05$?
- is t greater than 1.96 in two tailed test?
- are the confidence interval bounds the same sign?

Level of Confidence	Alpha	Since...	T-Stat
90% (.90)	.10	.90 = 1-.10	1.64
95% (.95)	.05	.95 = 1-.05	1.96
99% (.99)	.01	.99 = 1-.01	2.58

Confidence Intervals quantify uncertainty

If a certain interval is a 95% confidence interval for μ, that means that if I repeated the procedure of drawing random samples and computing confidence intervals, 95% of those confidence intervals would include the actual value of μ.

It is **incorrect** to say if [-2.7, 3.1] is a 95% confidence interval for μ, then P($-2.7 < \mu < 3.1$) = .95. This is a common but incorrect interpretation. Instead, I am $1 - \alpha$% confident that a confidence interval includes μ based not on this single confidence interval but rather as a result of what would happen if I repeated the process of drawing random samples and computing confidence intervals over and over.

Formula: $\bar{x} - 1.96 * SE, \bar{x} + 1.96 * SE$

Calculate CIs:

1. compute standard error for sample mean or proportion
2. choose a level of confidence and z-score, t-value if small n
3. calculate lower and upper bound

Conditions for Valid CIs:

1. Random sample
2. Normality
3. Independence

2.5 Measurement

We need to be as confident as possible that the concepts in our theory correspond as closely as possible to empirical observations. Measuring concepts with care is one of the most important aspects of social science. If empirical analysis is based on measures that do not capture the essence of our theory, we are unlikely to have confidence in our findings.

I.V. (concept) → causal theory → D.V. (concept) D.V. (concept) → D.V. (measured) → operationalization I.V. (measured) → hypothesis → D.V. (measured)
3 Issues of Measurement:

1. Conceptual clarity: what is the exact nature of the concept we are trying to measure?

2. Reliability: an operational measure of a concept is reliable to the extent it is repeatable or consistent (i.e., applying the same measurement rules produces same result)

3. Validity: a valid measure accurately represents the concept it is supposed to measure

Measuring D.V.: Identify the (1) time dimension, the point or points in time we would like to measure the variable, and (2) spatial dimension, the physical units we want to measure

The D.V. is then either (1) time series, where the spatial dimension is the same for all cases and D.V. is measured at multiple time points, or (2) cross sectional, where the time dimension is the same for all cases and D.V. is measured for multiple spatial units.

Longitudinal/panel data: multiple measurements on the same units over a long time, more credible than cross sections

Once measurement is conducted, it is important for the researcher to get a good idea of the types of values that the individual variables take on before moving to test causal connections, you can run Crosstabs or Histograms to do this.

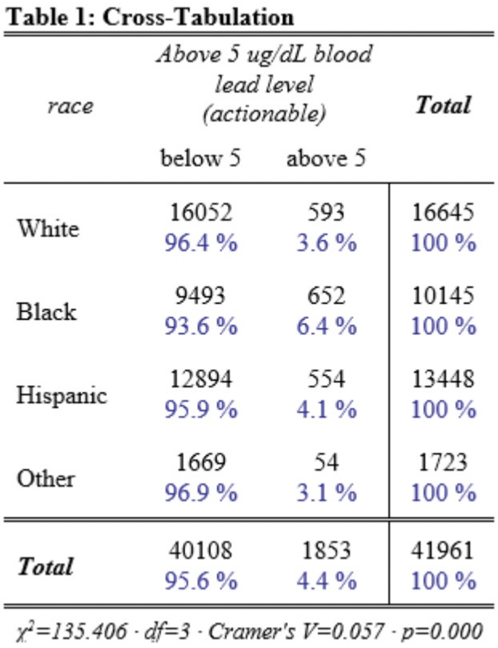


Table 1: Cross-Tabulation

race	*Above 5 ug/dL blood lead level (actionable)*		***Total***
	below 5	above 5	
White	16052 96.4 %	593 3.6 %	16645 100 %
Black	9493 93.6 %	652 6.4 %	10145 100 %
Hispanic	12894 95.9 %	554 4.1 %	13448 100 %
Other	1669 96.9 %	54 3.1 %	1723 100 %
Total	40108 95.6 %	1853 4.4 %	41961 100 %

χ^2=135.406 · df=3 · Cramer's V=0.057 · p=0.000

3 Distributions

Explaining Distributions:

1. Normal

2. Unimodal/bimodal

3. Outliers

4. Deviations
5. Symmetric, skewed

The total area under a curve is equal to 1 but the exact shape is determined by the variance of the random variable, while the placement of the curve is determine by the mean.

