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 are a tool for diagnosing possible multicollinearity problems

4. Standard errors tell you about statistical power
   \[ 2.8 \times \text{s.e.} = \text{MSE} \times \text{80\% power} \]
   and \( p < 0.05 \)

Bloom -- later!


3rd class Part I

1. Confidence intervals

\[ \text{confidence interval} \pm 2 \times \text{standard error} \]

Confidence interval

**Note:** if we were to take 100 samples, we would expect that the estimate of \( b_1 \) to fall into the confidence interval 95% of the time.

\[ b_{19} (1.15) \]

\[ b_{18} (1.05) \]

50% probability

\[ -1 \quad 0 \quad .1 \quad .2 \quad .3 \quad .4 \quad .5 \]

Can't be very confident that \( b_{1a} \) is different from 0.

Most confident that \( b_{1b} \) is different from 0.

Q: What is our best guess as to the value of \( b_1 \)?

\[ b_1 \neq 0 \rightarrow \text{No best guess} \]

Two aspects:

- \( b_{1a} = 0.4 \)
- \( b_{1b} = 0.20 \)

Regression analysis can take a bunch of \( b_{1b} \) estimates

Pool them and deliver on with a much more precise estimate of \( b_1 \) (e.g., class size)
How to make samples even as small as possible?

1. Minimise sample size
   S.E. falls with the square root of the sample size.

   See if this is the case with handout next sample.
   Full sample S.E. = 0.20
   1/4 sample S.E. = 0.40 perfectly on predicted

2. Add more to $R^2$ (i.e. reduce the unexplained sum of squares)

   $s.e. = \sqrt{\frac{\text{unexplained variance from}}{\text{sum of squares}}} = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$

   To increase $R^2$, 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(s) of forecast? (custom content)

Exogenous predictors
- allowed!!
- invariant demographic characteristics
- temporally precisely measured (surveys)

Endogenous predictors -- not allowed!! except in mediating analyses

Go to output
Student uses ex to test for 
observing multicollinearity

Traditional:

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

2. You don't want to introduce new error that blows up standard errors.

Way overblown!!

E.g., Fryer at Harv.

Fall K = 0.0663
Black = 0.099

\( t = 0.025 \)

Little change

Harvart K today - what happens when you add in a highly correlated predictor

0.076 \( \rightarrow \) 0.176

Not much!

Other, more sophisticated ways of diagnosing

But S.E.'s are trusty ways as well.
3rd class Part I

s.e. as guile to power

was later

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

BEY -- prob to detect a .23 sd drift/cuts

different in & Age 3 cognitive test score?

Pregno be this, but so do data

Fragile foils -- take 1,000 lm secs mins

and regen Age 3 IQ on child gender

s.e. = .07

Hypo:

2.8 * s.e.

.07 * 2.8 = .206 sd, which makes
Fun with standard errors

```bash
> # 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>zread6</td>
<td>11,265</td>
<td>-2.04e-10</td>
<td>1</td>
<td>-3.235986</td>
<td>2.012693</td>
</tr>
<tr>
<td>zread4</td>
<td>16,336</td>
<td>-1.30e-11</td>
<td>1</td>
<td>-2.208911</td>
<td>4.469528</td>
</tr>
<tr>
<td>zread1</td>
<td>17,622</td>
<td>-1.61e-10</td>
<td>1</td>
<td>-1.392766</td>
<td>10.12822</td>
</tr>
<tr>
<td>dfemale</td>
<td>21,396</td>
<td>0.4882221</td>
<td>0.4998729</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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

. su zread6 zread4 zread1 dfemale

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>zread6</td>
<td>9,065</td>
<td>0.0979408</td>
<td>0.9649163</td>
<td>-3.215906</td>
<td>2.012693</td>
</tr>
<tr>
<td>zread4</td>
<td>9,065</td>
<td>0.0979004</td>
<td>0.977377</td>
<td>-2.171627</td>
<td>4.469528</td>
</tr>
<tr>
<td>zread1</td>
<td>9,065</td>
<td>0.061892</td>
<td>0.9888895</td>
<td>-1.386883</td>
<td>10.12822</td>
</tr>
<tr>
<td>dfemale</td>
<td>9,065</td>
<td>0.501048</td>
<td>0.5000265</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

and here's the correlation matrix for the test score measures:
```
. corr zread6 zread4 zread2
(obs=9,065)

