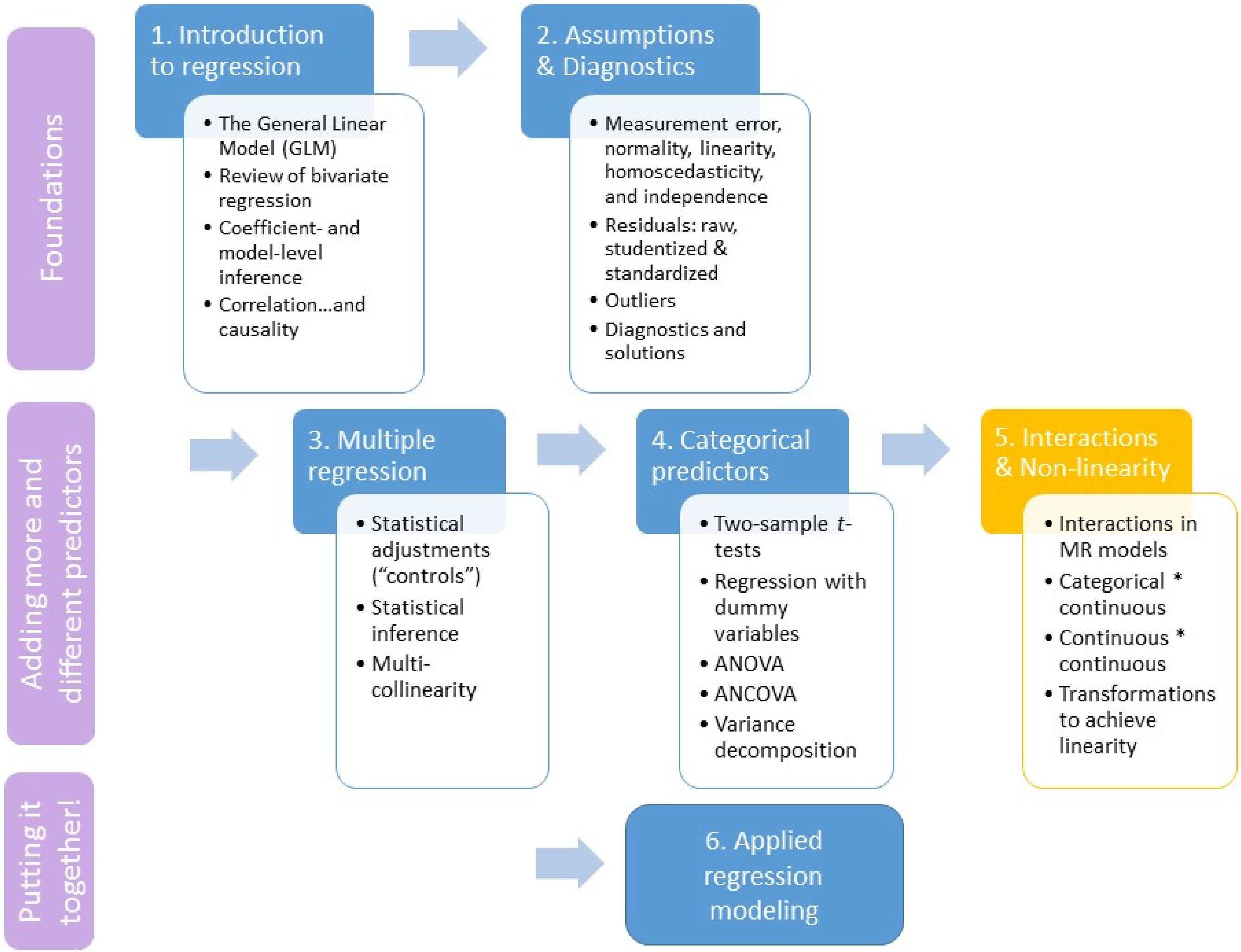
# Non-linearity

## EDUC 643: Unit 5 Part II

David D. Liebowitz

UNIVERSITY OF OREGON

# Roadmap



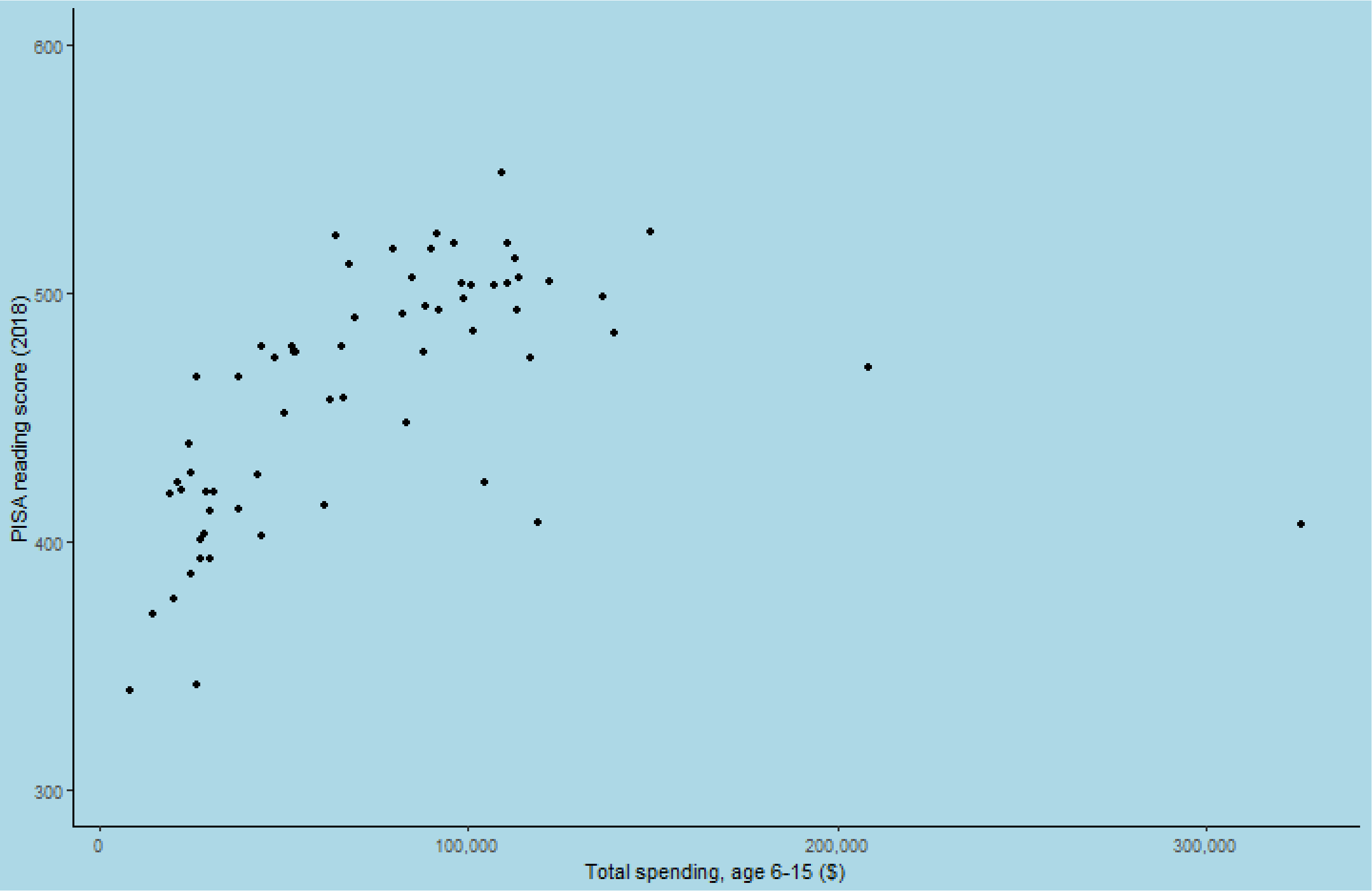
Foundations
1. Introduction to regression
• The General Linear Model (GLM)
• Review of bivariate regression
• Coefficient- and model-level inference
• Correlation…and causality
2. Assumptions & Diagnostics
• Measurement error, normality, linearity, homoscedasticity, and independence
• Residuals: raw, studentized & standardized
• Outliers
• Diagnostics and solutions
Adding more and different predictors
3. Multiple regression
• Statistical adjustments (“controls”)
• Statistical inference
• Multi-collinearity
4. Categorical predictors
• Two-sample t-tests
• Regression with dummy variables
• ANOVA
• ANCOVA
• Variance decomposition
5. Interactions & Non-linearity
• Interactions in MR models
• Categorical * continuous
• Continuous * continuous
• Transformations to achieve linearity
Putting it together!
6. Applied regression modeling

# Goals of the unit

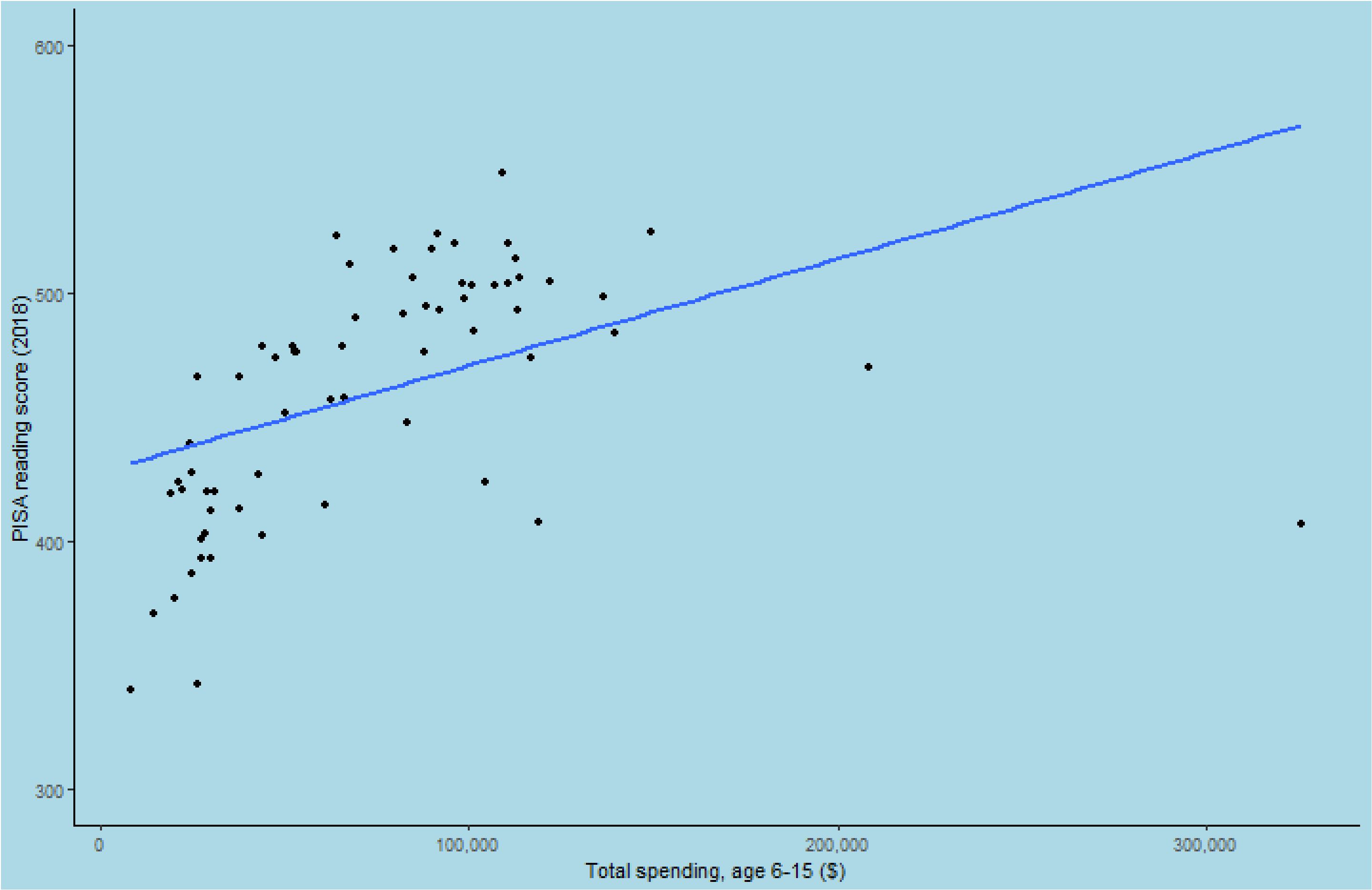
- Describe in writing and verbally the assumptions we violate when we fit a non-linear relationship with a linear model
- Transform non-linear relationships into linear ones by using logarithmic scales
- Estimate regression models using logarithmic scales and interpret the results
- Estimate models with quadratic and higher-order polynomial terms (special kinds of interactions)
- Select between transformation options