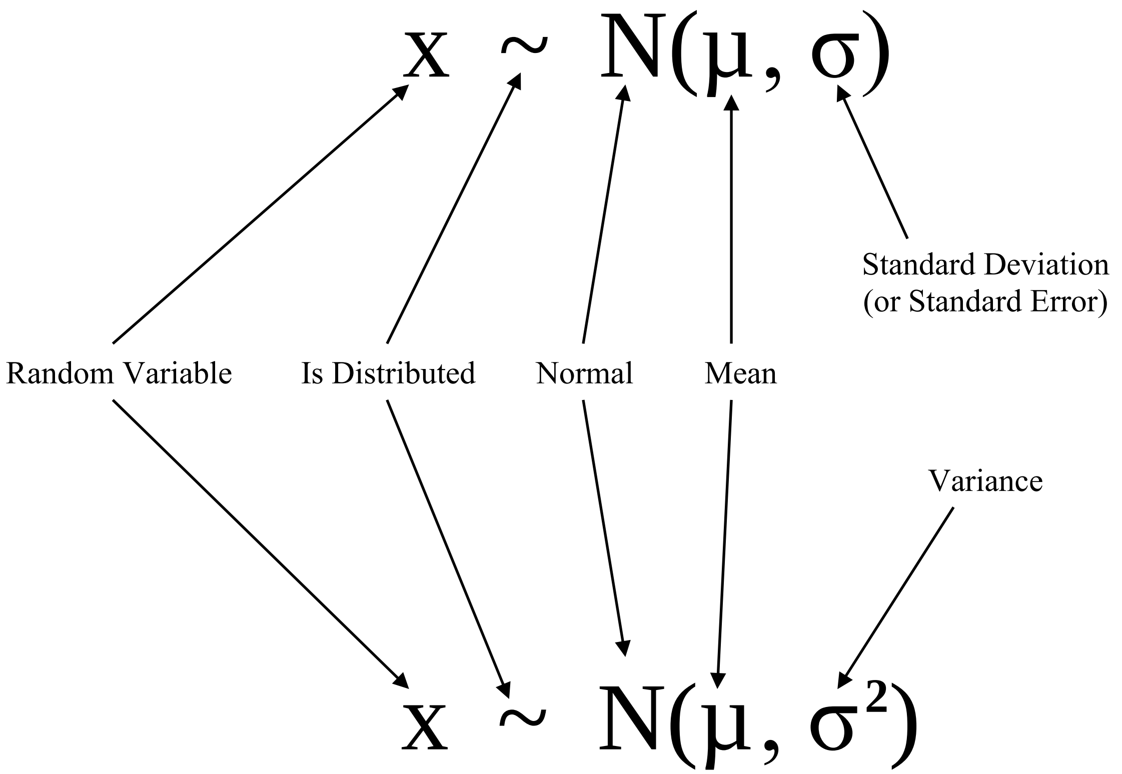
3.1 Normal Distribution

A bell-shaped curve that is symmetric around its mean.

A Brief Note on "Math Language" Notation

As with any language, in "Math Language" there are often different ways of expressing something.

For example: In Math Language there are two ways of stating that a variable has a Normal distribution and then specifying the properties that are sufficient for defining that Normal distribution.



As you can see from the Math Language presented above, sometimes the last argument is the value of the "Standard Deviation" (or "Standard Error") while other times it is the value of the "Variance."

It is generally no problem to figure out which dialect of Math Language is being used – IF you are aware of the general context in which it is being used.

Also: A *dot* over the ~ (i.e., "$\dot{\sim}$") translates as "...is *approximately* distributed as...."

We see this, for example, in the Central Limit Theorem:

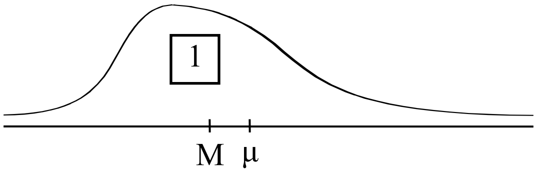
$\bar{x} \mathrel{\dot{\sim}} N(\mu, \sigma/\sqrt{n})$
"X-bar is distributed approximately normal,
with a mean of μ and a standard error of $\sigma/\sqrt{n}$."

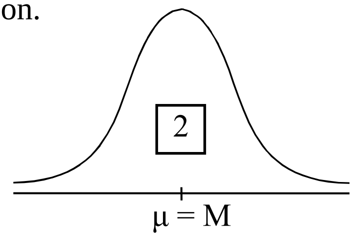
$\bar{x} \mathrel{\dot{\sim}} N(\mu, \sigma^2/n)$
"X-bar is distributed approximately normal,
with a mean of μ and a variance of σ^2/n."

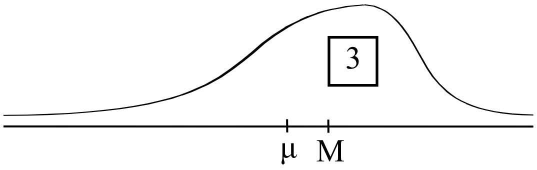
"Skewness" and "Kurtosis"

Sometimes a distribution is (or should be...) "almost Normal," but has some characteristics that result in a shape that is not quite exactly a Normal bell curve. Two common such characteristics are Skewness and Kurtosis.

Skewness -- Where the distribution is not symmetric because one tail is longer than the other, as if the bell curve has been stretched-out in that direction.







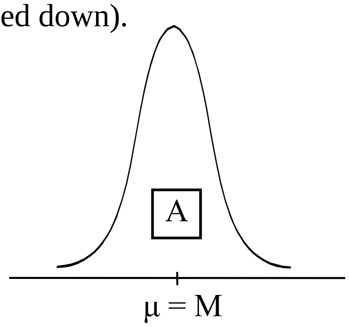
Case 1: **Positive (Right) Skewness** -- The curve is not symmetric.... Instead, one tail is stretched in the right ("positive") direction. This sometimes happens when the values have a lower bound and/or high-end outliers.

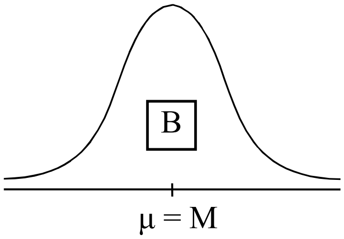
Case 2: **Zero (No) Skewness** -- The curve is symmetric and not stretched in either direction.

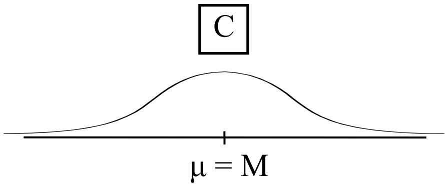
Case 3: **Negative (Left) Skewness** -- The curve is not symmetric.... Instead, one tail is stretched in the left ("negative") direction. This sometimes happens when the values have an upper bound and/or low-end outliers.

With skewness you should report the Mean ("μ"; the "balancing point" if the curve was a flat rigid surface of uniform thickness and density) *and* the Median ("M"; the "middle point" that divides the area [values] in half).

Kurtosis -- Where the distribution is symmetric, but either more peaked (stretched up) or less peaked (squashed down).







Case A: **Positive (Leptokurtic) Kurtosis** -- The curve is "stretched up" to have a higher peak and fatter tails. ("Lepto" means "Slender.") A *higher* probability of values near the mean and also values in the tails (outliers).

