3rd class Part I

Let us now praise standard errors!

Pay almost as much attention to standard errors as to coefficients

Why?

1. Standard errors & show how confident you should be about the size of your estimated "effect"

2. Try to make standard errors as small as possible

3. Standard errors as a tool for diagnosing possible multicollinearity problems

4. Standard errors tell you statistical power

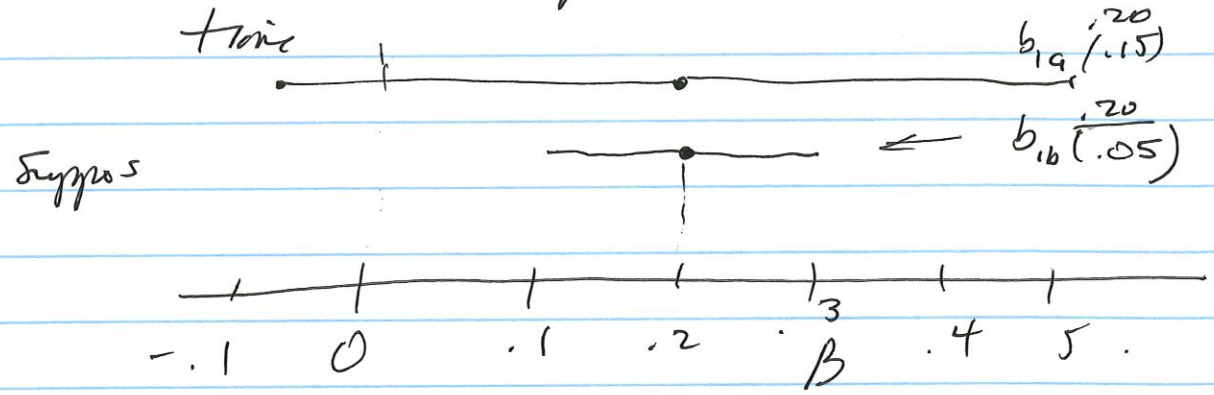
2.8 * s.e. = MDE with 80% power and $p < .05$

Bloom -- later!

1. Confidence intervals

coeffest ± 2 × standard error is ~95% confidence interval

Good interpretation: if we were to take 100 samples, we would expect that the estimate of b_1 to fall with that confidence interval 95% of the time



can be very confident that b_{1a} is different from zero more confident that b_{1b} is different from zero

Q: what is our best guess as to the value of b_1? $b_{1a} = 0$? NO Best guess is

the same $b_{1a} = b_{1b} = .20$

Meta-analysis can take a bunch of $\begin{matrix} .20 \\ (.15) \end{matrix}$ estimates, pool them and deliver ~~an est~~ a much more precise estimate of b_1 (e.g. class size)

How to make standard error as small as possible?

(1) increase sample size
s.e. falls with the square root of the increase in sample size

see if that is the case with handout
first example
Full sample s.e. = .020
1/4 sample s.e. = .040 perfectly as predicted

(2) ~~add~~ increase the R^2 (i.e. reduce the unexplained sum of squares

s.e. = ~~$\sqrt{\frac{\sigma}{n}}$~~ ~~$\sqrt{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}$~~ ~~$\frac{\sqrt{RSS}}{n}$~~

$\frac{\sigma}{\sum(x_i-\bar{x})^2}$ σ ← unexplained variance from regression (Error sum of squares)

⇒ increase the R^2, reduce the standard error

Note: you want predictors that correlate with the dependent variable (and increase R^2)

3rd class Part I

What is fair game to add predictor % outcome %

Exogenous predictors
- allowed!!

Endogenous predictors -- not allowed!!

except in mediation analyses

- invariant demographic characteristics

- temporally preceding measures. (sometimes)

Soak up variance

Go to output

3rd class Part II

Standard errors as tools for observing multicolinearity problems

Tradeoff:

(1) You want to control for a lot of stuff to reduce omitted variable bias

(2) you don't want to introduce ~~student error~~ multicolinearity that blows up standard errors.

Way overblown!!

E.g. Fryer and Levitt

	Full K	no controls	lots of controls
Black		−.663	−.099
		(.025)	(.026)

little change

Handout for today - what happens when you add in a highly correlated predictor

.0076 → .1176

not much!

other, more sophisticated ways of diagnosing, but S.E.'s are trusty ways as well

3rd class Part I

s.e. as guide to power

More later

Is my sample large enough for me to detect a reasonable effect size?

BFY -- power to detect a .20 sd treatment/control difference in Age 3 cognitive test score?

Program to this, but so do data

Fragile Families -- take 1,000 low SES moms and regress Age 3 IQ on child gender

s.e. = .07

.07 * 2.8 = .206 sd!