<table>
<thead>
<tr>
<th></th>
<th>zread6</th>
<th>zread4</th>
<th>zread2</th>
</tr>
</thead>
<tbody>
<tr>
<td>zread6</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zread4</td>
<td>0.6770</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>zread2</td>
<td>0.5318</td>
<td>0.7608</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

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
-------|-------|----|---
Model  | 36.7626256 | 1  | 36.7626256
Residual | 8402.39646 | 9,063 | .927109838
Total   | 8439.15909 | 9,064 | .931063447

F(1, 9063) = 39.65
Prob > F = 0.0000
R-squared = 0.0044
Adj R-squared = 0.0042
Root MSE = 0.96287

|       | 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 % of the data:

. gen random = runiform()
. gen quarter = random <= 0.25
. tab quarter

<table>
<thead>
<tr>
<th>quarter</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6,787</td>
<td>74.87</td>
<td>74.87</td>
</tr>
<tr>
<td>1</td>
<td>2,278</td>
<td>25.13</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>9,065</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
Repeat our first full sample regression and note the sample size and standard error:

`. reg zread6 dfemale`

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs  =</th>
<th>F(1, 9063) = 39.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>36.7626256</td>
<td>1</td>
<td>36.7626256</td>
<td>Prob &gt; F = 0.0000</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>8402.39646</td>
<td>9,063</td>
<td>0.927109838</td>
<td>R-squared = 0.0044</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8439.15909</td>
<td>9,064</td>
<td>0.951063447</td>
<td>Adj R-squared = 0.0042</td>
<td></td>
</tr>
</tbody>
</table>
|          |          |     |       | Root MSE = 0.96287 |}

| 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`

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs  =</th>
<th>F(1, 2276) = 2.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.95013154</td>
<td>1</td>
<td>1.95013154</td>
<td>Prob &gt; F = 0.1465</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>2103.27394</td>
<td>2,276</td>
<td>0.924109815</td>
<td>R-squared = 0.0009</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2105.22407</td>
<td>2,277</td>
<td>0.924560418</td>
<td>Adj R-squared = 0.0005</td>
<td></td>
</tr>
</tbody>
</table>
|          |          |     |       | Root MSE = 0.96131 |}

| zread6   | Coef.    | Std. Err. | t   | P>|t| | [95% Conf. Interval] |
|----------|----------|------------|-----|-----|---------------------|
| dfemale  | .0585201 | .0402342   | 1.45 | 0.146 | -.0204775 - .1375178 |
| _cons    | .0828865 | .0203474   | 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
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 9,065</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>36.7626256</td>
<td>1</td>
<td>36.7626256</td>
<td>F(1, 9063) = 39.65</td>
</tr>
<tr>
<td>Residual</td>
<td>8402.39646</td>
<td>9,063</td>
<td>0.927109838</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R-squared = 0.0044</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.0042</td>
<td>Root MSE = 0.96287</td>
</tr>
<tr>
<td>Total</td>
<td>8439.15909</td>
<td>9,064</td>
<td>0.931063447</td>
<td></td>
</tr>
</tbody>
</table>

|             | 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
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 9,065</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3868.89866</td>
<td>2</td>
<td>1934.44933</td>
<td>F(2, 9062) = 3835.66</td>
</tr>
<tr>
<td>Residual</td>
<td>4570.26043</td>
<td>9,062</td>
<td>.504332425</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R-squared    = 0.4584</td>
<td>Adj R-squared = 0.4583</td>
</tr>
<tr>
<td>Total</td>
<td>8439.15909</td>
<td>9,064</td>
<td>0.931063447</td>
<td>Root MSE = 0.71016</td>
</tr>
</tbody>
</table>

|             | Coef.       | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-------------|-------------|------------|------|------|----------------------|
| dfemale     | .0260252    | .0149631   | 1.74 | 0.082 | -0.0033057 - 0.0553562 |
| zread4      | .6672878    | .0076551   | 87.17 | 0.000 | .6522821 - .6822935  |
| _cons       | .0195732    | .0105609   | 1.85 | 0.064 | -0.0011285 - 0.0402749 |
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)