# Non-linearity

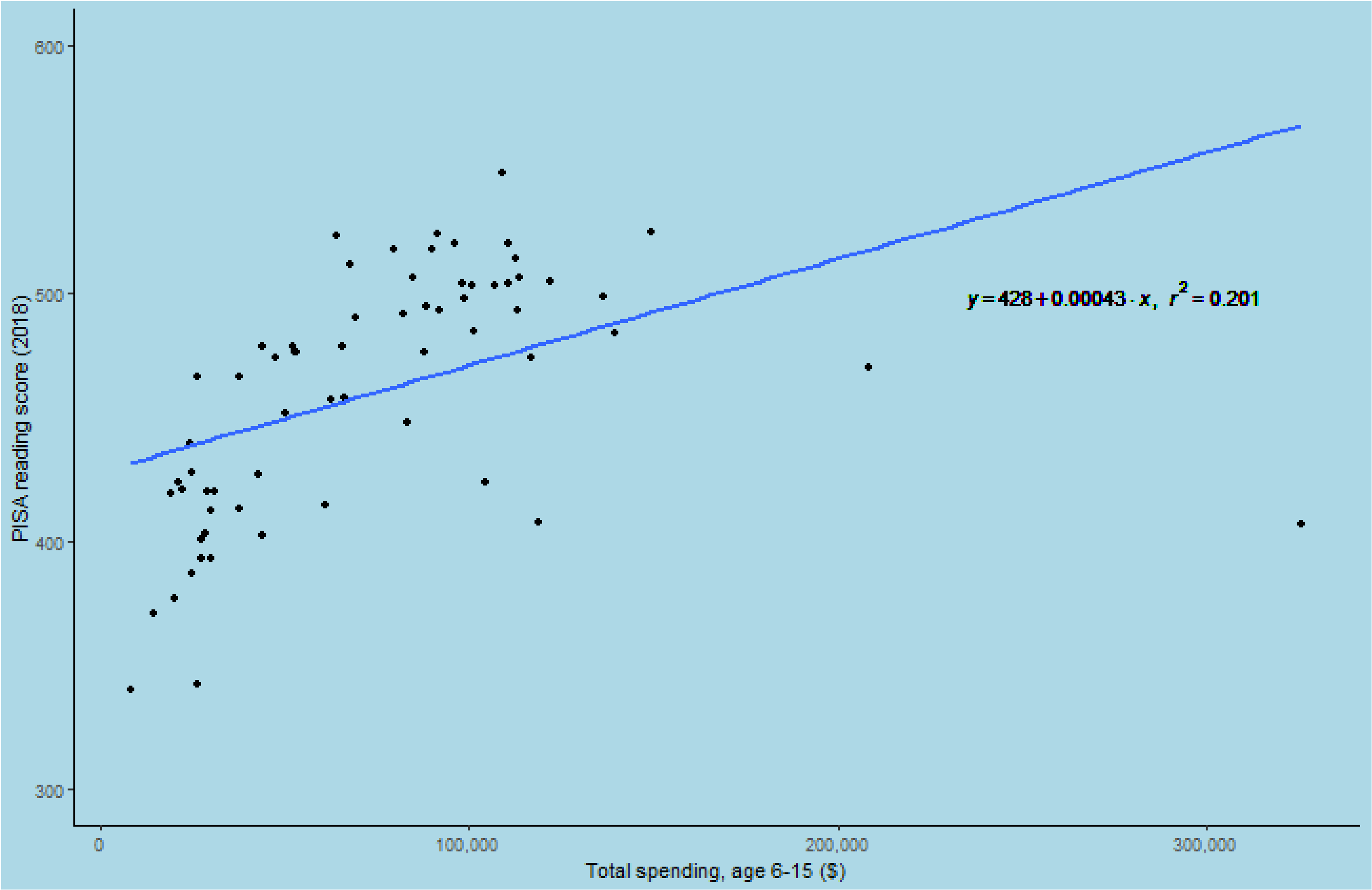
# $ and learning



# $ and learning



# $ and learning




***If assumptions hold***, each $10,000 diff in total spending associated, on average, with 4.3 scale score point difference in reading scores. **But do they?**

# Linear?

```r
# Fit the model
fit <- lm(read_score ~ total_spending, data=pisa)
# Generate residual vs fitted plot
pisa$resid <- resid(fit)
pisa$fitted <- fitted(fit)
ggplot(pisa, aes(fitted, resid)) + geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype="dashed")
```

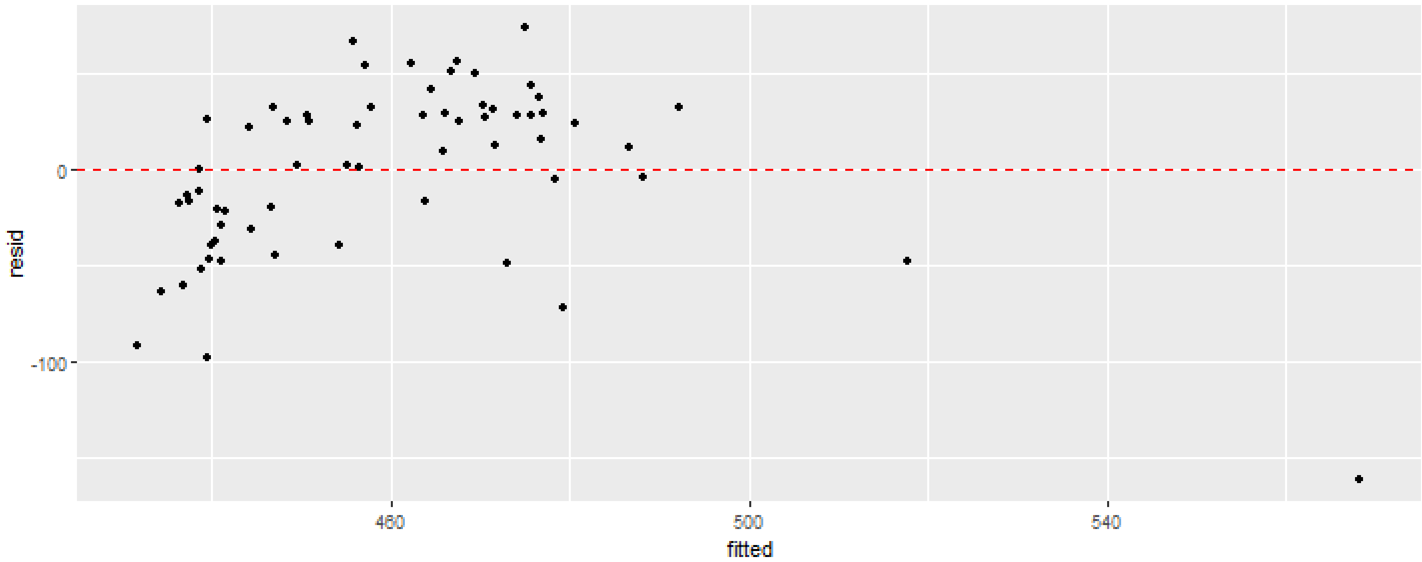
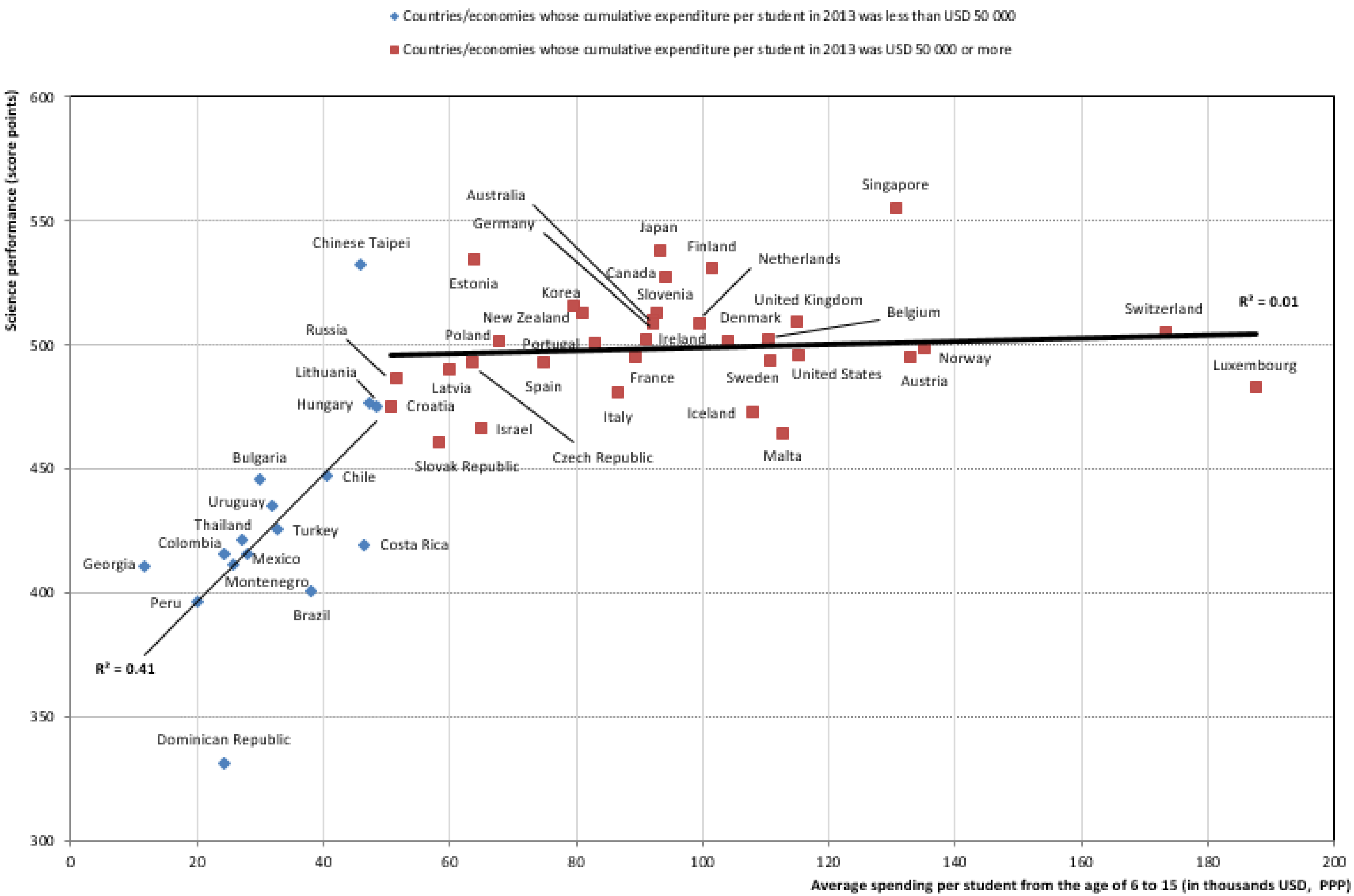
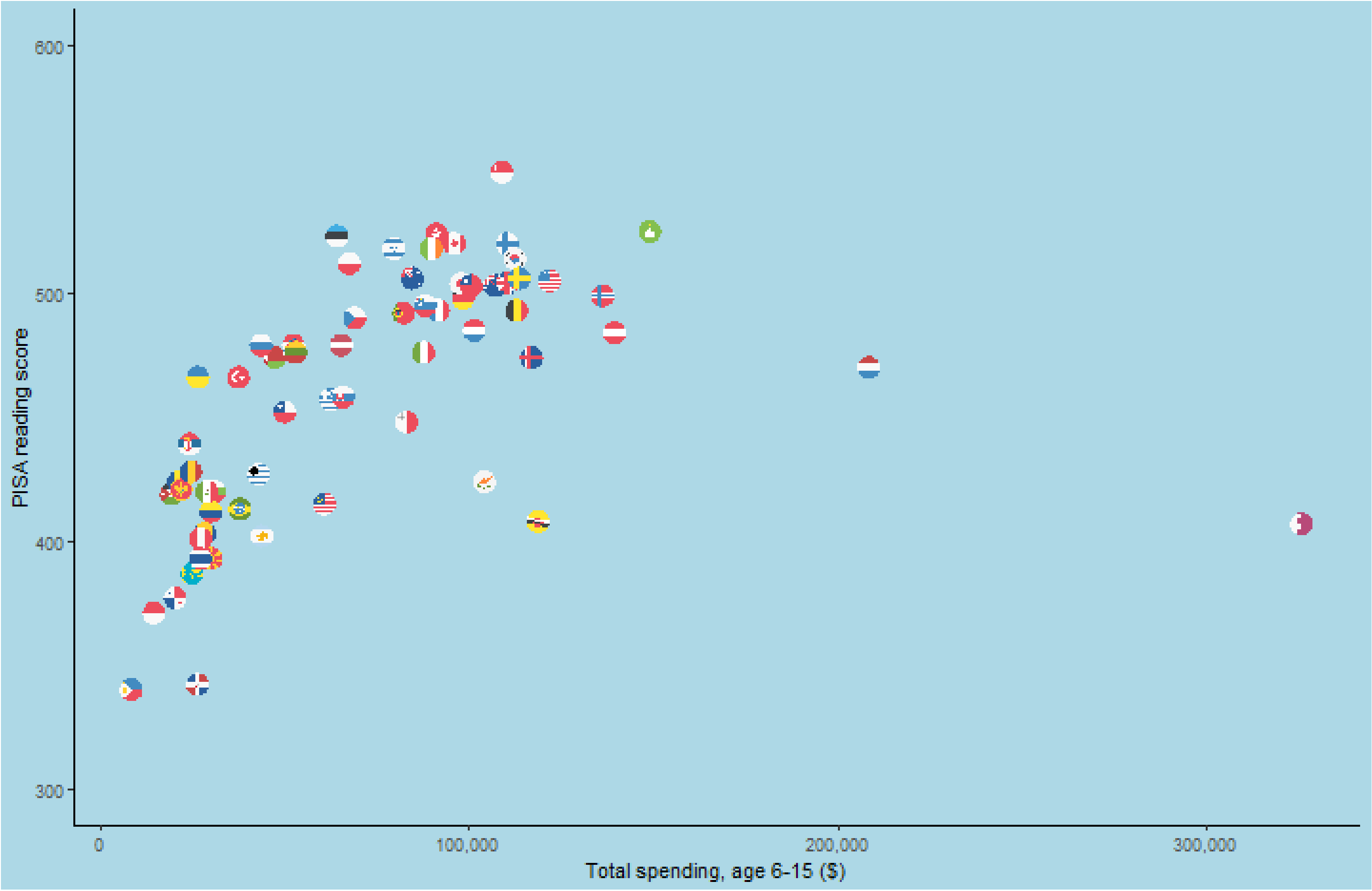