Bloom: 2.8 * s.e. imputes MDES

Fun with standard errors

```
. /*******************************************************************************
>  Lecture 3 example on standard errors
>  Programmer: Paul Yoo, pyyoo@uci.edu
>  *******************************************************************************/
.
. global project "C:\Users\Jee Hyung Park\Dropbox\UCI Regression class\2021\Labs and
Problem sets\"

. global data "${project}data\"

. global output "${project}output\"

.
. * load data
. use "${data}for_ps6.dta", clear

.
. * identify the variables we'll use for the example and apply a list-wise deletion.
```

The variable list includes an approximately randomly assigned variable - child gender - plus standardized test scores for Spring of Kindergarten (zread1), Spring of 1st grade (zread4) and 5th grade (zread6)

```
.
. // first standardize the reading variables we'll use
. egen zread6 = std(read6)
(10,145 missing values generated)

. egen zread4 = std(read4)
(5,074 missing values generated)

.
. su zread6 zread4 zread1 dfemale

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      zread6 |     11,265   -2.04e-10           1  -3.235986   2.012693
      zread4 |     16,336   -1.30e-11           1  -2.208911   4.469528
      zread1 |     17,622   -1.61e-10           1  -1.392766   10.12822
     dfemale |     21,396    .4882221    .4998729          0          1

. keep if !mi(zread6, zread4, zread1, zread2, dfemale )
(12,345 observations deleted)

. su zread6 zread4 zread1 dfemale

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      zread6 |      9,065    .0979408    .9649163  -3.215906   2.012693
      zread4 |      9,065    .0979004     .977377  -2.171627   4.469528
      zread1 |      9,065     .061892    .9888895  -1.386883   10.12822
     dfemale |      9,065     .501048    .5000265          0          1
```

and here's the correlation matrix for the test score measures:

```
. corr zread6 zread4 zread2
(obs=9,065)

             |   zread6   zread4   zread2
-------------+---------------------------
      zread6 |   1.0000
      zread4 |   0.6770   1.0000
      zread2 |   0.5318   0.7608   1.0000
.
```

Start with a simple regression of 5th grade test scores on child gender. Concentrate on the standard error.

```
. reg zread6 dfemale

      Source |       SS           df       MS      Number of obs   =     9,065
-------------+----------------------------------   F(1, 9063)      =     39.65
       Model |  36.7626256         1  36.7626256   Prob > F        =    0.0000
    Residual |  8402.39646     9,063  .927109838   R-squared       =    0.0044
-------------+----------------------------------   Adj R-squared   =    0.0042
       Total |  8439.15909     9,064  .931063447   Root MSE        =    .96287

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .1273651   .0202261     6.30   0.000     .0877173    .1670128
       _cons |   .0341248    .014317     2.38   0.017     .0060603    .0621894
------------------------------------------------------------------------------
```

Do standard errors change systematically with sample size? In particular do they change with the square root of changes in sample sizes? Let's explore this by throwing out a random ¾ of the data:

```
.
. gen random = runiform()

. gen quarter = random < = 0.25

. tab quarter

    quarter |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      6,787       74.87       74.87
          1 |      2,278       25.13      100.00
------------+-----------------------------------
      Total |      9,065      100.00
```

.

Repeat our first full sample regression and note the sample size and standard error:

```
. reg zread6 dfemale

      Source |       SS           df       MS      Number of obs   =     9,065
-------------+----------------------------------   F(1, 9063)      =     39.65
       Model |  36.7626256         1  36.7626256   Prob > F        =    0.0000
    Residual |  8402.39646     9,063  .927109838   R-squared       =    0.0044
-------------+----------------------------------   Adj R-squared   =    0.0042
       Total |  8439.15909     9,064  .931063447   Root MSE        =    .96287

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .1273651   .0202261     6.30   0.000     .0877173    .1670128
       _cons |   .0341248   .014317      2.38   0.017     .0060603    .0621894
------------------------------------------------------------------------------
```

Now run the same regression on the ¼ subsample. If standard errors change with the square root of the change in sample sizes, then a random ¼ of the sample should produce standard errors that are doubled in size:

```
. reg zread6 dfemale if quarter == 1

      Source |       SS           df       MS      Number of obs   =     2,278
-------------+----------------------------------   F(1, 2276)      =      2.11
       Model |  1.95013154         1  1.95013154   Prob > F        =    0.1465
    Residual |  2103.27394     2,276  .924109815   R-squared       =    0.0009
-------------+----------------------------------   Adj R-squared   =    0.0005
       Total |  2105.22407     2,277  .924560418   Root MSE        =    .96131

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .0585201   .0402342     1.45   0.146    -.0204775    .1375178
       _cons |   .0828665   .0283474     2.92   0.003     .0272771    .1384559
------------------------------------------------------------------------------
```

Voila!

The standard error should drop if we add in a covariate that is highly correlated with the dependent variable (in other words, that reduces the residual sum of squares). Let's add in the 1st grade reading score, ignore the coefficient and concentrate on the standard error of dfemale.

First repeat the original regression:

```
. reg zread6 dfemale

      Source |       SS           df       MS      Number of obs   =     9,065
-------------+----------------------------------   F(1, 9063)      =     39.65
       Model |  36.7626256         1  36.7626256   Prob > F        =    0.0000
    Residual |  8402.39646     9,063  .927109838   R-squared       =    0.0044
-------------+----------------------------------   Adj R-squared   =    0.0042
       Total |  8439.15909     9,064  .931063447   Root MSE        =    .96287

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .1273651   .0202261     6.30   0.000     .0877173    .1670128
       _cons |   .0341248    .014317     2.38   0.017     .0060603    .0621894
------------------------------------------------------------------------------
```

How add in an independent variable that is strongly correlated with the dependent variable in order to reduce the unexplained sum of squares:

```
. reg zread6 dfemale zread4

      Source |       SS           df       MS      Number of obs   =     9,065
-------------+----------------------------------   F(2, 9062)      =  3835.66
       Model |  3868.89866         2  1934.44933   Prob > F        =    0.0000
    Residual |  4570.26043     9,062  .504332425   R-squared       =    0.4584
-------------+----------------------------------   Adj R-squared   =    0.4583
       Total |  8439.15909     9,064  .931063447   Root MSE        =    .71016

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .0260252   .0149631     1.74   0.082    -.0033057    .0553562
      zread4 |   .6672878   .0076551    87.17   0.000     .6522821    .6822935
       _cons |   .0195732   .0105609     1.85   0.064    -.0011285    .0402749
------------------------------------------------------------------------------
```

A separate issue is about how standard errors can provide an early warning that your estimation equation may suffer from excessive multicollinearity. So the question is: does adding in a highly correlated variable cause trouble? Standard error changes are a very useful indicator of trouble - do standard errors blow up when you add in the extra predictors?