<table>
<thead>
<tr>
<th></th>
<th>zread6</th>
<th>zread4</th>
<th>zread2</th>
</tr>
</thead>
<tbody>
<tr>
<td>zread6</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zread4</td>
<td>0.6770</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>zread2</td>
<td>0.5318</td>
<td>0.7608</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
```

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
--------|-----------|----|-----------|------------------------
Model   | 3868.89866| 2  | 1934.44933| F(2, 9062) = 3835.66
Residual| 4570.26043| 9,062| .504332425| Prob > F = 0.0000
--------|-----------|----|-----------|------------------------
Total   | 8439.15909| 9,064| .931063447| R-squared = 0.4583
         |           |    |           | Adj R-squared = 0.4583
         |           |    |           | Root MSE = 0.71016

|         | 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`

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 9,065</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3874.40394</td>
<td>3</td>
<td>1291.46798</td>
<td>F(3, 9061) = 2563.55</td>
</tr>
<tr>
<td>Residual</td>
<td>4564.75515</td>
<td>9,061</td>
<td>0.503780504</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>8439.15909</td>
<td>9,064</td>
<td>0.931063447</td>
<td>Adj R-squared = 0.4589</td>
</tr>
</tbody>
</table>

| zread6 | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|--------|-----------|-------|-----|---------------------|
| dfemale | 0.0247772 | 0.0149596 | 1.66 | 0.098 | -.0045471 to 0.0541014 |
| zread4  | 0.6377669 | 0.0117594 | 54.23 | 0.000 | .6147158 to 0.6608181 |
| zread2  | 0.0386455 | 0.0116904 | 3.31 | 0.001 | .0157297 to 0.0615614 |
| _cons   | 0.0204129 | 0.0105581 | 1.93 | 0.053 | -.0002834 to 0.0411093 |
```

⇒ ~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 multicollinearity problems.
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.]

<table>
<thead>
<tr>
<th>Name</th>
<th>Comments</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw score</td>
<td>Both X and Y are in natural units</td>
<td>$\Delta Y / \Delta X$, where $\Delta X$ is a one raw unit change in X</td>
</tr>
<tr>
<td>Effect size – in experiments.</td>
<td>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.</td>
<td>$\frac{\Delta Y}{\sigma_Y}$, where T is (1,0) indicator of treatment status</td>
</tr>
<tr>
<td>Standardized coefficients (&quot;effect sizes&quot;, $\beta$, “Beta weight&quot;)</td>
<td>Both X and Y are in standard deviation units. In a bivariate regression, $\beta^2 = \text{the explained variance (R}^2\text{)}$ of the regression, so $\beta$s are taken to indicate relative (explanatory) importance. In this important sense, $\beta$s are comparable indicators of relative explanatory power across independent variables in an equation.</td>
<td>$\frac{\Delta Y}{\sigma_Y} / \frac{\Delta X}{\sigma_X}$ (note that this is the same as: $\frac{\Delta Y}{\Delta X} \cdot \frac{\sigma_Y}{\sigma_X}$, where $\Delta Y / \Delta X$ is the raw score coefficient)</td>
</tr>
<tr>
<td>Percent change coefficients</td>
<td>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 $\Delta \ln Y$ approximates percentage change in Y for small changes in X.</td>
<td>$\frac{\Delta Y}{Y} / \Delta X$ (note that this is the same as: $\frac{\Delta Y}{\Delta X} \cdot \frac{X}{Y}$, where $\Delta Y / \Delta X$ is the raw score coefficient)</td>
</tr>
<tr>
<td>Elasticity coefficients</td>
<td>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.</td>
<td>$\frac{\Delta Y}{Y} / \Delta X$</td>
</tr>
<tr>
<td>Months of learning (or development)</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>interpretation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benefit-cost coefficients</td>
<td>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) bang for the buck.</td>
<td></td>
</tr>
</tbody>
</table>