Figure II.6.2
**Spending per student from the age of 6 to 15 and science performance**

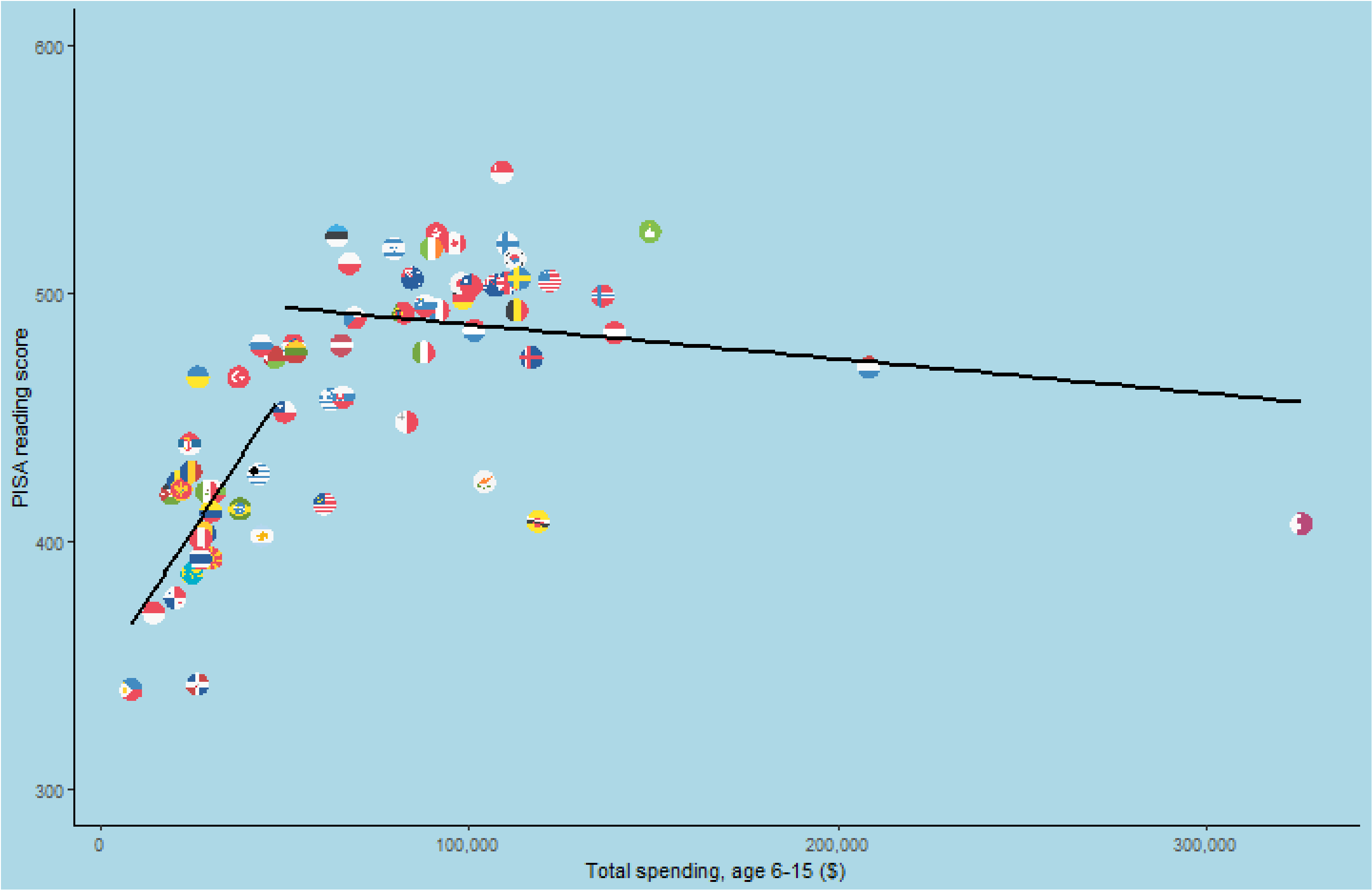


# Make it nice



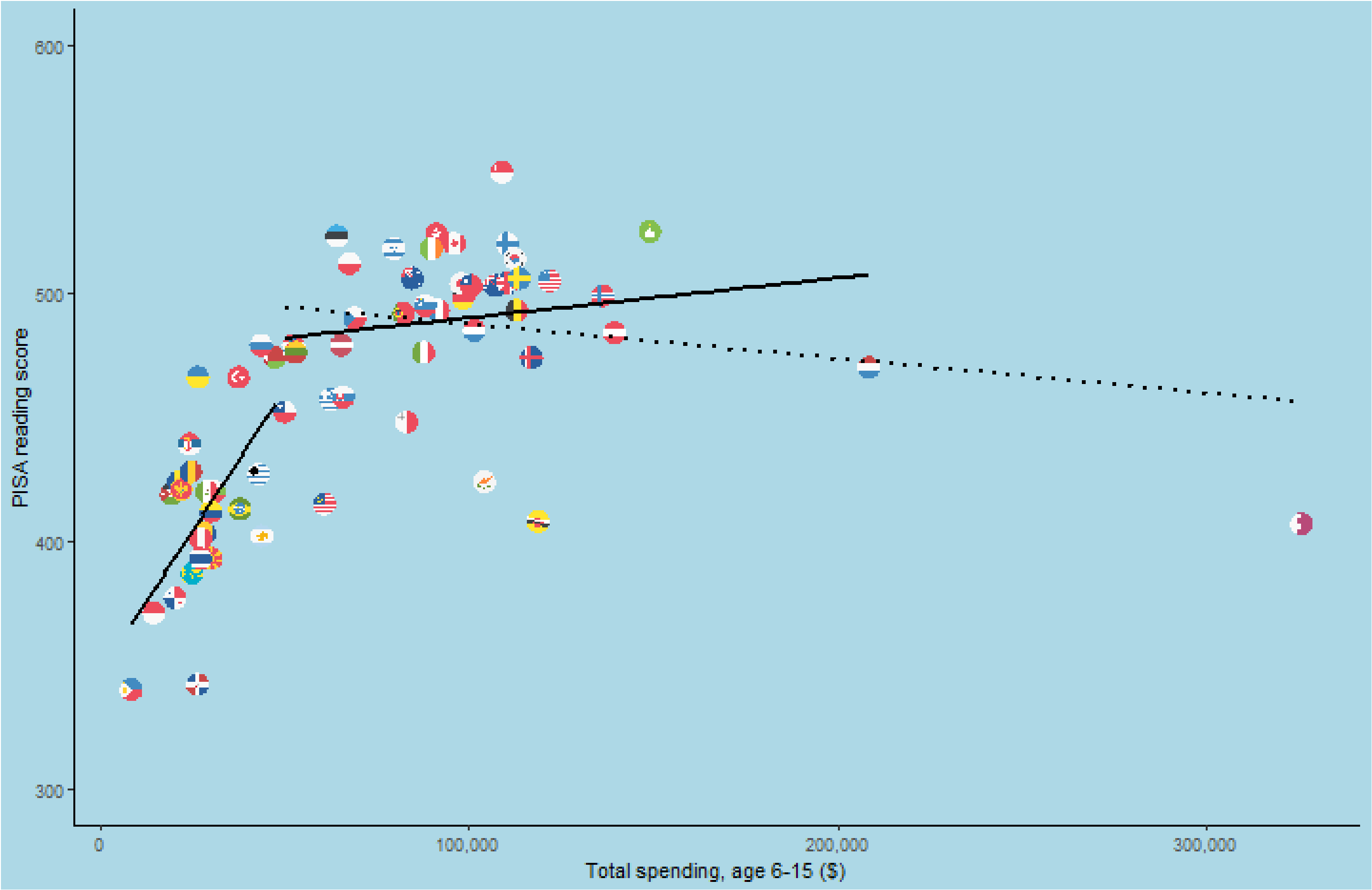

At low levels of spending the relationship between ***total_spending*** and ***read_score*** has a big magnitude. At higher levels of spending, it seems much more modest (negative?).

# Piecewise



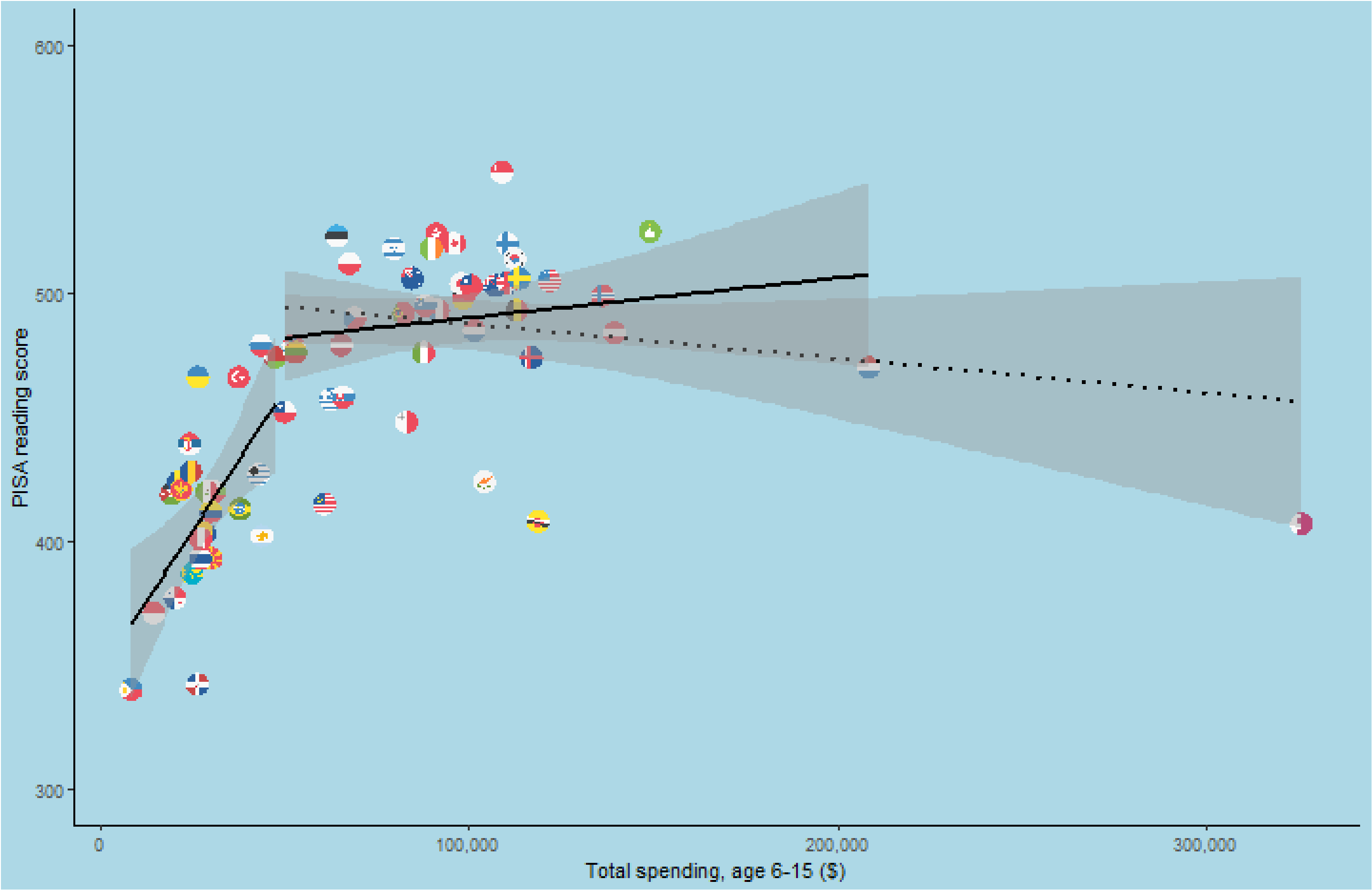
600
500
400
300
PISA reading score
0
100,000
200,000
300,000
Total spending, age 6-15 ($)