To generate possible conditions of multicollinearity, let's add in another test score - one from kindergarten. The model doesn't make sense substantively, but it does illustrate how standard errors change in the presence of correlated variables.

Remember that zread2 zread4 and zread6 are highly correlated:

```
.
. corr zread6 zread4 zread2
(obs=9,065)

             |   zread6   zread4   zread2
-------------+---------------------------
      zread6 |   1.0000
      zread4 |   0.6770   1.0000
      zread2 |   0.5318   0.7608   1.0000
```

Now see what happens to the standard error on zread4 when you add in zread2.

First repeat the original regression:

```
. reg zread6 dfemale zread4

      Source |       SS           df       MS      Number of obs   =     9,065
-------------+----------------------------------   F(2, 9062)      =   3835.66
       Model |  3868.89866         2  1934.44933   Prob > F        =    0.0000
    Residual |  4570.26043     9,062  .504332425   R-squared       =    0.4584
-------------+----------------------------------   Adj R-squared   =    0.4583
       Total |  8439.15909     9,064  .931063447   Root MSE        =    .71016

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .0260252   .0149631     1.74   0.082    -.0033057    .0553562
      zread4 |   .6672878   .0076551    87.17   0.000     .6522821    .6822935
       _cons |   .0195732   .0105609     1.85   0.064    -.0011285    .0402749
------------------------------------------------------------------------------
```

Now add in the additional correlated predictor:

```
. reg zread6 dfemale zread4 zread2

      Source |       SS           df       MS      Number of obs   =     9,065
-------------+----------------------------------   F(3, 9061)      =   2563.55
       Model |  3874.40394         3  1291.46798   Prob > F        =    0.0000
    Residual |  4564.75515     9,061  .503780504   R-squared       =    0.4591
-------------+----------------------------------   Adj R-squared   =    0.4589
       Total |  8439.15909     9,064  .931063447   Root MSE        =    .70977

------------------------------------------------------------------------------
      zread6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dfemale |   .0247772   .0149596     1.66   0.098    -.0045471    .0541014
      zread4 |   .6377669   .0117594    54.23   0.000     .6147158    .6608181
      zread2 |   .0386455   .0116904     3.31   0.001     .0157297    .0615614
       _cons |   .0204129   .0105581     1.93   0.053    -.0002834    .0411093
------------------------------------------------------------------------------
```

⇨ **~50% increase. An increase to be sure, but not a doubling, tripling or worse increase that indicate real trouble. More generally, adding in a bunch of theoretically-appropriate control variables has the benefit of reducing omitted-variable bias and rarely (but not always!) causes multicolinearity problems.**

```
.

.
end of do-file

.
```

Education 265: Applied Regression Analysis
Winter, 2021

Ways of calculating and communicating about regression coefficients

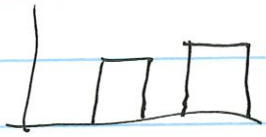
[Coefficients are always about $\frac{\Delta Y}{\Delta X}$, the change in Y associated with a one-unit change in X, which is the slope of the regression line you are estimating]

Name	Comments	Definition
Raw score ("unstandardized")	Both X and Y are in natural units	$\Delta Y / \Delta X$, where ΔX is a one raw unit change in X
Effect size – in experiments.	Y is standardized by dividing natural units by the standard deviation of Y. X is a (0,1) indicator of whether in the treatment group. X is kept in its "natural units" of 0 and 1.	$\frac{\Delta Y}{\sigma_Y} / T$, where T is (1,0) indicator of treatment status
Standardized coefficients ("effect sizes", β, "Beta weight")	Both X and Y are in standard deviation units. In a bivariate regression, β^2 = the explained variance (R^2) of the regression, so βs are taken to indicate relative (explanatory) importance. In this important sense, βs are comparable indicators of relative explanatory power across independent variables in an equation.	$\frac{\frac{\Delta Y}{\sigma_Y}}{\frac{\Delta X}{\sigma_X}}$ *(note that this is the same as: $\frac{\Delta Y}{\Delta X} \cdot \frac{\sigma_X}{\sigma_Y}$, where $\frac{\Delta Y}{\Delta X}$ is the raw score coefficient)*
Percent change coefficients	An advantage of percentage change is that it is units-free. Percentage changes are usually calculated from the mean of Y in the sample, but you can plug in other values if they are more meaningful. In the equation: $ln\ Y = a + b_1 X$, b_1 has a percentage change interpretation because Δ ln Y approximates percentage change in Y for small changes in X.	$\frac{\Delta Y}{\bar{Y}} / \Delta X$
Elasticity coefficients	In this case, both the numerator and denominator are expressed as percentage change. In the equation: $ln\ Y = a + b_1\ ln\ X$, b_1 has an elasticity interpretation.	$\frac{\frac{\Delta Y}{\bar{Y}}}{\frac{\Delta X}{\bar{X}}}$ *(note that this is the same as: $\frac{\Delta Y}{\Delta X} \cdot \frac{\bar{X}}{\bar{Y}}$, where $\frac{\Delta Y}{\Delta X}$ is the raw score coefficient)*
Months of learning (or development) interpretation	This is really an interpretation of the first three coefficients and often helps readers understand the magnitude of coefficients. Suppose, for example, a tutoring program boosted reading scores of second graders by .30 sd. Hill et al. (2008) show that the average gain in reading scores across second grade is .60sd, so the effect of the tutoring program translates into a half-year of additional learning. Percentiles of learning from a normed assessment are an alternative way of doing this.	
Benefit-cost coefficients	This rarely-used coefficient wins the "best intentions" award from me because it calculates the impacts of, say, a $1,000 expenditure on X on some outcome of interest. It best serves policy maker needs because it shows (educational) bank for the buck.	