Case B: **Zero (Mesokurtic) Kurtosis** -- The curve is neither "stretched up" nor "squashed down" relative to a Normal distribution. ("Meso" means "Intermediate / Middle.")

Case C: **Negative (Platykurtic) Kurtosis** -- The curve is "squashed down" to have a lower peak and thinner tails. ("Platy" means "Flat.") A *lower* probability of values near the mean and also values in the tails (outliers).

Sometimes, using a particular strict mathematical formula, Mesokurtic distributions have a kurtosis value of three; but generally folks subtract three from this kurtosis value so they have a kurtosis value equal to zero.

N(0,1) is a special distribution called the standard normal distribution (or the z distribution)

3.1.1 Skewness & Kurtosis

Skewness: the distribution is not symmetric because one tail is longer than the other. Whatever way the tail is longer is the direction of skew. With skew you should report the mean and median. For left skew, the mean is smaller than the median. For right skew, the mean is larger than the median.

Kurtosis: The distribution is symmetric but more or less peaked. If the curve is stretched up with a higher peak and fatter tails it is Laptokurtic, there is a higher probability of values near the mean and values in the tails (outliers). A Mesokurtic is neither stretched or squashed. A Platykurtic is squashed down with a lower peak and thinner tails, there is a lower probability of values near the mean and in the tails (outliers).

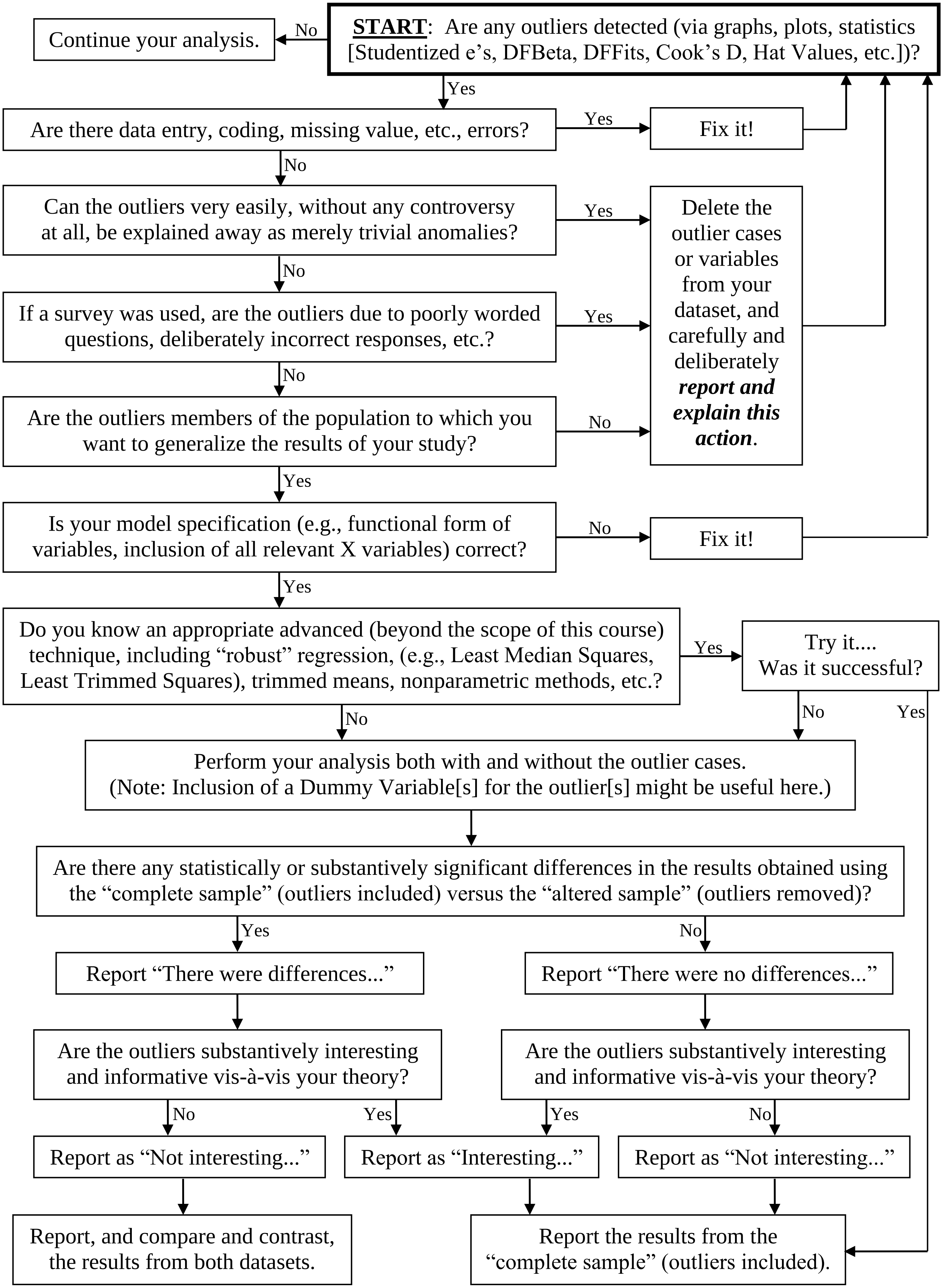
RMSE (Root Mean Square Error):$\sqrt{\frac{(residual)^2+(residual_2)^2...+(residual_n)^2}{n}}$ or $\sqrt{\frac{RSS}{n}}$

A higher RMSE means higher variance (flatter curve). RMSE is the standard deviation of the error, how spread out residuals are from line.

Outliers: 1.5 * IQR above upper quartile or below lower quartile.
1st quartile is first 25% of observations x * .25, 3rd quartile is 75%.

Outliers:
1. Unusual IV values = leverage
2. Large residual values
3. Both leverage and large residuals
Influence = leverage and residual

Tim McDaniel's Guide to Dealing with Outliers



START: Are any outliers detected (via graphs, plots, statistics [Studentized e's, DFBeta, DFFits, Cook's D, Hat Values, etc.])?