# Piecewise



600
500
400
300
PISA reading score
0
100,000
200,000
300,000
Total spending, age 6-15 ($)

# Piecewise

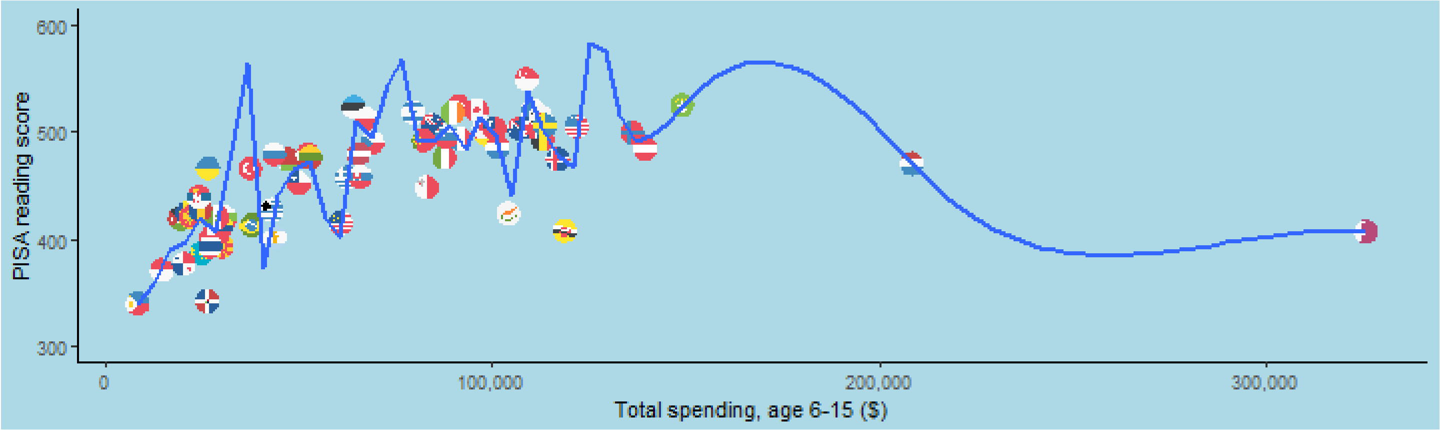


600
500
400
300
PISA reading score
0
100,000
200,000
300,000
Total spending, age 6-15 ($)

# Piecewise

While it is true, as we've said before that *locally all relationships are linear*, we've identified some emerging issues:

- Cut points arbitrary and these choices may substantially alter nature of observed relationship
- With large data "eyeballing" linear sub-segments impossible
- Increasing loss of power (larger standard errors and confidence intervals, greater influence of outliers)
- **Overfitting** risks increase
    - Analysis conforms to particularly to your specific data, but generalizes poorly to population of inference



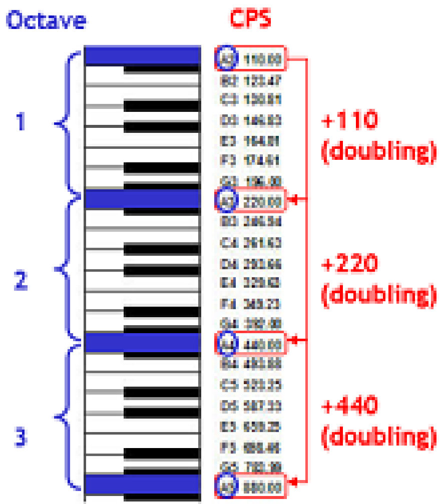

**Solutions**: transformations and polynomials

# Logarithmic transformations in X

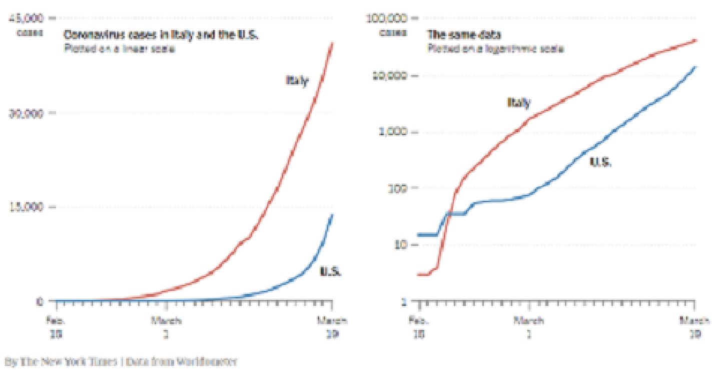
# Log transformations

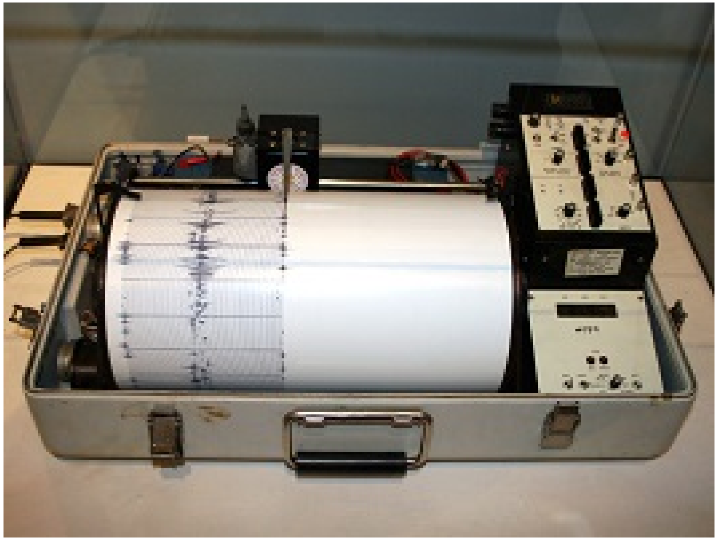
- We can posit a **non-linear relationship** between X and Y *in the population*
- Any non-linear relationship implies that the relationship between X and Y is relative to a particular value of X and/or Y, not absolute (the slope is non-constant)
- **Transformations** (i.e., spreading out in some cases and compressing in others the values of our X and Y variables) allow us to fit non-linear relationships within the existing machinery of the general linear model

# Log transformations in life




↑ 1 octave = doubling of cycles-per-second



| Seismic-wave amplitude | Location | Richter Scale |
|---|---|---|
| 1,000,000 | Christchurch, 2010 | 6.0 |
| 10,000,000 | Port-au-Prince, 2010 | 7.0 |
| 100,000,000 | Sichuan, 2008 | 8.0 |
| 1,000,000,000 | Sumatra, 2004 | 9.0 |

↑ 1 Richter = 10x ↑ SWA

# A log 🌳 you say??

Logs are the function we can perform to "undo" raising a number to a power. If a number is equal to a base raised to a power ($x = base^{power}$), then a logarithim of a given base is the number you would have to raise to that power to get $x$:

**Exponents**

$10 = 10^1$

$100 = 10^2$

$1,000 = 10^3$

$10,000 = 10^4$

$100,000 = 10^5$

**Logarithms**

$\log_{10}(10) = 1$

$\log_{10}(100) = 2$

$\log_{10}(1,000) = 3$

$\log_{10}(100,000) = 4$

$\log_{10}(100,000) = 5$

Each 1 unit increase in a base-10 logarithm represents a 10-fold increase in $x$. Can have logarithms of different base.

# A log 🌳 you say??

Logs are the function we can perform to "undo" raising a number to a power. If a number is equal to a base raised to a power ($x = base^{power}$), then a logarithim of a given base is the number you would have to raise to that power to get $x$:

**Exponents**

$2 = 2^1$

$4 = 2^2$

$8 = 2^3$

$16 = 2^4$

$32 = 2^5$

**Logarithms**

$\log_2(2) = 1$

$\log_2(4) = 2$

$\log_2(8) = 3$

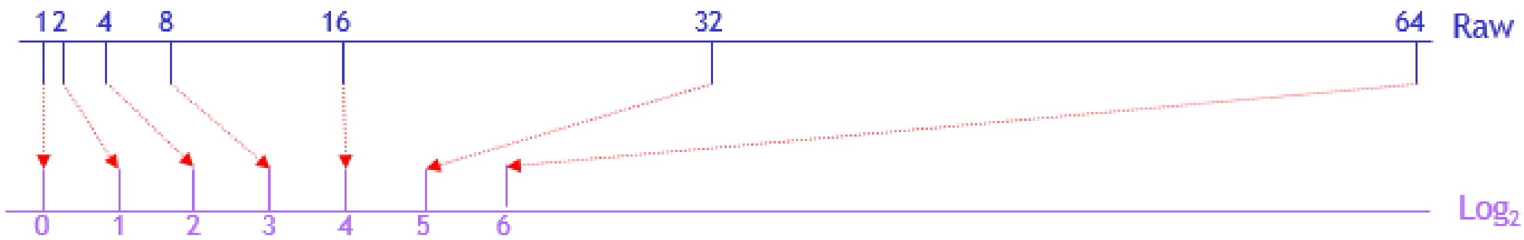
$\log_2(16) = 4$

$\log_2(32) = 5$

Each 1 unit increase in a base-2 logarithm represents a doubling of $x$.

Can say this as: “Log base 2 of 32 is 5” or “Log base 10 of 1,000 is 3”

# Understanding logs




**Some key concepts:**

- Taking logs spreads out the distance between small (closer to 0) values and compresses the distance between large (further from zero) values.
- Log base anything(1) is = 0
- Log base anything(0) is undefined (can't raise anything to a power and get 0)
- Log base anything>0(negative number) is undefined (technically a complex number)
- Taking logs is a **monotonic** transformation; doesn't change the order of any of the underlying raw values

# $ and scores?

Let's try transforming our X variable (*total_spending*) on a logarithmic scale; can do this directly in our plot:

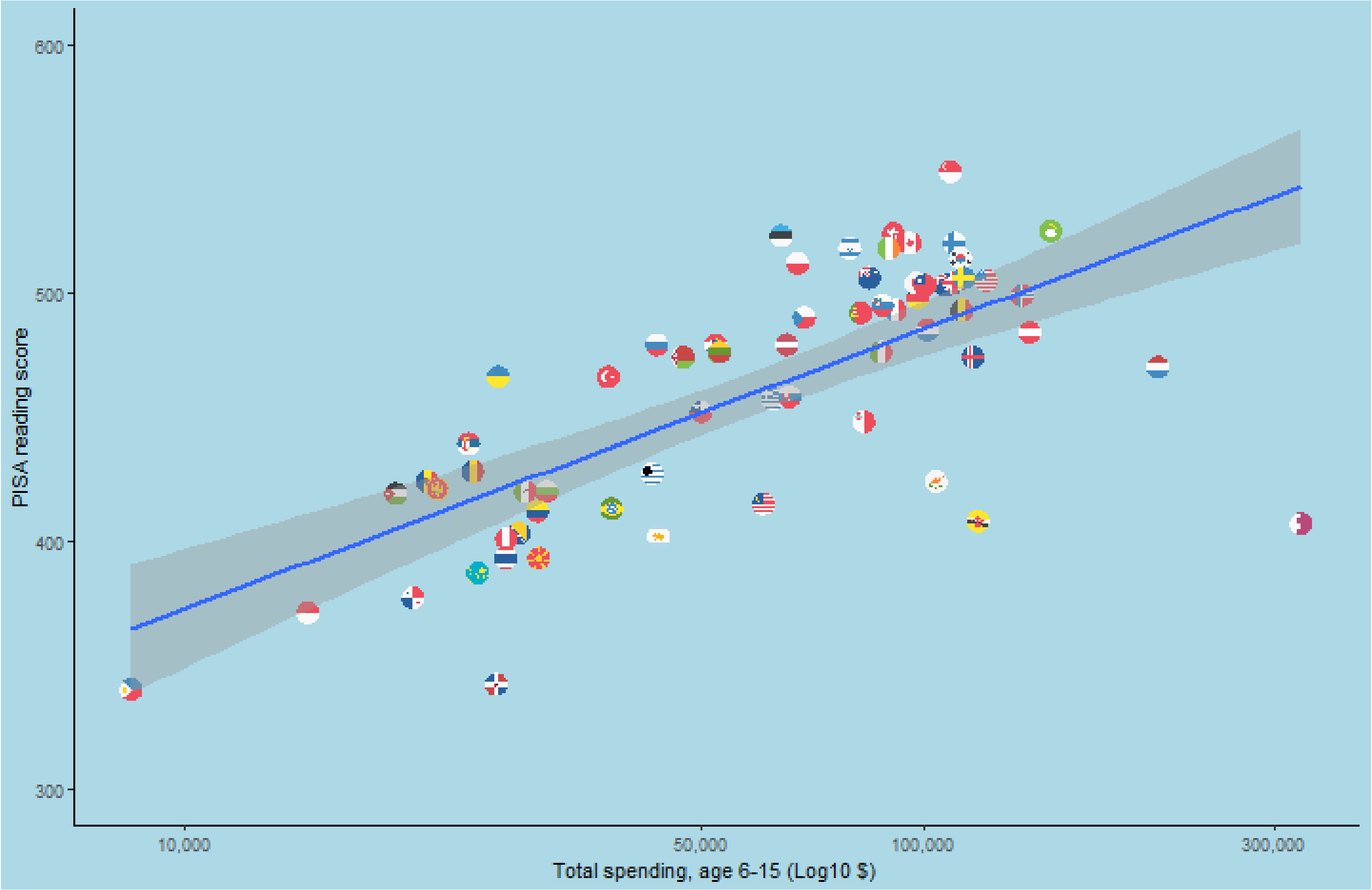
```
log_flag <- flag +
              xlab("Total spending, age 6-15 (Log10 $)") +
              scale_x_log10(breaks=c(10000, 50000, 100000, 300000),
                            label=scales::comma)
```

# $ and scores?

Let's try transforming our X variable (*total_spending*) on a logarithmic scale; can do this directly in our plot:

# $ and scores?

Let's try transforming our X variable (*total_spending*) on a logarithmic scale; can do this directly in our plot:

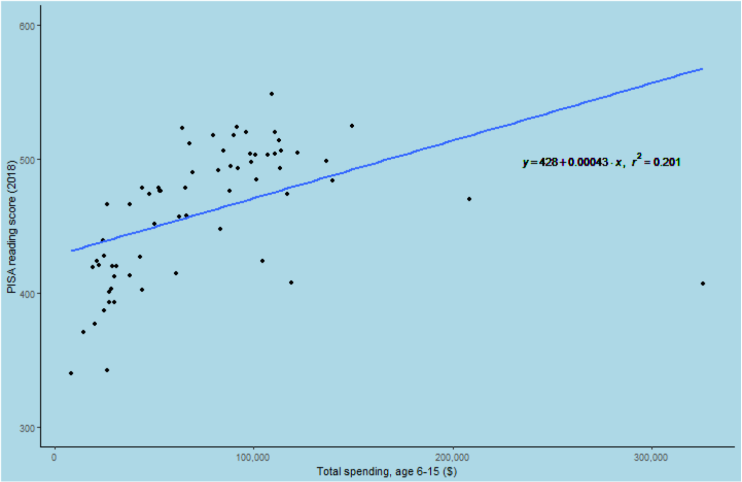


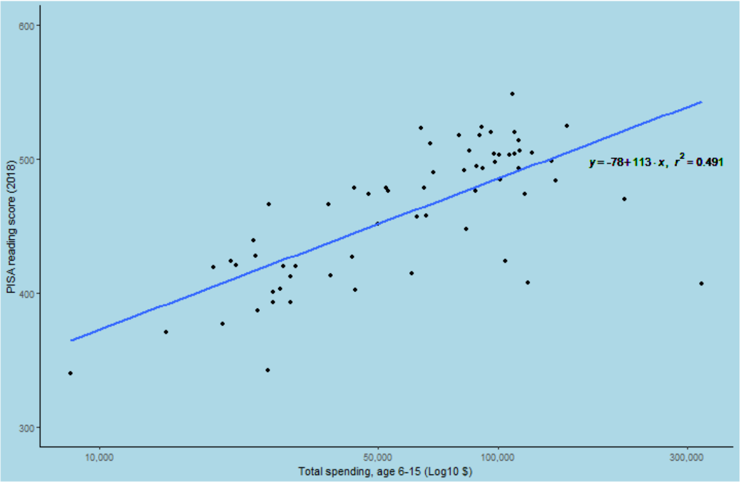
# Regress read on $\log_{10}(spend)$

```
summary(lm(read_score ~ log10(total_spending), data=pisa))
```

```
##
## Call:
## lm(formula = read_score ~ log10(total_spending), data = pisa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -136.50  -20.83   11.00   22.42   59.11
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -78.03      69.14  -1.129    0.263
## log10(total_spending)   112.74      14.46   7.798 8.06e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.59 on 63 degrees of freedom
## Multiple R-squared:  0.4911,	Adjusted R-squared:  0.4831
## F-statistic:  60.8 on 1 and 63 DF,  p-value: 8.062e-11
```

# Conceptually



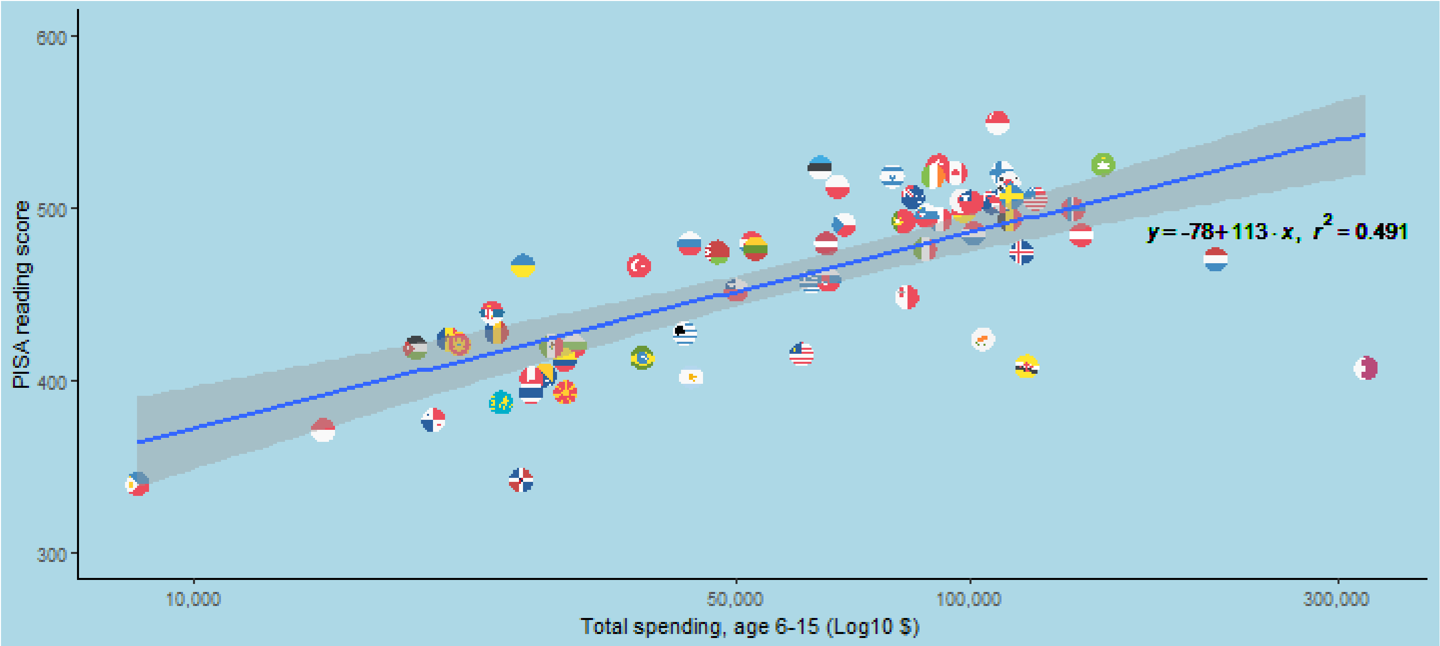




$$\hat{READ}_j = 428 + 0.00043 \times SPEND_j$$

$$\hat{READ}_j = -78.03 + 112.74 \times \log_{10}(SPEND_j)$$

- In ed/dev psych this kind of curve is called “learning curve”; represents standard rate of learning
- More broadly, increasing exponential decay or diminishing marginal returns

# Interpret



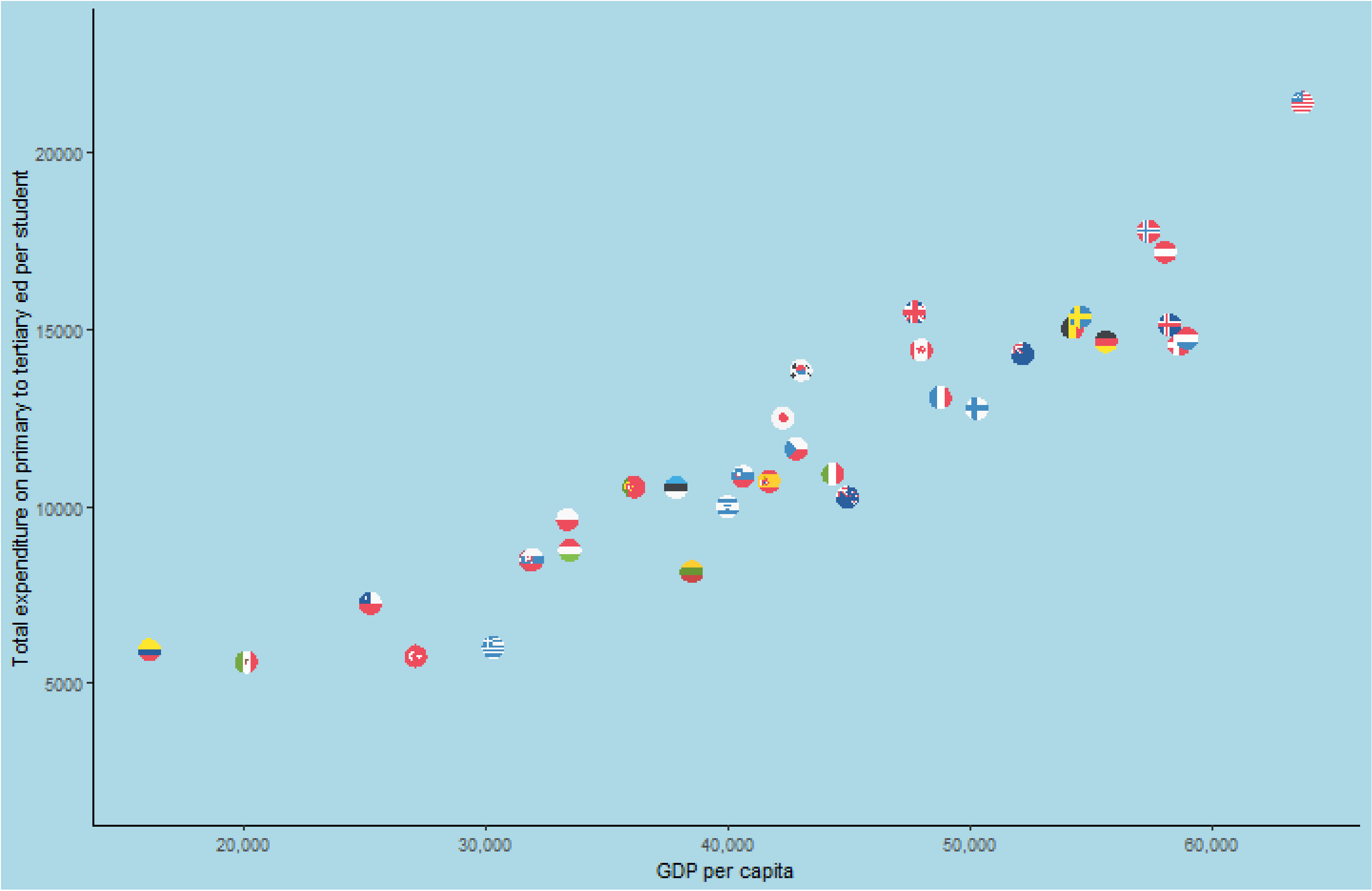

**Some alternative ways to describe this relationship:**

- Average reading scores in the population of countries sitting for the 2018 PISA reading test were 112.7 points higher for every ten-fold increase in cumulative educational spending on children aged 6-15.
- As cumulative education spending on children aged 6-15 is ten times higher, reading scores in the population of countries sitting for the 2018 PISA reading test were 112.7 points higher, on average.
- We predict that two countries that spend an order of magnitude (e.g., $10,000 vs. $100,00) apart on cumulative educational expenditures on children aged 6-15 will have PISA reading scores 112.7 points apart.