2nd class Part II

Effect sizes for 10

1. Raw score
   Both X, Y
   ("instantaneous" in natural units)
   \[
   \frac{\Delta y}{\Delta x} = \frac{10}{1} = 10
   \]
   [.14 years per 10k $\Delta Y$]

2. Effect size
   in experiment
   Y is standardized
   X in natural units
   (usually 1/0)
   \[
   \frac{.67}{1} = .67 \text{ sd}
   \]

3. Standardized
   coefficient
   (also called "effect
   sizes / $\beta$"

4. Percent change
   in Y
   Y is 20 change
   relation to the
   mean
   \[
   \frac{\Delta y}{\Delta x} = .10 \%
   \]

5. Elasticity
   in X
   Y and X are changed
   as 20 change

6. Months of learning
   Express $\Delta Y$ as a matter of learning

467
   benefit / cost
   ratio
   $\frac{Y}{X}$
   $\frac{\text{value of change in } Y}{\text{value of cost in } X}$
   $\frac{Y}{\text{cost of change intervention}}$
Suppose $Motivation = a + b \times Age$ for middle schoolers

<table>
<thead>
<tr>
<th>Age in years</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.58</td>
<td>.64</td>
<td></td>
</tr>
<tr>
<td>Mot 1-5 scale</td>
<td>3.61</td>
<td>.61</td>
</tr>
</tbody>
</table>

In fact $Mot = 5.25 - .14 \times Age$

$\begin{pmatrix}
(.38) \\
(.03)
\end{pmatrix}$

$\Rightarrow$ motivation declines across middle school

First of all $5.25$? $Mot$ at age $0$

Focus on $- .14$ big or small?

$\Rightarrow$ raw scale weights are often difficult to interpret

In fact if age was scaled in months rather than years

$Mot = 5.25 - \frac{14/12}{.03} = - .012 \times Age$

1. Standardized numerator $\frac{- .14}{.61} = 12.3$ rd per year of age

2. Standardized effects “$\beta$” “Beta weight”

$\Rightarrow$ divide $- .14$ in std of $\frac{\Delta Y}{\Delta x}$, how $\frac{\Delta Y}{\Delta x}$ over $\frac{\Delta Y}{\Delta x}$ $\sigma_x$
Part II

In our example

\[
\frac{\Delta Y}{\Delta X} \cdot \frac{1}{\sqrt{\sigma_x}} = -1.14 \cdot \frac{1}{.64} = -1.5
\]

\[
\frac{\Delta Y}{\Delta X} = \frac{\Delta Y}{\Delta X} \cdot \frac{\sigma_x}{\sigma_y} = \text{raw score} \times \frac{\sigma_x}{\sigma_y}
\]

Virtue of standardized effects

1. Facilitates comparisons across coefficients in a regression \( \beta = \text{"relative importance"} \)

in bivariate \( Y = \alpha + \beta x \), \( \beta ^2 = R^2 \)

in \( Y = \alpha + \beta_1 x_1 + \beta_2 x_2 \) (and \( x \)'s are uncorrelated)

\[ \beta_1 ^2 + \beta_2 ^2 + \ldots = R^2 \]
Problems with student computer.

(1) Sometimes near score change are much easier to think about.

\[ \text{Cog} = a + b \text{ Treatment} \]

but not in \( \frac{1}{10} \) value

\( \text{Standard} \) in \( \text{SD} \)

\( \text{Int} \)

(2) What SD to use?

\( \text{Abecedarian had high by injected on IQ at age 19} \)

4 pts, \( \frac{1}{2} = 0.40 \) SD

n, 40 SD using the sample SD.