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

2nd class Part II

Effect sizes for 100



See cheat sheet

1. Raw score ("unstandardized") ~~Hedges~~ Both X & Y in natural units $\frac{\Delta y}{\Delta x} = \frac{10}{1} = 10$ [.14 years per *10K ΔY]

2. "Effect size" in experiments Y is standardized X in natural units (usually 1,0) $\frac{.67}{1} = .67$ sd

3. Standardized coefficients (also called "effect sizes", β) Y and X are standardized

4. Percent change in Y Y is % change relative to the mean $\frac{\frac{\Delta y}{y}}{\Delta x} = 10\%$

5. Elasticity (% chg in both X & Y) Y & X are expressed as % change

6. months of learning Express ΔY as a months of learning

47 Benefit/cost ratio

$$\frac{\text{\$ value of change in } y}{\text{\$ value of cost of changing } X}$$

$$\frac{\text{\$ value of a .67sd increase in } Y}{\text{cost of intervention}}$$

covid paper / graph

2nd class Part II

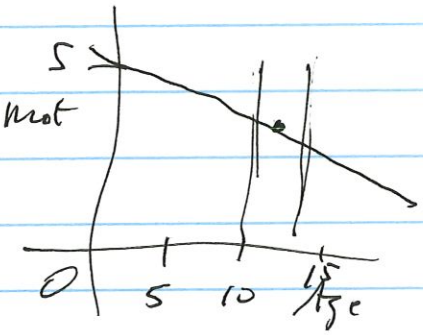
Suppose Motivation = a + b, Age for middle schoolers

	Mean	SD
Age in years	12.58	.64
Mot 1-5 scale	3.61	.61

in fact $\widehat{Mot} = 5.25 - .14\ Age$
$(.38)\quad(.03)$

⇒ motivation declines across middle school

First of all 5.25? Mot at age 0



Focus on -.14 big or small?
⇒ raw score coefficients are often difficult to interpret

In fact if Age was scaled in months rather than years

$$\widehat{Mot} = 5.25 - \frac{.14/12}{.03/12} = -.012\ Age \quad (.003)$$

① Standing numerator $\frac{-.14}{.61} = .23$ sd per year of age

② Standardized coefficients "β" "Beta weights"

⇒ divide -.14 instead of $\frac{\Delta Y}{\Delta X}$, how $\frac{\frac{\Delta Y}{\sigma_y}}{\frac{\Delta X}{\sigma_x}}$

in our example $\dfrac{\frac{\Delta Y}{\sigma_Y} \; \frac{.14}{.61}}{\frac{\Delta X}{\sigma_X} \; \frac{1}{.64}} = -.14 \times \frac{.64}{.61} = -.15$

First

$$\frac{\frac{\Delta Y}{\sigma_y}}{\frac{\Delta X}{\sigma_x}} = \frac{\Delta Y}{\Delta X} \cdot \frac{\sigma_x}{\sigma_y} = \text{raw score coeff} \times \frac{\sigma_x}{\sigma_y}$$

Virtues of standardized coffits

1. Facilitates comparisons across coffits in a regression β = "relative importance"

in bivariate $Y = a + \beta X$, $\beta^2 = R^2$

in $Y = a + \beta_1 X_1 + \beta_2 X_2$ (and X's are uncorrelated)

$$\beta_1^2 + \beta_2^2 + \dots = R^2$$

Problems with fully standards copper.

(1) Sometimes raw score changes are much easier to think about (1,0)

$$Cog = a + b_1 Treatment$$

b_1 ↑ but not this

↑ Studying this

SD of a 1,0 variable is $\sqrt{p(1-p)}$

(2) What S.d. to use?

Abecedarian had high by inputs on IQ at age 19

.40sd using the sample S.d.

but <.30 sd using national norms.

4 pts $\sigma = 10$ ⇒ .40sd

4pt $\sigma = 15$ ⇒ .28 sd

Months of learning

Example of covid survey - question.