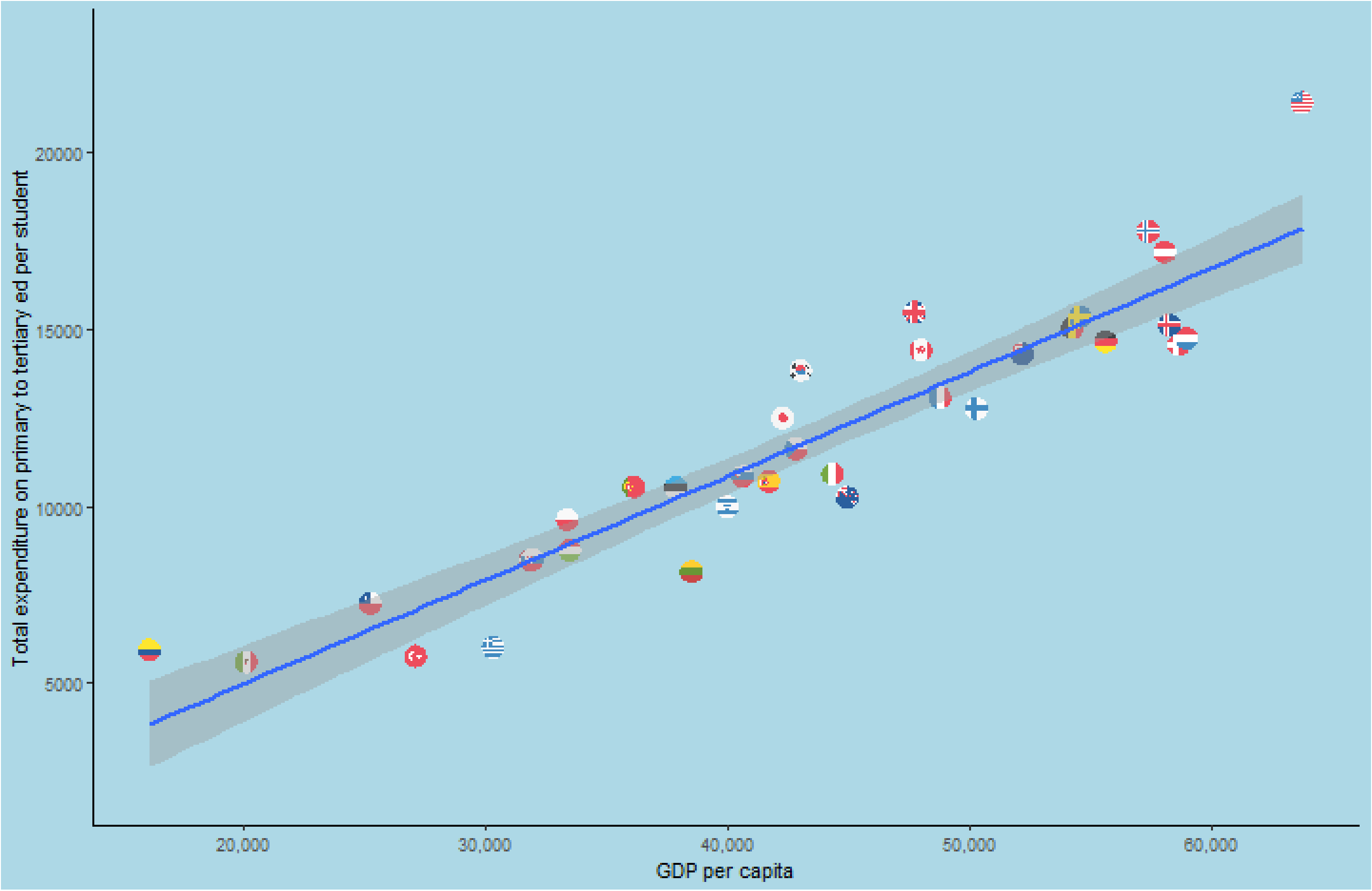
# Log transformations in Y

## aka Exponential growth curve

# GDP and PPE


Total expenditure on primary to tertiary ed per student
20000
15000
10000
5000
20,000
30,000
40,000
50,000
60,000
GDP per capita

# GDP and PPE


Total expenditure on primary to tertiary ed per student
20000
15000
10000
5000
20,000
30,000
40,000
50,000
60,000
GDP per capita

# An alternative model

The relationship of GDP and PPE are relative to their respective values. The relationship has a smaller magnitude when GDP per capita is smaller and a larger magnitude when GDP per capita is larger. Can use a log transformation to capture the non-absolute (non-constant) nature of the slope:

$$PPE_j = \beta_0 * 2^{(\beta_1 GDP_j + \varepsilon)}$$

$$\log_2(PPE_j) = \log_2 \beta_0 + \beta_1 GDP_j + \varepsilon$$

# Interpreting this

Can interpret log outcomes as percent changes because:

$$Y_1 = \beta_0 2^{\beta_1 X_1}$$

$$Y_2 = \beta_0 2^{\beta_1 (X+1)} = \beta_0 2^{\beta_1 X} 2^{\beta_1}$$

$$\frac{Y_2}{Y_1} = \frac{\beta_0 2^{\beta_1 X} 2^{\beta_1}}{\beta_0 2^{\beta_1 X}} = 2^{\beta_1}$$

So, $Y_2$ is $2^{\beta_1}$ times larger than $Y_1$! Depends on key properties of logs:

- log(xy) = log(x) + log(y)
- $\log(x^p)$ = p*log(x)

Percent growth rate = $100(2^{\beta_1} - 1)$

Regress log(Y) on X and substitute the estimated slope into the equation for the percent growth rate to obtain the estimated percent growth rate per unit change in X.

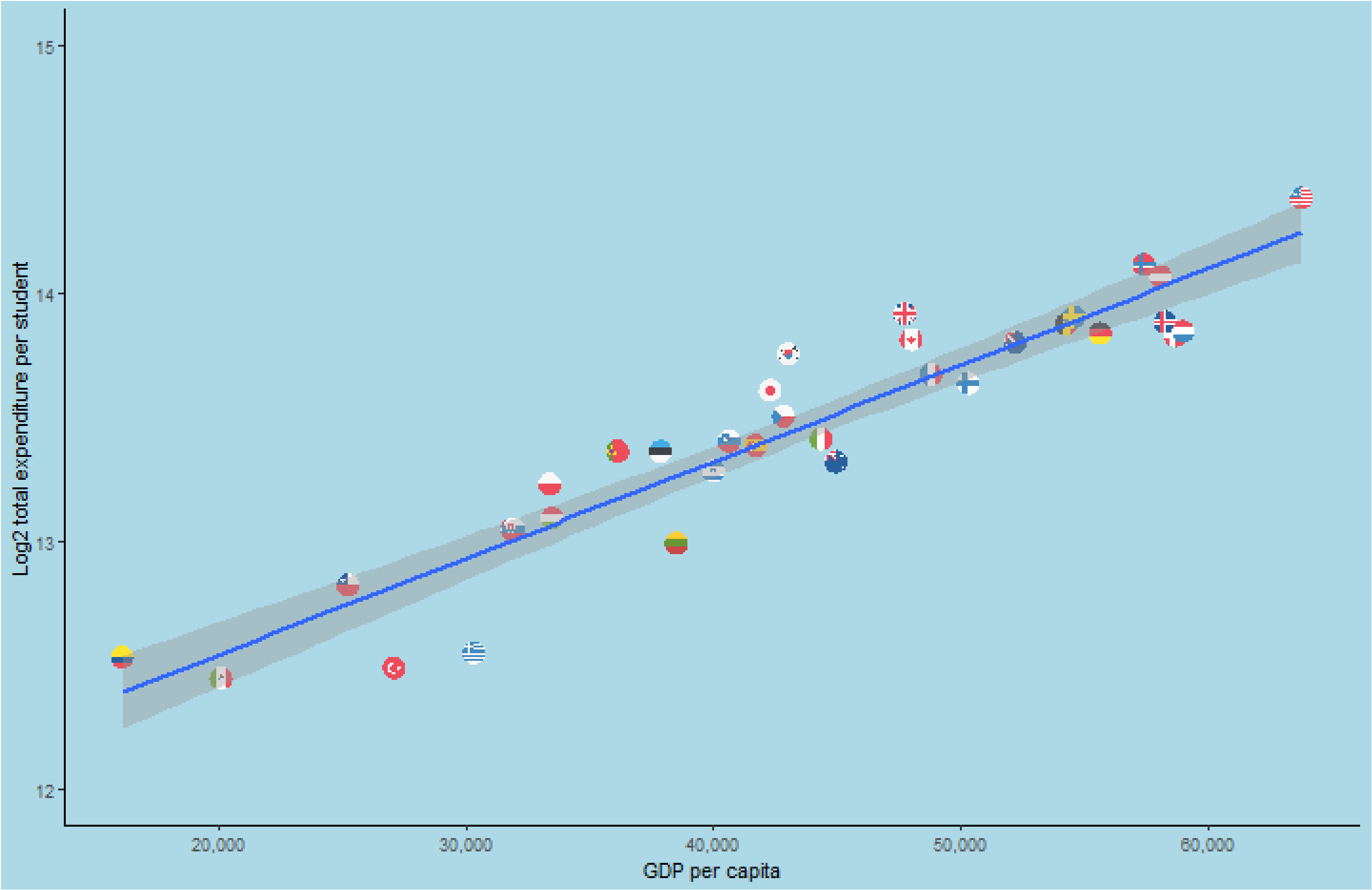
$Y_2 = 2^{\beta_1} Y_1$ is the same thing as saying the percent growth rate is $100(2^{\beta_1} - 1)$

# Visualized Y transformation

```
oecd$log2ppe <- log2(oecd$ppe)

log_ppe <- ggplot(oecd, aes(x=gdp, y=log2ppe))
```

# Visualized Y transformation



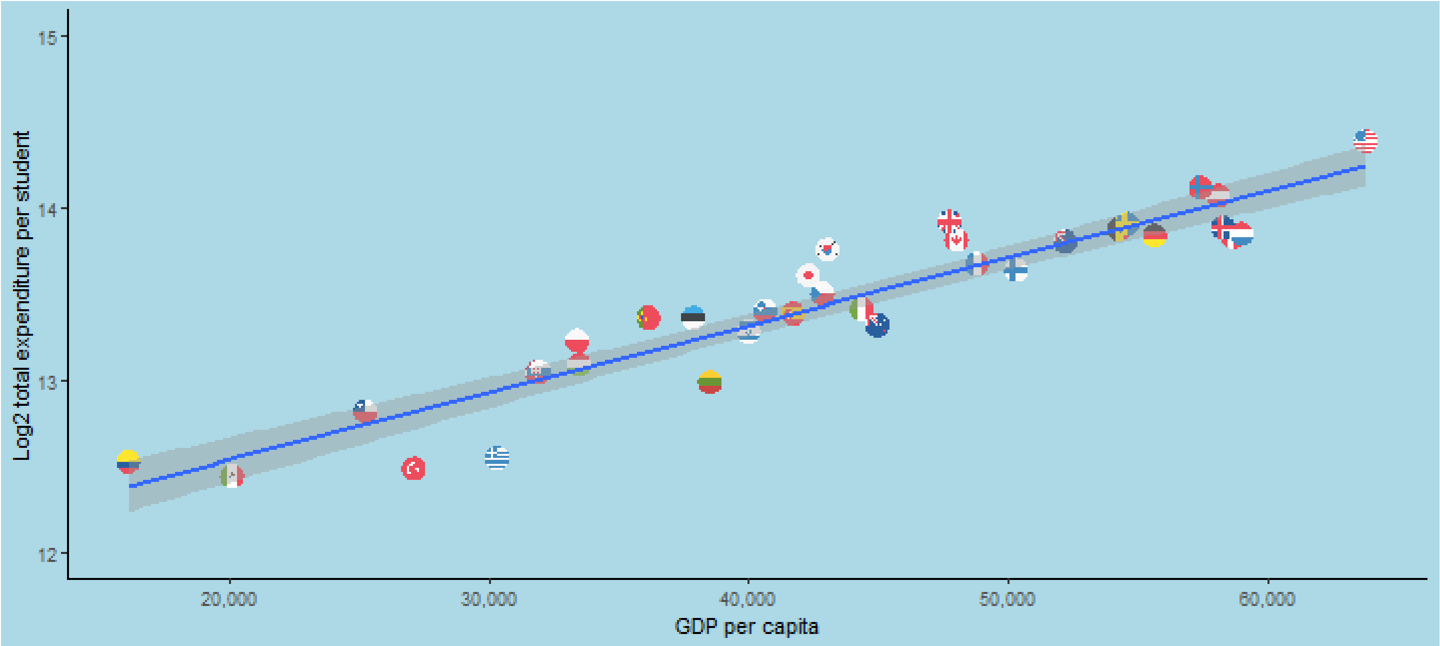
# Regress $\log_2(ppe)$ on gdp

```
summary(lm(log2(ppe) ~ gdp, oecd))
```

```
...
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39728 -0.09378  0.01867  0.11920  0.31357
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.176e+01  1.113e-01   105.7   <2e-16 ***
## gdp         3.899e-05  2.484e-06    15.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1712 on 32 degrees of freedom
## Multiple R-squared:  0.8851,    Adjusted R-squared:  0.8815
## F-statistic: 246.5 on 1 and 32 DF,  p-value: < 2.2e-16
...
```

**Percent growth rate**: $100(2^{0.000039} - 1) = 0.0027\%$ ; for each $1 more of GDP per person, PPE is 0.0027% higher; or for each $1,000 more of GDP per person, PPE is 2.7% higher

# Interpreting log Y results




$$\log_2(\hat{PPE_j}) = 11.8 + 0.000039 * GDP_j$$

> Per capita gross domestic product (GDP) is a strong predictor of yearly per-student expenditure from primary through tertiary education. In particular, if we compare two countries whose GDPs differ by $1,000, we would predict that the wealthier country would have per pupil expenditure that is 2.7 ***percent*** higher than the country with the smaller economy.