Continue your analysis.

Are there data entry, coding, missing value, etc., errors?

Fix it!

Can the outliers very easily, without any controversy at all, be explained away as merely trivial anomalies?

If a survey was used, are the outliers due to poorly worded questions, deliberately incorrect responses, etc.?

Are the outliers members of the population to which you want to generalize the results of your study?

Delete the outlier cases or variables from your dataset, and carefully and deliberately ***report and explain this action***.

Is your model specification (e.g., functional form of variables, inclusion of all relevant X variables) correct?

Fix it!

Do you know an appropriate advanced (beyond the scope of this course) technique, including "robust" regression, (e.g., Least Median Squares, Least Trimmed Squares), trimmed means, nonparametric methods, etc.?

Try it....
Was it successful?

Perform your analysis both with and without the outlier cases.
(Note: Inclusion of a Dummy Variable[s] for the outlier[s] might be useful here.)

Are there any statistically or substantively significant differences in the results obtained using the "complete sample" (outliers included) versus the "altered sample" (outliers removed)?

Report "There were differences..."

Report "There were no differences..."

Are the outliers substantively interesting and informative vis-à-vis your theory?

Are the outliers substantively interesting and informative vis-à-vis your theory?

Report as "Not interesting..."

Report as "Interesting..."

Report as "Not interesting..."

Report, and compare and contrast, the results from both datasets.

Report the results from the "complete sample" (outliers included).

4 Probability Theory

Probability is a way of measuring uncertainty.
Since we will deal with sets containing a finite number of members, we can think of a probability as a proportion.

Example: You have a basket of 10 identical ping pong balls, each with a number, and exactly 3 have the number 7. The probability of drawing a ball at random with the number 7 is 3/10 or .3. This is written as P(drawing a 7) = .3

Combinations: higher denominator = smaller amount, selects objects without regard to their arrangement, AB and BA are the same
Permutations: AB and BA are different

Probability density function (PDF): how likely is it that x takes a particular value? Describes the distribution of the whole population of the probability of selecting someone at random, bulk will be in middle.

Cumulative density function (CDF): what is the probability that a random variable x takes a value equal to or less than x?

4.1 Basic Rules

For any probability:

1. P(event occurring) ≥ 0 and ≤ 1, where 0 = never and 1 = always
2. P(event occurring) = 1 - P(event not occurring)
3. If there are exactly n possible events, and no two of the events can happen simultaneously, then the sum of the probabilities is 1

Probability of not = 1-P(always)
P(at least 1) = 1-P(none)

4.2 Conditional

Conditional Probability: The probability of an event occurring given another event has already occurred or the probability of event A happening changes probability of B.

P(A|B) = "The probability of A, given B". | (vertical straight line) translates to given.

Conditional Hypothesis: The outcome will happen conditional on a second variable

Example: The probability that it will rain is less than the probability that it will rain given there are black clouds in the sky. So P(will rain) < (will rain | black clouds in sky)

4.3 Or & And

If we say P(A or B), we mean the probability of:

1. A happens, B does not
2. or B happens, A does not
3. or A and B both happen

Rule of Addition: P(A or B) = P(A) + P(B) - P(A) * P(B)
P(A and B) is the probability both A and B happen

If events are not independent, P(A and B)=P(A|B)P(B)

4.4 Independence

Independence: If x and y are independent RVs, there is no relationship at all between x and y. Any information about x gives you absolutely no clues about any characteristics of Y.

Two events are independent if P(A and B)=P(A)*P(b), P(A|B)=P(A)

Example: X = Price of rice in China; Y = Price of cigarettes in North Carolina

Whereas: X = Presidential Approval Rating; Y = Success of President's Party in Congressional Elections are not independent, they are dependent as information about x gives you some information about y - not perfect information but some

Notation: $X \perp\!\!\!\perp Y \rightarrow$ X and Y are independent

$X \not\perp\!\!\!\perp Y \rightarrow$ X and Y are not independent

As Corr(x,y) gets closer and closer to 0, this indicates a probable lack of any bivariate relationship between x and y. That is, the closer Corr(x,y) is to 0, the more evidence we have that $X \not\perp\!\!\!\perp Y$. But remember this is only measuring linear relationships.

So Corr(X, Y) = 0 could be observed when x and y are not independent.

Mutually Exclusive: events that cannot happen at the same time

4.4.1 Frequentest & Bayesian

Frequentest: repeated experiments → approximation, repeatable events

Bayesian: subjective belief → personal measure of uncertain

4.5 Random Variables

Variable: an empirical measure of a concept/characteristic that varies across observations

X = the independent, explanatory variable, predictor

Y = dependent, outcome, response variable

Visualization:

Unit of Observation: individual units, how you observe the unit of analysis or unique observations (i.e., individual-wave, state-month)

Unit of Analysis: making inferences on, the thing you are studying (i.e., individual, household, district), what you wish to say something about

Random Variables: Variables which take on values with some probability, generally a known (or closely estimated) probability

Variable Type	Definition	Example
Continuous	Exact Number	Exact Weight
Discrete	An estimation/categorization of the true value	Weight Rounded
Categorical	Observations belong to a discrete set of categories	Race, Gender
Nominal	No Order	Name, PhD Program
Ordinal	Ordered	Socioeconomic Status
Interval Level	Change from one category to the next is identical across all values	Temperature

It is usually impossible to make a random variable continuous but continuous random variables often have properties which are more desirable than discrete random variables. We sometimes estimate the actual value of a random variable as close as possible then pretend it is continuous. For example, rounding a person's weight to the nearest tenth of a pound may be good enough to treat the corresponding random variable as continuous although it is technically discrete.

Standard Random Variables: It is sometimes useful to know how many standard deviation units a particular random variables is from its mean, so we transform the random variable into a standardized random variable.