Percentile of learning

2nd class Part II

Percent change

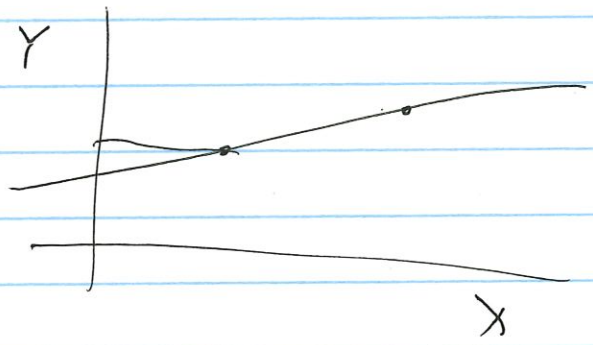
mean Y

$-.14 / 3.61$

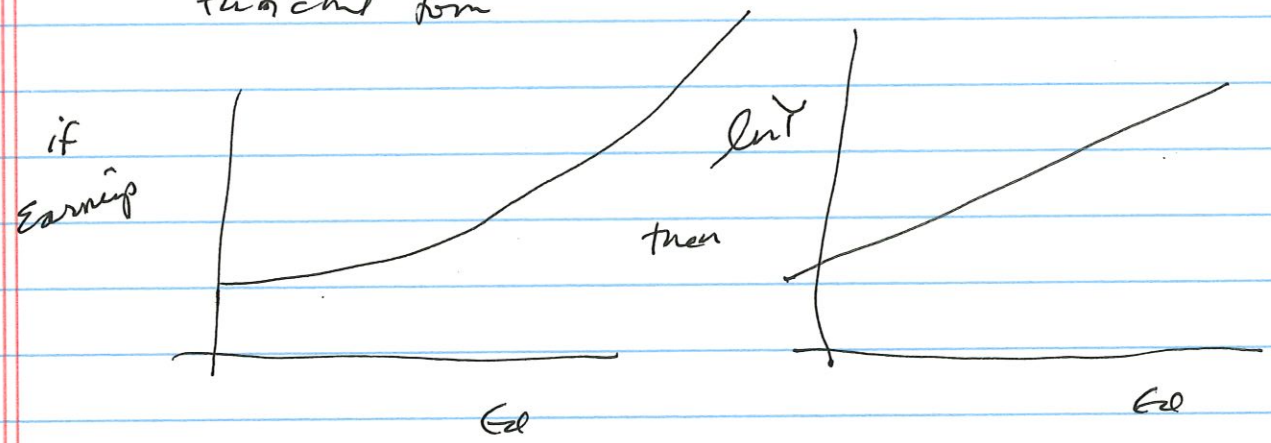
if $\Delta Y / \Delta X$ = 10 and ~~X=2~~ 3.82 per year

then 20% increase associated with a one unit change in X

Base doesn't have to be $\bar{Y}$, it could be some other Y



Percent change interpretation also comes from using a log Y functional form



$\ln Y = a + b_1 Ed$. Going from Ed_1 to $Ed_1 + 2$

takes you from $\ln Y_1$ to $\ln Y_2$

$\Delta \ln Y = \ln Y_2 - \ln Y_1$

$= \ln\left(\frac{Y_2}{Y_1}\right) = \frac{Y_2}{Y_1} - 1$ for small chgs

b_1 has the interpretation of % change if b_1 is small

Achievement Gaps in the Wake of COVID-19 (Paper here)

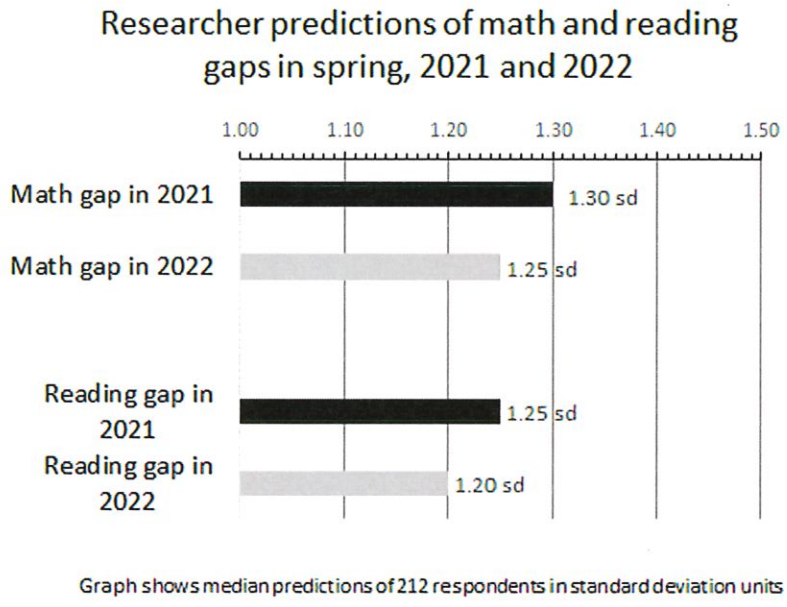
Drew Bailey, Greg J. Duncan, Richard J. Murnane, and Natalie Au Yeung

Appendix 1: Qualtrics Survey

Suppose that, in early 2020 before the pandemic hit, the achievement gap on a NAEP-type **math** test for children attending elementary school was +1.00 standard deviations when children in the top income quintile are compared with children in the bottom income quintile (in other words, roughly what Sean Reardon and others have found).

1. Suppose that those same children were all somehow able to take comparable math achievement tests **this coming spring (i.e., Spring, 2021)**. What is your best estimate of the gap estimated from those data?