# Log-log transformations

## aka proportional growth

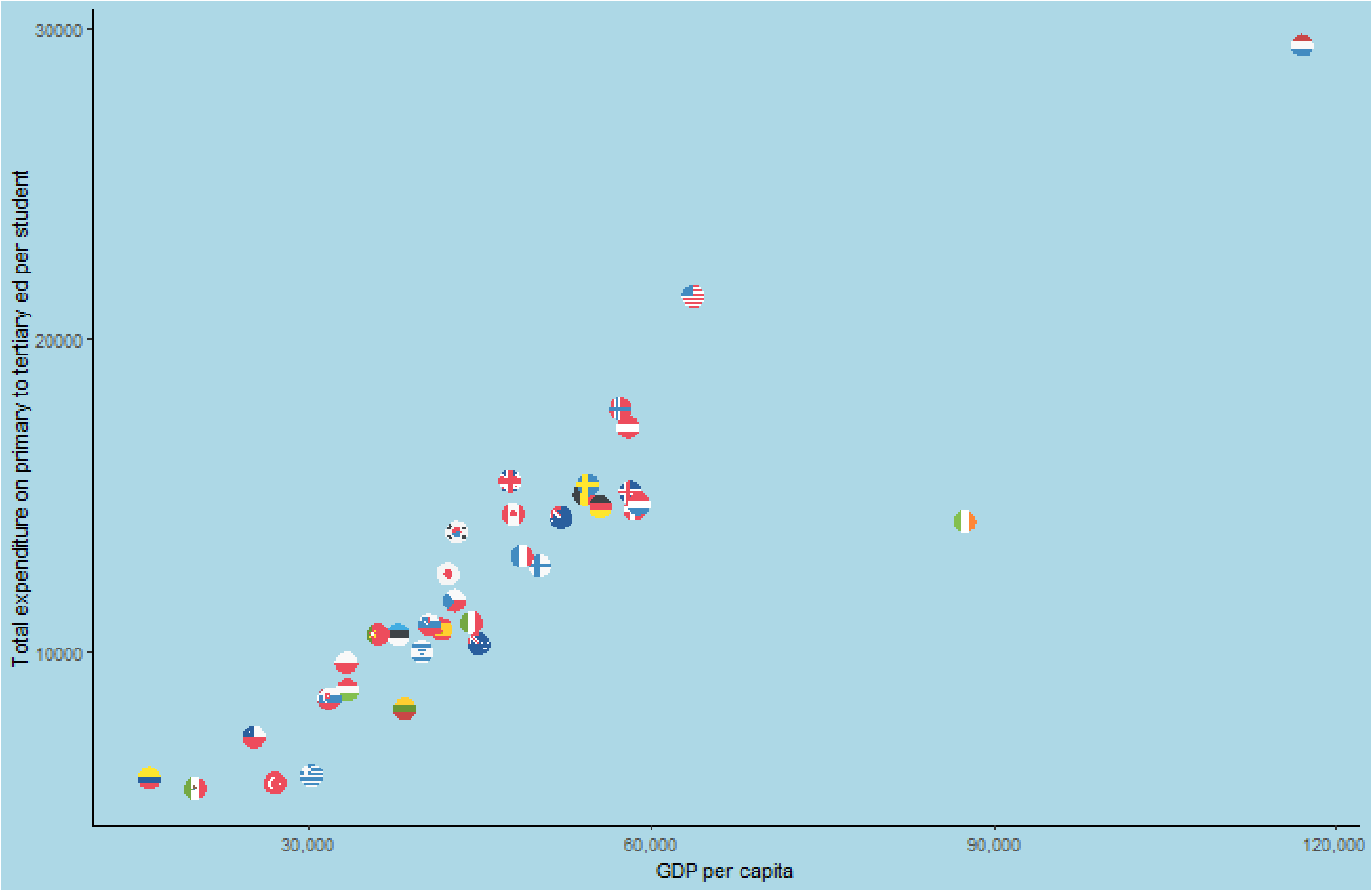
# Which 🌲 to harvest?

- Could theoretically select a log of any base to transform outcome or predictor or both to a linear relationship
- Much more sensible to restrict yourself to base_10, base_2 or the natural log; comes from Euler's number $(e)$

$$e = \lim_{n\to\infty}(1 + \frac{1}{n})^n \approx 2.718281828459...$$

- Natural log: $\log_{2.718...}(x) = \log_e(x) = \ln(x)$

# All the countries

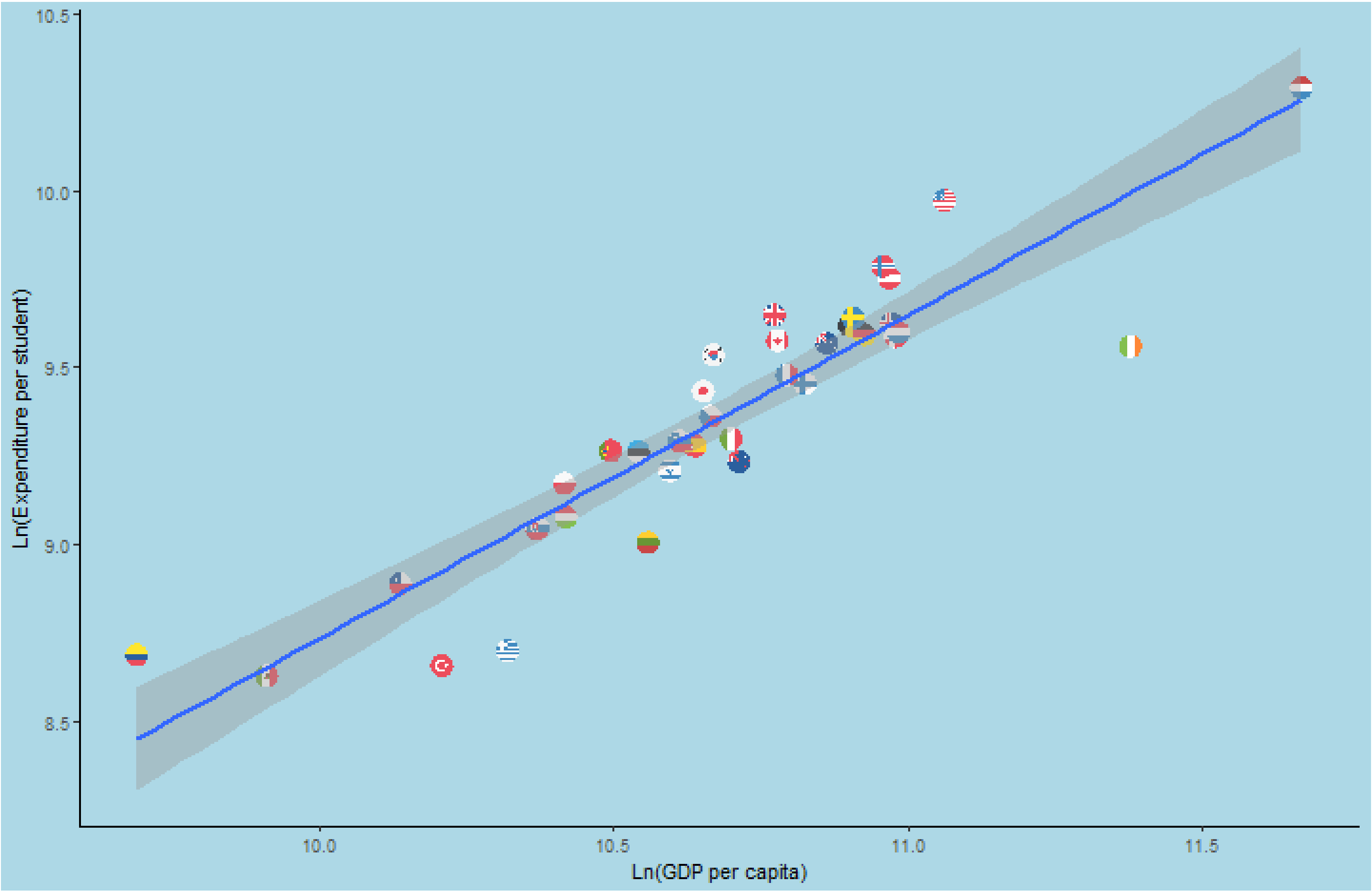


# Log-log transformations

```
oecd2$lngdp <- log(oecd2$gdp)
oecd2$lnppe <- log(oecd2$ppe)

ln_ppe <- ggplot(oecd2, aes(x=lngdp, y=lnppe))
```

# Log-log transformations



# Regress $\ln(ppe)$ on $\ln(gdp)$

```
summary(lm(log(ppe) ~ log(gdp), oecd2))
```

```
...
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43570 -0.04076  0.01302  0.07489  0.26542
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.39273    0.72674   -0.54    0.592
## log(gdp)     0.91274    0.06801   13.42 3.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1509 on 34 degrees of freedom
## Multiple R-squared:  0.8412,    Adjusted R-squared:  0.8365
## F-statistic: 180.1 on 1 and 34 DF,  p-value: 3.826e-15
...
```

$$Ln\hat{P}PE_j = -0.39 + 0.91 * LnGDP_j$$

# Interpreting this

Can interpret log-log relationships in percent terms. $\beta_1$ **represents the % change in Y per 1% change in X.**

**Postulated model:**

- $Y = \beta_0 X^{\beta_1} e^{\varepsilon}$
- $\ln(Y) = \ln(\beta_0 X^{\beta_1} e^{\varepsilon})$
- $\ln(Y) = \ln(\beta_0) + \ln(X^{\beta_1}) + \ln(e^{\varepsilon})$
- $\ln(Y) = \ln(\beta_0) + \beta_1 \ln(X) + \varepsilon$

**Imagine `(Y_1)` and `(Y_2)` are 1% (or 0.01) apart:**

- $Y_1 = \beta_0 X^{\beta_1}$
- $Y_2 = \beta_0 (1.01X)^{\beta_1} = \beta_0 X^{\beta_1} (1.01)^{\beta_1}$
- $\frac{Y_2}{Y_1} = \frac{\beta_0 X^{\beta_1}}{\beta_0 X^{\beta_1}} = (1.01)^{\beta_1}$

So $Y_2$ is $(1.01)^{\beta_1}$ times larger than $Y_1$

Regress ln(Y) on ln(X) and the slope estimate is the estimated percent difference in Y per 1 percent difference in X