Z-score: the standardized random variable, tells us how many standard deviation units the corresponding random variable value is from the mean

$\frac{x-\mu}{\sigma}$

4.6 Central Limit Theorem

Central Limit Theorem: Given a population with **any** distribution and taking random independent samples of size 'n' from that distribution, the sample means of those independent samples will be approximately normally distributed with a mean equal to the mean of the population and variance equal to the variance of the population divided by n. The higher n is, the closer the distribution will be to normal. The distribution of the sample mean approaches a more normal distribution as the sample size increases. For any population with known mean = μ and known variance = σ^2, random sample of size n can be drawn. The mean of these independent samples approach a N(μ, $\frac{\sigma}{\sqrt{n}}$) as n increases.

The central limit theorem is important for calculating confidence intervals because it tells us how confident we can be that, over repeated random sampling, the population parameter will be in the confidence interval. The CLT is how we can use what we know about the sample to infer about the population, but it only applies to random samples.

4.7 Law of Large Numbers

Law of Large Numbers: As the sample size increases, it converges closer to the true parameter, also as repeated experiments increases, the experimental probability converges to the theoretical probability

5 Ordinary Least Squares

5.1 Bivariate Regression

$y = \hat{\alpha} + \hat{\beta}x + \hat{e}$
α is an estimate of the true intercept of the line of best fit
β is an estimate of the true slope of the line of best fit
e is an estimate of the true errors inherent in the line of best fit

Bivariate regression is between two variables

The goal is to fit the best line through a scatterplot of data, the line is defined by its slope and y-intercept

Line of Best Fit: We want the residuals (the vertical distance between observations and lines) to be as small as possible, it minimizes the sum of squared residuals

Slope/coefficient: a 1 unit change in x leads to a beta-unit change in y

$R\frac{s.d.y}{s.d.x}$

α is the intercept, it is what y equals when x = 0

$\alpha = \bar{y} - \beta(\bar{x})$

Scatterplots: Describe the:

- Form: Linear, nonlinear
- Direction: Positive, Negative
- Strength: strong, moderately strong, weak
- Outliers: any points unusually far
- Clusters

5.2 Assumptions

Tim McDaniel

A Very Quick-and-Dirty Overview of the Assumptions of Regression

The multiple regression model: $Y_i = a + b_1X_{1_i} + b_2X_{2_i} + ... + b_jX_{j_i} + ... + b_kX_{k_i} + e_i$ (Sample size = n)

Family	More Formally Stated	Less Formally Stated	Way Less Formally Stated
Measurement	No measurement error in Y or in any of the X’s.	The values in your sample do not systematically differ from the true unobserved population values.	Each of the X’s and Y are measured and recorded correctly, with no mistakes or inaccuracies.
Multicollinearity	**X** is of Full Rank. Rank(**X**) = k + 1 ($\mathbf{X}^T\mathbf{X}$) is nonsingular. $(\mathbf{X}'\mathbf{X})^{-1}$ exists.	No X variable is a perfect linear combination of all the other X’s. Each X variable is linearly independent from the set of all of the remaining X variables.	Each X is measuring something above-and-beyond different than what is being measured by all of the other X’s.
Specification	A linear relationship between each X_j and Y. dY/dX_j is a constant $\forall$ j.	For each X_j, when X_j increases by one then Y changes at a fixed rate – no matter if you are considering low, medium, or high X_j values.	The scatterplot of dots for each X_j and Y is straight as opposed to curvy.
	All relevant X’s are included in the model. No “omitted variables.”	If an X truly does belong in the model then it is in the model.	You have included all of the X’s that predict or explain Y.
	No irrelevant X’s are included in the model.	If an X truly does not belong in the model, then it is not in the model.	You have not included any X’s that do not predict or explain Y.
Error Terms	$E[e_i] = 0$ $\forall\ 1 \leq i \leq n$	Theoretically, in the long run, the residuals average out to (that is, have an expected value of) zero.	The sample regression equation is not shifted up or down from where it truly belongs.
	$Var(e_i) = E[e_i^2]$ is constant $\forall\ 1 \leq i \leq n$ Homoskedasticity No heteroskedasticity	The spread of the residuals is the same across all values (e.g., low, medium, and high) of an X.	The model does not do a better (i.e., more precise) job explaining or predicting Y for some values of an X than it does for others.
	$Cov(e_i, e_j) = E[e_i e_j] = 0$ $\forall\ 1 \leq i \neq j \leq n$ No autocorrelation No serial correlation	There is no linear relationship among different residual values. The residuals are linearly independent of each other.	The residuals have nothing to do with one another. Knowing something about one of the residual values gives you no clue about any other residual’s value.
	$Cov(e, X_j) = E[e X_j] = 0$ $\forall\ 1 \leq j \leq k$	There is no linear relationship between each X_j and e. Each set of X_j values and the set of residual values are linearly independent of each other.	The estimated slope between each X and Y is not more or less steep than it truly is. All of Y that is linearly related to these X’s is actually explained and predicted by these X’s.
	e ~ Normal	The residual values have a Normal distribution.	If you plot all of the residual values they form a bell curve.

Note: For every sample regression model, the “$E[e_i] = 0$” and “$Cov(e, X_j) = 0$” assumptions are *always* mathematically forced to be true. They are known as “Artifacts of Regression.”