[ ] *sd*



The median forecast for the increase in the gap in math achievement in elementary school was a change from 1.00 to 1.30 standard deviations – *fill in the best interpretation!*

3rd class Part III

Dummy variables
~~Effect size~~

1. Equivalence of ANOVA & dummy variable regression

go over hand out

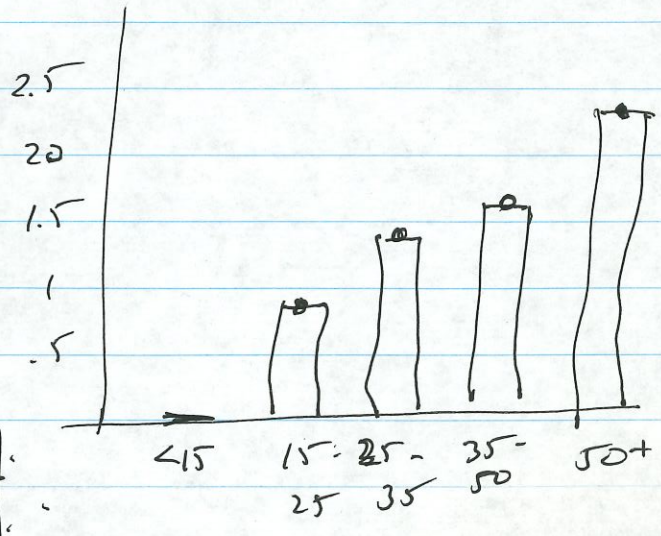
Why bother with regression?
- adjust for covariates

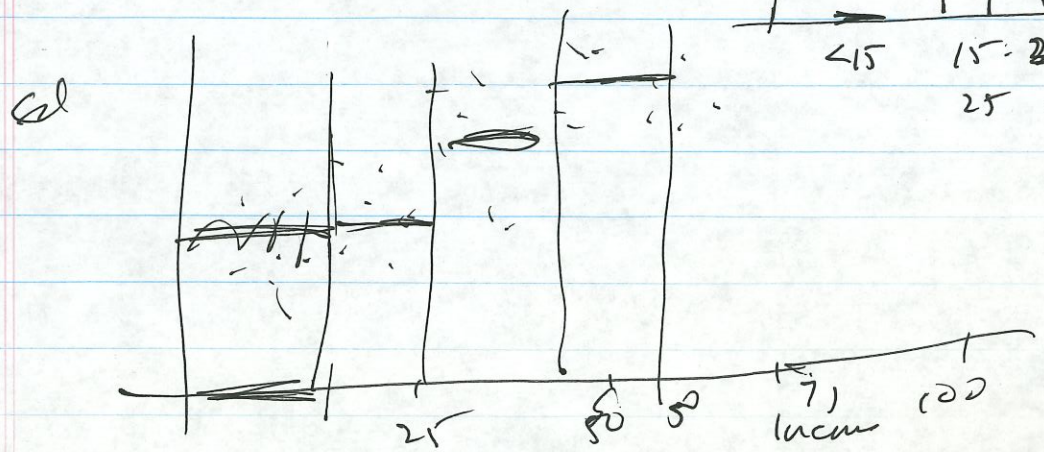
2. Flexible way of examining relationship between X and Y

Table 3 from ASR article

Completed Education

omitted	0
15-25	.82
25-35	1.41
35-50	1.69
50+	2.35





Regression and ANOVA

(Educ 265 week 3)

Table 2. Estimated Racial Achievement Gap over the First Four Years of School, Math

	(1) Fall-K	(2) Spring-K	(3) Spring-1st	(4) Spring-3rd	(5) Fall-K	(6) Spring-K	(7) Spring-1st	(8) Spring-3rd
Black	−0.663 (0.025)	−0.724 (0.027)	−0.758 (0.029)	−0.882 (0.031)	−0.099 (0.026)	−0.209 (0.028)	−0.279 (0.031)	−0.382 (0.033)
Hispanic	−0.738 (0.024)	−0.681 (0.025)	−0.568 (0.026)	−0.539 (0.026)	−0.197 (0.024)	−0.189 (0.026)	−0.122 (0.027)	−0.078 (0.028)
Asian	0.11 (0.059)	0.088 (0.056)	−0.025 (0.052)	0.066 (0.054)	0.258 (0.050)	0.226 (0.050)	0.092 (0.047)	0.163 (0.049)
Other race	−0.495 (0.047)	−0.481 (0.048)	−0.497 (0.050)	−0.541 (0.050)	−0.158 (0.040)	−0.175 (0.043)	−0.21 (0.046)	−0.244 (0.046)
Age (in months)					0.058 (0.002)	0.053 (0.002)	0.037 (0.002)	0.019 (0.002)
Birth weight (in ounces)					0.003 (0.000)	0.003 (0.000)	0.003 (0.000)	0.003 (0.000)
Female					0.005 (0.017)	−0.005 (0.017)	−0.044 (0.018)	−0.175 (0.018)
Number of children's books					0.006 (0.001)	0.006 (0.001)	0.005 (0.001)	0.006 (0.001)
Number of children's books (squared) (×1000)					−0.021 (0.002)	−0.02 (0.003)	−0.019 (0.003)	−0.020 (0.003)
Mother over 30 at first birth					0.165 (0.026)	0.107 (0.025)	0.086 (0.027)	0.083 (0.024)
Socioeconomic status measure					0.306 (0.016)	0.282 (0.015)	0.256 (0.015)	0.208 (0.015)
Mother receives Wic benefits					−0.212 (0.021)	−0.191 (0.022)	−0.19 (0.023)	−0.208 (0.024)
Mother a teenager at first birth					−0.114 (0.021)	−0.118 (0.022)	−0.131 (0.025)	−0.132 (0.025)
Constant	0.307 (0.013)	0.304 (0.013)	0.266 (0.012)	0.275 (0.012)	−4.357 (0.154)	−3.952 (0.160)	−2.795 (0.168)	−1.576 (0.168)
Observations	11201	11201	11201	11201	11201	11201	11201	11201
R-squared	0.11	0.11	0.1	0.12	0.32	0.29	0.24	0.26

Notes. The dependent variable is the math test score at various points in a student's career. Test scores are IRT scores, normalized to have a mean of 0 and a standard deviation of 1 in the full, unweighted sample. Non-Hispanic Whites are the omitted race category, so all of the race coefficients are relative to that group. The unit of observation is a student. Standard errors in parentheses. Estimation is done using weighted least squares, using sample weights provided in the data set.