4 pts, \( \frac{1}{2} = 0.28 \) SD

\( \text{March of learning} \)

\( \text{Couple of survey result is given.} \)

\( \text{Readouts of learning} \)
2nd class Part II

\[
\frac{\Delta Y}{\Delta x} = 0 \quad \text{and} \quad \Delta \sigma = \frac{3.82}{\text{per year}}
\]

then 2D variance accord with a one unit change in \(x\)

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

\[
\begin{array}{c}
\text{Percent change in } Y \\
\text{also comes from using a log } Y \\
\text{function from}
\end{array}
\]

\[
\ln Y = a + b \cdot Ed. \quad \text{Going from } Ed_1 \text{ to } Ed_2 + 2
\]

takes you from \(\ln Y_1\) to \(\ln Y_2\)

\[
\Delta \ln Y = \ln Y_2 - \ln Y_1
\]

\[
\Delta \ln Y = \left(\frac{Y_2}{Y_1}\right) - 1 \quad \text{for small change in } \frac{Y_2}{Y_1}
\]
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!
1. Equation of ANOVA in dummy variable regression

   go over hand out

   Why bother with regression?
   - adapt for context

2. Flexible way of examining relative between
   X and Y

   Table 3 from ASR article

   Complete Educa

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>15-25</th>
<th>25-35</th>
<th>35-50</th>
<th>50+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited</td>
<td>.0</td>
<td>.82</td>
<td>1.41</td>
<td>1.69</td>
<td>2.35</td>
</tr>
<tr>
<td>25-35</td>
<td>2.5</td>
<td>2.0</td>
<td>1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35-50</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   Col
Regression and ANOVA

(Almost identical) Regression

ANOVA

Regression v. ANOVA

Different reference/omitted group
Table 3. Coefficients from the Regression of Child’s Outcome Variables on Family Income at Ages 0 to 15: Panel Study of Income Dynamics

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Years of Completed Schooling*</th>
<th>High School Completionb</th>
<th>Hazard of Nonmarital Birthc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)  (2)  (3)  (4)</td>
<td>(1)  (2)  (3)  (4)</td>
<td>(1)  (2)  (3)  (4)</td>
</tr>
<tr>
<td><strong>Family Income at Child’s Ages 0 to 15</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear function</td>
<td>.14* (.02)</td>
<td>.23* (.07)</td>
<td>-.43* -.43*</td>
</tr>
<tr>
<td>Spline function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income &lt; $20,000</td>
<td>.150* (.29)</td>
<td>1.97* (.44)</td>
<td>-.50 (.41)</td>
</tr>
<tr>
<td>Difference between</td>
<td>-.17* (.30)</td>
<td>-.84* (.46)</td>
<td>-.08 (.44)</td>
</tr>
<tr>
<td>income &lt; $20,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and &gt; $20,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural logarithm</td>
<td></td>
<td>1.16* (.11)</td>
<td>1.35* (.26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-1.18* (.26)</td>
</tr>
<tr>
<td><strong>Dummy Variables for Family Income</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$15,000 to $24,999</td>
<td></td>
<td>1.41* (.27)</td>
<td>-.54 (.35)</td>
</tr>
<tr>
<td>$25,000 to $34,999</td>
<td></td>
<td>1.83* (.28)</td>
<td>-.94 (.41)</td>
</tr>
<tr>
<td>$35,000 to $49,999</td>
<td></td>
<td>2.48* (.28)</td>
<td>-1.44* (.43)</td>
</tr>
<tr>
<td>$50,000 and over</td>
<td></td>
<td>2.64* (.29)</td>
<td>-2.40* (.49)</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>.192 .201 .219 .216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2 Log likelihood</td>
<td>-718.9 -702.6 -701.1 694.6</td>
<td>1266.1 1266.1 1271.1 1267.3</td>
<td></td>
</tr>
</tbody>
</table>

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).

*OLS models; N = 1,323.

bLogistic models; N = 1,323.

cCox models; N = 620.

*p < .05 (two-tailed tests)