5.3 Estimates

Estimand: Value of interest
Estimator: Method to compute estimate
Estimate: Approximation of a value

How good an estimate is depends on sample size, size of standard error, coefficient

Statistical Inference: Guessing/estimating what we do not observe from what we do

Sample Analogue Principle: use sample mean to infer population mean

Estimate Properties:

Unbiased: a sample statistic which estimates a population parameter is an unbiased estimate if its expected value is equal to the population parameter

Consistent: A sample statistic is a consistent estimator of a population parameter if as the sample size n gets larger, the expected value of the sample statistic approaches the actual value of the population parameter and its variance approaches 0. It converges to the parameter as data points increase. It becomes more accurate

Efficient: Assume we have two unbiased sample statistics which estimate the same population parameter. The sample statistic with the smaller variance is more efficient

Uniform Minimum Variance Unbiased (UMVU) estimate: The most efficient estimate is defined as the estimate with the least variance out of all unbiased estimates of a population parameter

Sufficient: A sample statistic is a sufficient estimator of a population parameter if it contains all of the information in the data about the value and other characteristics of that population parameter

5.4 Problems with OLS

Tim McDaniel

Some Potential Problems When Using Ordinary Least Squares (OLS) Regression

Situation	Reported Slope Coefficient Value (b_j)	Reported Standard Error of b_j (s_{b_j})	x_j Appears _?_ Significant than it "Really" Is	Some Possible Mistakes: We Might...	Probability of Type _?_ Error is Increased	OLS Regression Assumption Violated:
Multicollinearity (MC) Involving x_j	Unbiased	Larger than if Less MC	Less	Omit Truly Relevant x's	II	*With <u>Perfect</u> MC:* Rank($\mathbf{X}$) = $k+1$ (i.e., $(\mathbf{X}'\mathbf{X})^{-1}$ exists)
Non-Linearity[*] Between y and x_j	Biased and Inconsistent	Larger than if Linear	Less	Misspecify Model; Wrongly Estimate y	I or II	Cov(e, x_j) = E[e x_j] = 0
Omit a Relevant x	Biased and Inconsistent	Biased	More or Less (Either Way)	Misspecify Model; Wrongly Estimate y	I or II	Cov(e, x_j) = E[e x_j] = 0
Include an Irrelevant x	Unbiased	Larger than Otherwise	Less	Omit Truly Relevant x's	II	Optimally Efficient b's (Similar to "High MC")
Measurement Error in y	Unbiased	Larger than Otherwise	Less	Omit Truly Relevant x's	II	No Measurement Error (Optimally Efficient b's)
Measurement Error in x_j	Biased[**] and Inconsistent	Biased	More or Less (Either Way)	Misspecify Model; Wrongly Estimate y	I or II	Cov(e, x_j) = E[e x_j] = 0
Heteroscedasticity: Low Var(e) Close to $\bar{x}_j$	Unbiased	Biased: Usually too Low	More	Include Truly Irrelevant x_j	I	Var(e_i) = E[e_i^2] = σ_ε^2 (a constant) for all i
Heteroscedasticity: High Var(e) Close to $\bar{x}_j$	Unbiased	Biased: Usually too High	Less	Omit Truly Relevant x_j	II	Var(e_i) = E[e_i^2] = σ_ε^2 (a constant) for all i
"+" Autocorrelation in Residuals (e's) and x_j	Unbiased	Biased: Usually too Low	More	Include Truly Irrelevant x_j	I	Cov(e_i, e_j) = E[e_i e_j] = 0 for all $i \neq j$
"-" Autocorrelation in Residuals (e's) and x_j	Unbiased	Biased: Usually too High	Less	Omit Truly Relevant x_j	II	Cov(e_i, e_j) = E[e_i e_j] = 0 for all $i \neq j$

[*] E.g., incorrect functional form for a continuous x_j variable; not interval-level for a categorical x_j variable.

[**] The slope coefficient is biased toward zero (i.e., attenuated) in bivariate regression.

5.5 Models

So we keep talking about the concept of a true population parameter value. But absent a divine intervention, we will never know this true parameter value with 100% certainty ($\alpha = 0$) or perfect precision. So a population parameter does have a true value, but that true value will always be unknown to us.

But we still can, and do, use a parameter's true value as a useful conceptual target when considering how well a sample statistic estimates it. So we can determine how desirable it is to make an inference that this sample statistic value is probably pretty close to the true population parameter value. A sample statistic value will never perfectly tell us a population parameter's true value, but it can do a pretty good job estimating it.

Model: A mathematical expression involving multiple variables that describes relationships involving those variables.

Example: In a regression model, a set of (independent) variables is used to explain or predict another (dependent) variable by estimating its value. We use our sample model to better (not perfectly) reflect and then make inferences about reality. But there is a big difference between population parameters and population models:

Even though a population parameter value will be unknown to us, it does exist and is true and perfect. A parameter value is a mathematically defined and focused stationary target at which we take aim. But there is not a true population model, there is no true set of variables that truly and perfectly describes relationships involving those variables. A model reflects our personal theory, not universal reality. Unlike a parameter, a model is a theoretically influenced fuzzy and moving target that does not truly exist.

We sometimes use the false notion of a true model for pedagogical purposes. As a conceptual tool, we strive to use our sample data to develop a better statistical model but we are really trying to get better estimates of the true population parameter values that are associated with the variables in our personal theory based sample model. A statistical model is a pathway that guides us to a better understanding of the relationships between variables.

6 Residuals

ε_1 is the true error term for case i.
If there is a true, perfect population value, why does the population still have an error term?
It is important to think of the error term not as a mistake but as a deviation, disturbance, stochastic shock.
The observation deviates from the true population due to randomness.

$\hat{\varepsilon}_i$ is the estimated error term for case i, the residual.
The residual is the error term and is the gap between the actual, observed data points in the sample and the predicted point by the line of best fit.

$e_i = y_i - (a + bx_i) = y_i - \hat{y}_i$ where $\hat{y}_i = a + bx_i$. $\hat{y}_i$ is the predicted value of y given the OLS model for case i.

6.1 Expected Value

The expected value: the value that you would expect a random variable to take, based on its distribution.

The expected value can also be thought of as a long-term average: the average value of a random variable if you drew an infinite number of samples from the population.

Example: We had a basket with five balls numbered 1, 1, 2, 3, 3. You reach in and grab one of the balls at random and your choice of grabbing any one of the balls is the same as the probability of selecting any other. Without looking at the ball you selected, what would you expect its value to be?