(Almost identical) Regression

```
. reg zmath1 dblack dhispanic dasian dother if replicate_sample==1
```

Source	SS	df	MS	Number of obs	=	10,605
				F(4, 10600)	=	288.00
Model	1094.62228	4	273.655571	Prob > F	=	0.0000
Residual	10072.1599	10,600	.950203761	R-squared	=	0.0980
				Adj R-squared	=	0.0977
Total	11166.7821	10,604	1.05307263	Root MSE	=	.97478

zmath1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dblack	-.6707628	.0304056	-22.06	0.000	-.7303636	-.6111621
dhispanic	-.7139294	.0257034	-27.78	0.000	-.7643128	-.663546
dasian	.1213123	.047709	2.54	0.011	.0277937	.2148309
dother	-.4451654	.0425755	-10.46	0.000	-.5286213	-.3617094
_cons	.3325539	.0120684	27.56	0.000	.3088975	.3562103

ANOVA

```
. oneway zmath1 race_simp if replicate_sample == 1, tab
```

Race, simplified	Summary of (Standardized) Math-Kindergarten Fall Mean	Std. Dev.	Freq.
White	.33255388	1.0347184	6,524
Black	-.33828897	.73177843	1,220
Hispanic	-.38137551	.81640887	1,845
Asian	.45386615	1.1896878	446
Other	-.1126115	1.0067506	570
Total	.11235875	1.0261933	10,605

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	1094.62228	4	273.655571	288.00	0.0000
Within groups	10072.1599	10600	.950203761		
Total	11166.7821	10604	1.05307263		

Bartlett's test for equal variances: chi2(4) = 360.3519 Prob>chi2 = 0.000

Regression v. ANOVA

Difference is .67076287

```
. reg zmath1 dblack dhispanic dasian dother if replicate_sample==1
```

Source	SS	df	MS	Number of obs	=	10,605
				F(4, 10600)	=	288.00
Model	1094.62228	4	273.655571	Prob > F	=	0.0000
Residual	10072.1599	10,600	.950203761	R-squared	=	0.0980
				Adj R-squared	=	0.0977
Total	11166.7821	10,604	1.05307263	Root MSE	=	.97478

zmath1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dblack	-.6707628	.0304056	-22.06	0.000	-.7303636	-.6111621
dhispanic	-.7139294	.0257034	-27.78	0.000	-.7643128	-.663546
dasian	.1213123	.047709	2.54	0.011	.0277937	.2148309
dother	-.4451654	.0425755	-10.46	0.000	-.5286213	-.3617094
_cons	.3325539	.0120684	27.56	0.000	.3088975	.3562103

Should be White mean

```
. oneway zmath1 race_simp if replicate_sample == 1, tab
```

Race, simplified	Summary of (Standardized) Math-Kindergarten Fall Mean	Std. Dev.	Freq.
White	.33255388	1.0347184	6,524
Black	-.33828897	.73177843	1,220
Hispanic	-.38137551	.81640887	1,845
Asian	.45386615	1.1896878	446
Other	-.1126115	1.0067506	570
Total	.11235875	1.0261933	10,605

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	1094.62228	4	273.655571	288.00	0.0000
Within groups	10072.1599	10600	.950203761		
Total	11166.7821	10604	1.05307263		

Bartlett's test for equal variances: chi2(4) = 360.3519 Prob>chi2 = 0.000

Different reference/omitted group

```
. reg zmath1 dwhite dhispanic dasian dother if replicate_sample==1
```

Source	SS	df	MS	Number of obs	=	10,605
				F(4, 10600)	=	288.00
Model	1094.62228	4	273.655571	Prob > F	=	0.0000
Residual	10072.1599	10,600	.950203761	R-squared	=	0.0980
				Adj R-squared	=	0.0977
Total	11166.7821	10,604	1.05307263	Root MSE	=	.97478

zmath1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dwhite	.6707628	.0304056	22.06	0.000	.6111621	.7303636
dhispanic	-.0431665	.0359704	-1.20	0.230	-.1136753	.0273422
dasian	.7920751	.0539305	14.68	0.000	.6863456	.8978046
dother	.2255975	.0494558	4.56	0.000	.1286547	.3225402
_cons	-.338209	.027908	-12.12	0.000	-.3929139	-.2835041

Using the "test" function

Difference is .0431666

```
. reg zmathl dwhite dhispanic dasian dother if replicate_sample==1

      Source |       SS       df       MS              Number of obs =  10,605
-------------+------------------------------           F(4, 10600)   =  288.00
       Model |  1094.62228     4  273.655571            Prob > F      =  0.0000
    Residual |  10072.1599 10,600 .950203761            R-squared     =  0.0980
-------------+------------------------------           Adj R-squared =  0.0977
       Total |  11166.7821 10,604 1.05307263            Root MSE      =  .97478

------------------------------------------------------------------------------
      zmathl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dwhite |   .6787628   .0304056    22.06   0.000     .6191621    .7383636
   dhispanic |  -.0431665   .0359704    -1.20   0.230    -.1136755    .0273423
      dasian |   .7939751   .0539385    14.68   0.000     .6882456    .8997046
      dother |   .2255975   .0494558     4.56   0.000     .1286547    .3225402
       _cons |  -.3382109   .027908   -12.12   0.000    -.3929139   -.2835081
------------------------------------------------------------------------------

. reg zmathl dblack dhispanic dasian dother if replicate_sample==1

      Source |       SS       df       MS              Number of obs =  10,605
-------------+------------------------------           F(4, 10600)   =  288.00
       Model |  1094.62228     4  273.655571            Prob > F      =  0.0000
    Residual |  10072.1599 10,600 .950203761            R-squared     =  0.0980
-------------+------------------------------           Adj R-squared =  0.0977
       Total |  11166.7821 10,604 1.05307263            Root MSE      =  .97478

------------------------------------------------------------------------------
      zmathl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dblack |  -.6787628   .0304056   -22.06   0.000    -.7383636   -.6191621
   dhispanic |  -.7139394   .0257034   -27.78   0.000    -.7643128   -.663566
      dasian |   .1152137   .047769     2.54   0.011     .0277937    .2026337
      dother |  -.4531654   .0425755   -10.46   0.000    -.5380213   -.3683094
       _cons |   .3325539   .0120684    27.56   0.000     .3088975    .3562103
------------------------------------------------------------------------------

. test dblack = dhispanic

 ( 1)  dblack - dhispanic = 0

       F(  1, 10600) =    1.44
            Prob > F =    0.2301
```

Fun fact: $F = t^2$

414 AMERICAN SOCIOLOGICAL REVIEW

Table 3. Coefficients from the Regression of Child's Outcome Variables on Family Income at Ages 0 to 15: Panel Study of Income Dynamics

	Dependent Variables/Models											
	Years of Completed Schooling[a]				High School Completion[b]				Hazard of Nonmarital Birth[c]			
Independent Variable	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Family Income at Child's Ages 0 to 15												
Linear function	.14* (.02)	—	—	—	.23* (.07)	—	—	—	−.43* (.10)	—	—	—
Spline function												
Income < $20,000	—	1.30* (.29)	—	—	—	1.97* (.44)	—	—	—	−.50 (.41)	—	—
Difference between income < $20,000 and > $20,000	—	−1.17* (.30)	—	—	—	−1.84* (.46)	—	—	—	−.08 (.44)	—	—
Natural logarithm	—	—	1.16* (.11)	—	—	—	1.35* (.26)	—	—	—	−1.18* (.26)	—
Dummy Variables for Family Income												
$15,000 to $24,999	—	—	—	.82* (.27)	—	—	—	1.41* (.38)	—	—	—	−.54 (.35)
$25,000 to $34,999	—	—	—	1.41* (.28)	—	—	—	1.83* (.43)	—	—	—	−.94 (.41)
$35,000 to $49,999	—	—	—	1.69* (.28)	—	—	—	2.48* (.45)	—	—	—	−1.44* (.43)
$50,000 and over	—	—	—	2.35* (.29)	—	—	—	2.64* (.49)	—	—	—	−2.40* (.54)
Adjusted R^2	.192	.201	.219	.216	—	—	—	—	—	—	—	—
−2 Log likelihood	—	—	—	—	718.9	702.6	701.1	694.6	1,266.1	1,266.1	1,271.1	1,267.3

Table 3. Coefficients from the Regression of Child's Outcome Variables on Family Income at Ages 0 to 15: Panel Study of Income Dynamics

	Dependent Variables/Models											
	Years of Completed Schooling[a]				High School Completion[b]				Hazard of Nonmarital Birth[c]			
Independent Variable	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Family Income at Child's Ages 0 to 15												
Linear function	.14* (.02)	—	—	—	.23* (.07)	—	—	—	−.43* (.10)	—	—	—
Spline function												
Income < $20,000	—	1.30* (.29)	—	—	—	1.97* (.44)	—	—	—	−.50 (.41)	—	—
Difference between income < $20,000 and > $20,000	—	−1.17* (.30)	—	—	—	−1.84* (.46)	—	—	—	−.08 (.44)	—	—
Natural logarithm	—	—	1.16* (.11)	—	—	—	1.35* (.26)	—	—	—	−1.18* (.26)	—
Dummy Variables for Family Income												
$15,000 to $24,999	—	—	—	.82* (.27)	—	—	—	1.41* (.38)	—	—	—	−.54 (.35)
$25,000 to $34,999	—	—	—	1.41* (.28)	—	—	—	1.83* (.43)	—	—	—	−.94 (.41)
$35,000 to $49,999	—	—	—	1.69* (.28)	—	—	—	2.48* (.45)	—	—	—	−1.44* (.43)
$50,000 and over	—	—	—	2.35* (.29)	—	—	—	2.64* (.49)	—	—	—	−2.40* (.54)
Adjusted R^2	.192	.201	.219	.216	—	—	—	—	—	—	—	—
−2 Log likelihood	—	—	—	—	718.9	702.6	701.1	694.6	1,266.1	1,266.1	1,271.1	1,267.3

Note: Numbers in parentheses are standard errors. In Model 4, the omitted category for family income is "less than $15,000." The mean years of schooling completed was 13.5 (S.D. = 2.1); the mean rate of high school completion was .90 (S.D. = .30).

[a] OLS models; N = 1,323.

[b] Logistic models; N = 1,323.

[c] Cox models; N = 620.

*$p < .05$ (two-tailed tests)