The answer is the expected value of the ball, also written as E[ball], it is the same as the average of all the balls which would be $\frac{1+1+2+3+3}{5} = 2$. So what does this mean? Sometimes we will draw a ball a bit lower than 2, sometimes a bit higher. But in the long run, if we repeat the drawing process over and over, the average value of our ball drawn will be 2. This is our best guess to answer the question "what is E[ball] on any one draw", which is 2.

Also notice that on any one draw the P(ball = 2) is 1/5, while P(ball = 1) = 2/5. So on any single draw, it is less likely that we will draw exactly a 2, it is more likely we will get a 1 or 3. However, this is not how expected value works, because E[ball] depends on the long-run average.

What about if we have $y = 5 + 7x$?
X = 0 → y = 5
x = 1 → y = 12
x = $\frac{-1}{2}$ → y = $1\frac{1}{2}$

What is E[y]? If we know E[x] = 0?

E[y] = 5 + 7*0 = 5
This means in the long run, given the equation y = 5 + 7x, we expect x = 0, so we expect y = 5

6.2 Model Fit

Model fit: how accurately our model predicts observations

6.2.1 ESS/TSS/RSS

Total Sum of Squares (TSS): Everything there is to explain or all that could be explained by the model
$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2$ or ESS + RSS

Sum of Squared Residuals (RSS): variation we cannot explain or model did not explain
$(Y_i - \hat{Y})^2 + (Y_2 - {}_2)^2 + (\widehat{Y_n} - \hat{Y}_n)^2$

Explained Sum of Squares (ESS): explained variation, all variation explained by y,
ESS = 1-$\frac{RSS}{TSS}$

6.2.2 R-Squared

The coefficient of determination (R^2) measures the proportion of explained variance (i.e., the proportion of variation in y explained by the model). It is the explanatory power of model.

$1 - \frac{RSS}{TSS}$ or $\frac{ESS}{TSS}$.

It is a measure of the linear relationship between the IV and DV, the more linear the higher it is.

It does not measure the strength of the model, only the linear relationship. Remember linearity means that ΔY is the same ("b") across all low, medium, and high values of x.

Problems of R^2:

1. it is sample specific: it depends on sd(y) and sd's of the x's, so we shouldn't compare it across equations, since differences in the characteristics of the samples themselves will pollute R^2

2. it is not a measure of predictive power, not even for a single equation. So models with a higher R^2 are not necessarily better.

Visualization:

7 Multiple Regression

There are multiple variables predicting the outcome

The interpretation of Beta remains the same, but we are now holding all other variables constant

For the OLS equation: $Y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i$
$b_1 =$ predicted Δ Y for a one unit increase in x_1, controlling for or removing the impact of or holding constant x_2

b_2 is the predicted Δ Y for a one unit increase in x_2, controlling for/removing the impact of/holding constant x_1

*Remember Δ = change

Visualization:

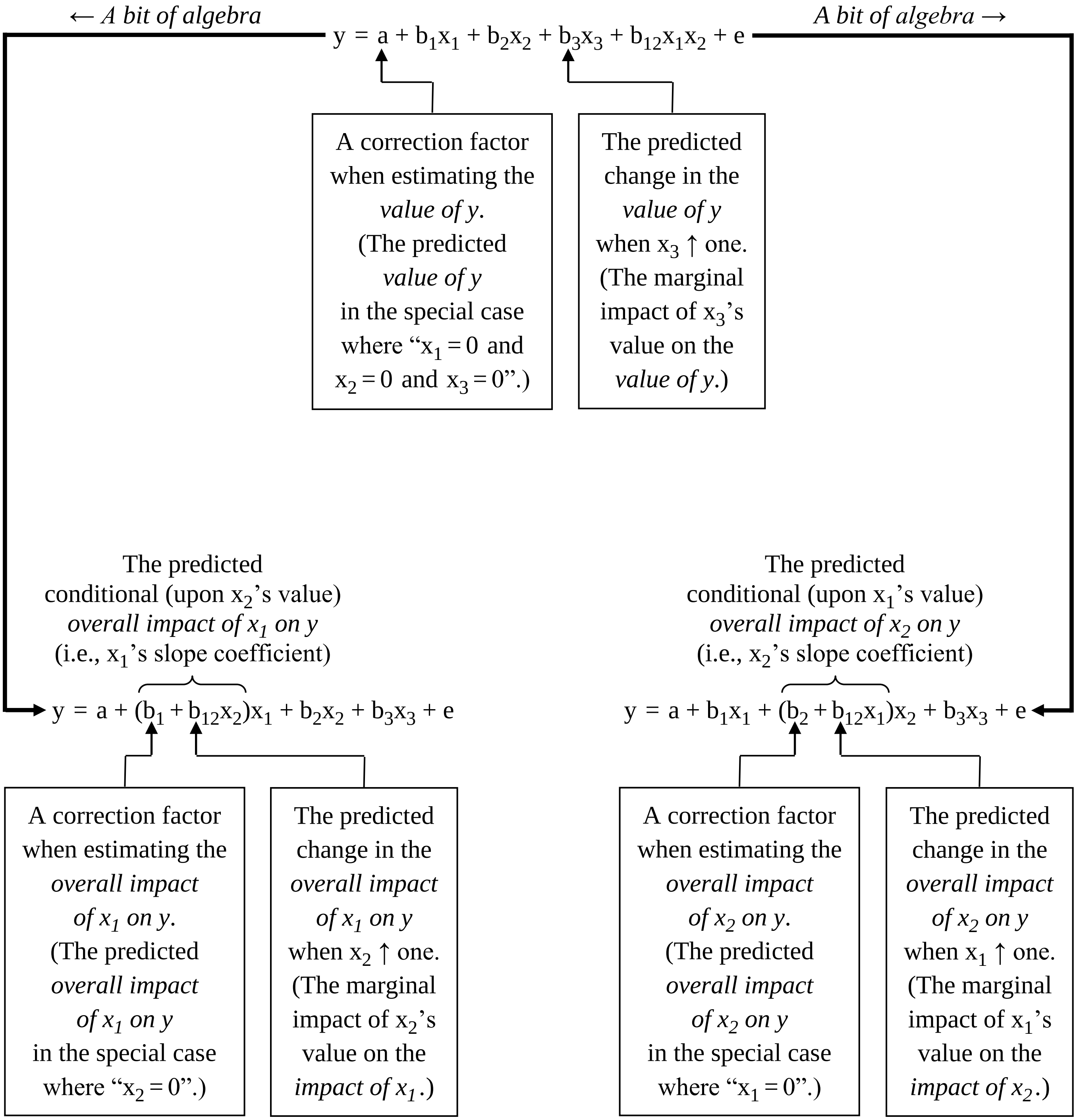
8 Interactions

The interactive effect is a function of all terms (i.e., democracy, GDP, inequality). The interaction term gives the slope of the conditional coefficient.

Interactive Models contain at least one IV that we create by multiplying together 2+ IVs. They are useful for testing theories about how the effects of one IV on a DV that may be contingent on value of another IV.

The Components of an Interaction Model: A Comparative Diagram

Here is a “Picture Language” presentation of what we also see in “English” and “Math” language.



← *A bit of algebra*

A bit of algebra →

$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + e$

A correction factor when estimating the *value of y*. (The predicted *value of y* in the special case where “$x_1 = 0$ and $x_2 = 0$ and $x_3 = 0$”.)

The predicted change in the *value of y* when x_3 ↑ one. (The marginal impact of x_3’s value on the *value of y*.)

The predicted conditional (upon x_2’s value) *overall impact of x_1 on y* (i.e., x_1’s slope coefficient)

$y = a + (\overbrace{b_1 + b_{12}x_2})x_1 + b_2x_2 + b_3x_3 + e$

A correction factor when estimating the *overall impact of x_1 on y*. (The predicted *overall impact of x_1 on y* in the special case where “$x_2 = 0$”.)

The predicted change in the *overall impact of x_1 on y* when x_2 ↑ one. (The marginal impact of x_2’s value on the *impact of x_1*.)

The predicted conditional (upon x_1’s value) *overall impact of x_2 on y* (i.e., x_2’s slope coefficient)

$y = a + b_1x_1 + (\overbrace{b_2 + b_{12}x_1})x_2 + b_3x_3 + e$

A correction factor when estimating the *overall impact of x_2 on y*. (The predicted *overall impact of x_2 on y* in the special case where “$x_1 = 0$”.)

The predicted change in the *overall impact of x_2 on y* when x_1 ↑ one. (The marginal impact of x_1’s value on the *impact of x_2*.)

Note: b_{12} is performing double-duty here. It is the “change in x_1’s slope when x_2 increases by one” AND ALSO the “change in x_2’s slope when x_1 increases by one.” Oops; that is probably not true.

Removing that constraint is methodological and statistical work waiting to be done…perhaps by you!

9 Nonlinearity

Remember how so many of our model fit statistics only work for linear relationships and our assumption of linearity?
What do we do if the relationship is nonlinear?

9.1 Transformations

Theory should drive pretty much everything we do. When we use methodology our goal is to have a model that reflects substantive theory. Functional transformations can help with this.

One of the assumptions we make when using regression is that there is no specification error, one component of this assumption stipulates a linear relationship between IV and DV in our model. Violations of this assumption result in slope coefficients that are biased and inconsistent.

Suppose our theory is that when x increases by one unit, y changes at a rate that is the same across low, medium, and high values of x.

Visualization:

But suppose is that when x increases by one unit then y changes at a rate that is not the same (different) across low, medium and high values of x. Example: When x increases by one unit then y changes by a large amount for low values of x but y changes by smaller and smaller amounts for higher and higher values of x. In this case, a model with an untransformed IV reflects our substantive theory.

Visualization:

But to meet our OLS assumption, we induce linearity by applying a functional transformation to the IV to have the model reflect our theory while avoiding specification error.

The most common transformations are the log, square root and square transformations. These are the simplest transformations, making them the easiest to interpret for your audience.

9.2 Log

ln(x) denotes the natural log
y = a + b(x) + e means that an increase of one unit in x produces an expected change of “b” units in y.

So, grocery bill = a + b (income) + e means a one unit increase in income produces a “b” unit change in expected grocery bill. But we’d expect the grocery bill of a family whose income increased from $5000 to $25000 to go up a whole lot more than a family whose income increased from $10 million to $10,020,000. We expect a $20,000 increase to have a different effect, it will be a big change of the family whose income is low, but small for a family whose income is high.

So we log income, y_i = a + b (log(x_i)) + e_i

Visualization:

Y and log(x) become linearly related so we will not be violating the linearity assumption. But now we have to explain it:

A one unit change in log(x) produces an expected “b” change in y.

Easier way is if x is small, then we do not have to change x by much to get log(x) to change by a unit and E(y) to change by “b”. If x is large, then we have to change x by a substantial amount to get log(x) to change by a unit, and E(y) to change by “b”.

Low values of x: 1 unit increase in x → large change in E(y)

High values of x: 1 unit increase in x → small change in E(y)

When income is small, a one unit increase in income will produce a large change in the expected grocery bill. But when income is large, a one unit increase in income will produce a small change in the expected grocery bill.

9.3 LPM/Logit/Probit

What do we do if we have a dummy variable as our DV in OLS? We can't treat this dummy variable as continuous because it is discrete.

Maximum likelihood: Seeing where the likelihood of DV = 1 is maximized

LPM: If we just use an OLS model, it describes the conditional probabilities but the error violate the homoskedasticity and normality of errors assumptions.

Visualization:

Logistic Regression: Produces a latent index and where we are on the latent index depends on our observed covariates. Only does sign and significance, you have to exponeniate to get the log odds to interpret magnitude. Used when y is a binary variable

Visualization:

Probit regression: 1.8x larger than logit but same, only certain advanced methods can be done on logit, etc.

Can also make out of sample predictions.