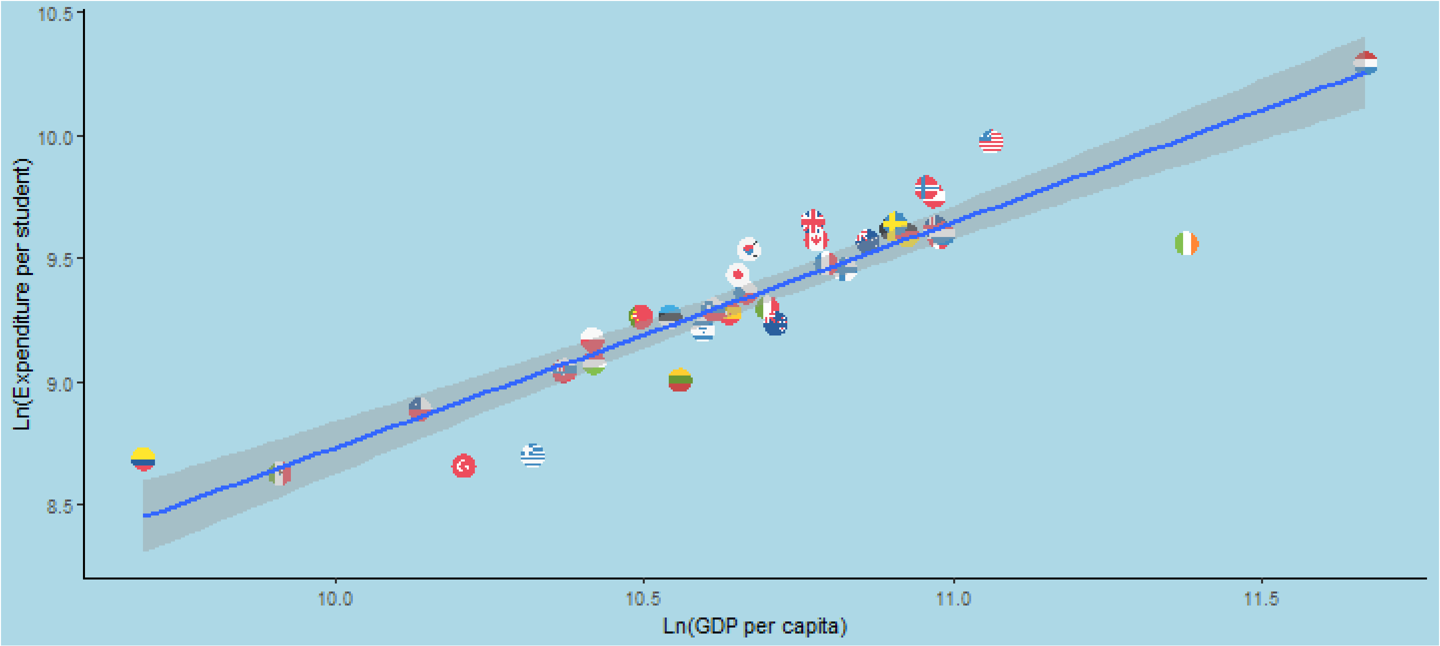
# Interpret log-log relationship

```
summary(lm(log(ppe) ~ log(gdp), oecd2))
```

```
...
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43570 -0.04076  0.01302  0.07489  0.26542
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.39273    0.72674   -0.54    0.592
## log(gdp)     0.91274    0.06801   13.42 3.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1509 on 34 degrees of freedom
## Multiple R-squared:  0.8412,    Adjusted R-squared:  0.8365
## F-statistic: 180.1 on 1 and 34 DF,  p-value: 3.826e-15
...
```

"1 percent change in GDP predicts 0.91 percent change in PPE"

# Interpret log-log relationship

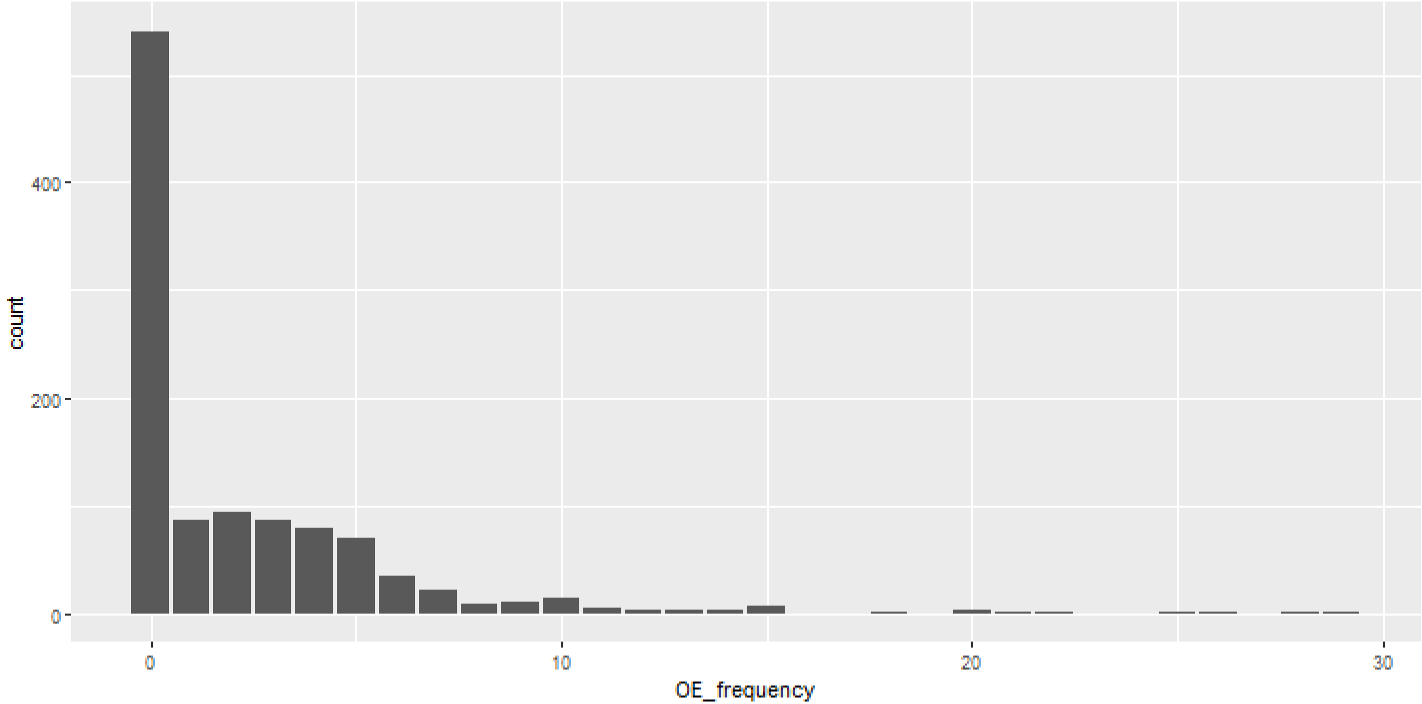



$$\ln(P\hat{P}E_j) = \ln(\beta_0) + \beta_1 \ln(GDP_j) + \varepsilon$$

> We predict that, on average, comparing two countries with GDP per capita separated by 1 percent the wealthier country will spend 0.91 percent more on its pupils across primary through tertiary education.

# "Forbidden" log transformations

So far, we've been dealing with situations in which all the variables we needed to transform were non-zero. In fact this is often not the case:




Many other instances: counts of behaviors, individual income, absences, scale scores, etc.

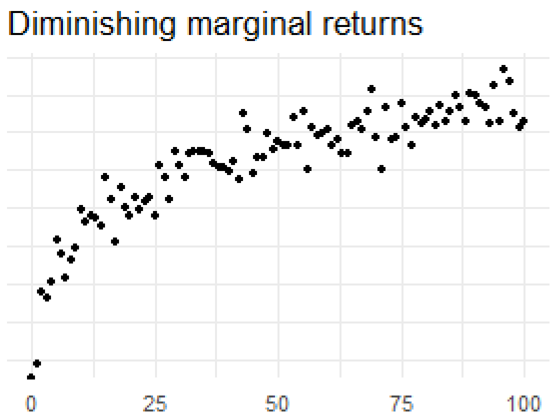
# "Forbidden" log transformations

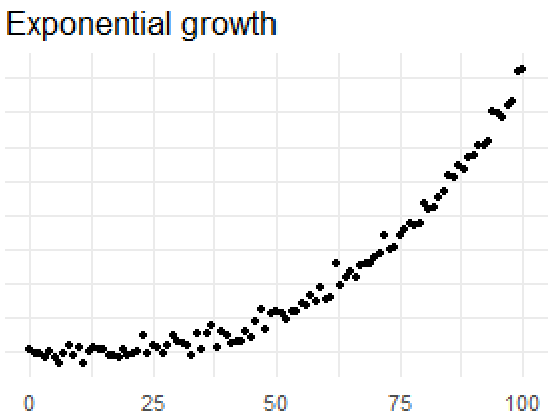
Traditional approach:

- Add a small "starter" value to all raw values (+1, +0.1, +0.01, +0.001, etc.)
- Take log of this "zero-inflated" variable

DO NOT DO THIS!!!

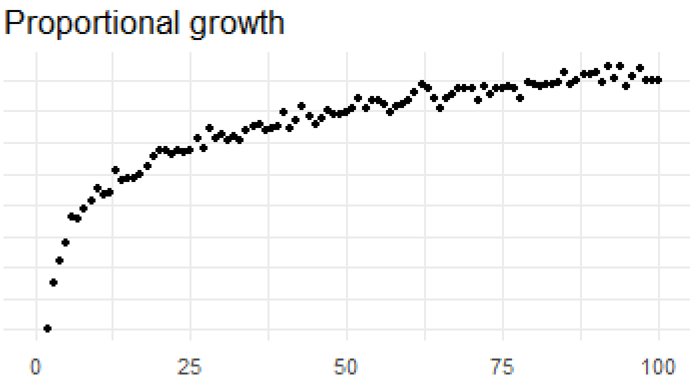
- Value selected for starter and proportion of 0s in your data can results in wildly inconsistent coefficient estimates
- You'll address this issue in EDUC 645 with Poisson regression




- Regress Y on log(X)
- $Y = \hat{\beta}_0 + \hat{\beta}_1 \log(X)$
- "every doubling (or whatever base) of X associated with $\hat{\beta}_1$ diff in Y"

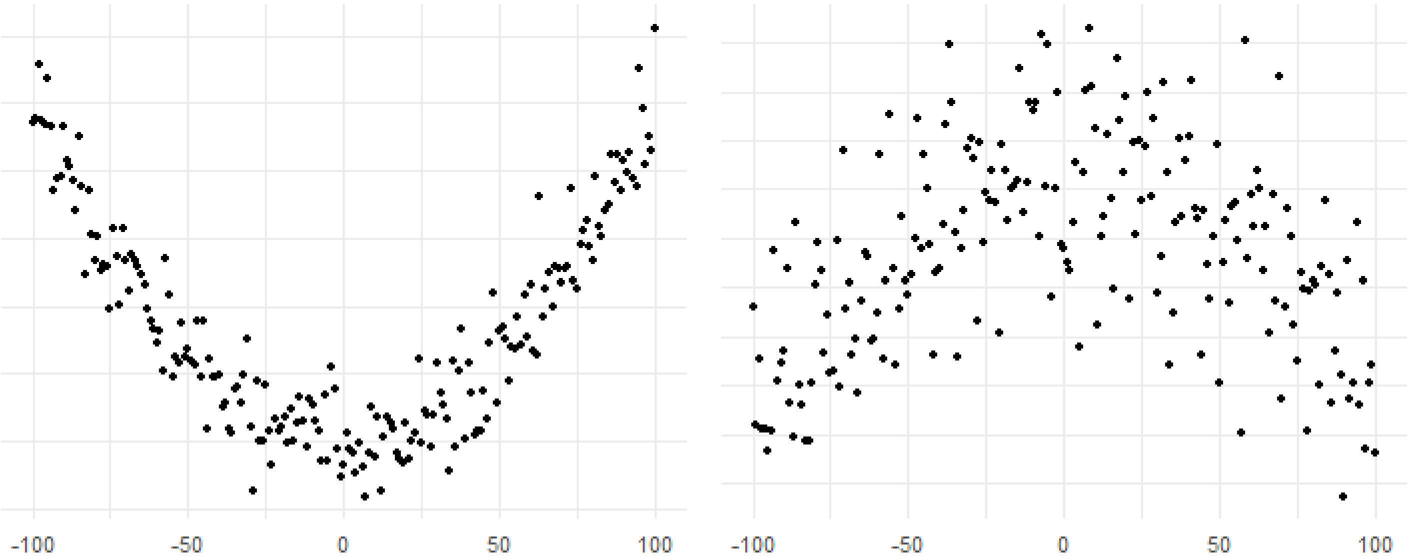
- Regress log(Y) on X
- $\log(Y) = \hat{\beta}_0 + \hat{\beta}_1 X$
- Every 1 unit diff in X associated with $100(e^{\hat{\beta}_1} - 1)$ % diff in Y




- Regress log(Y) on log(X)
- $\log(Y) = \hat{\beta}_0 + \hat{\beta}_1 \log(X)$
- Every 1% diff in X associated with $\hat{\beta}_1$ percent diff in Y

# Quadratic terms: a special kind of interaction

# Quadratic model



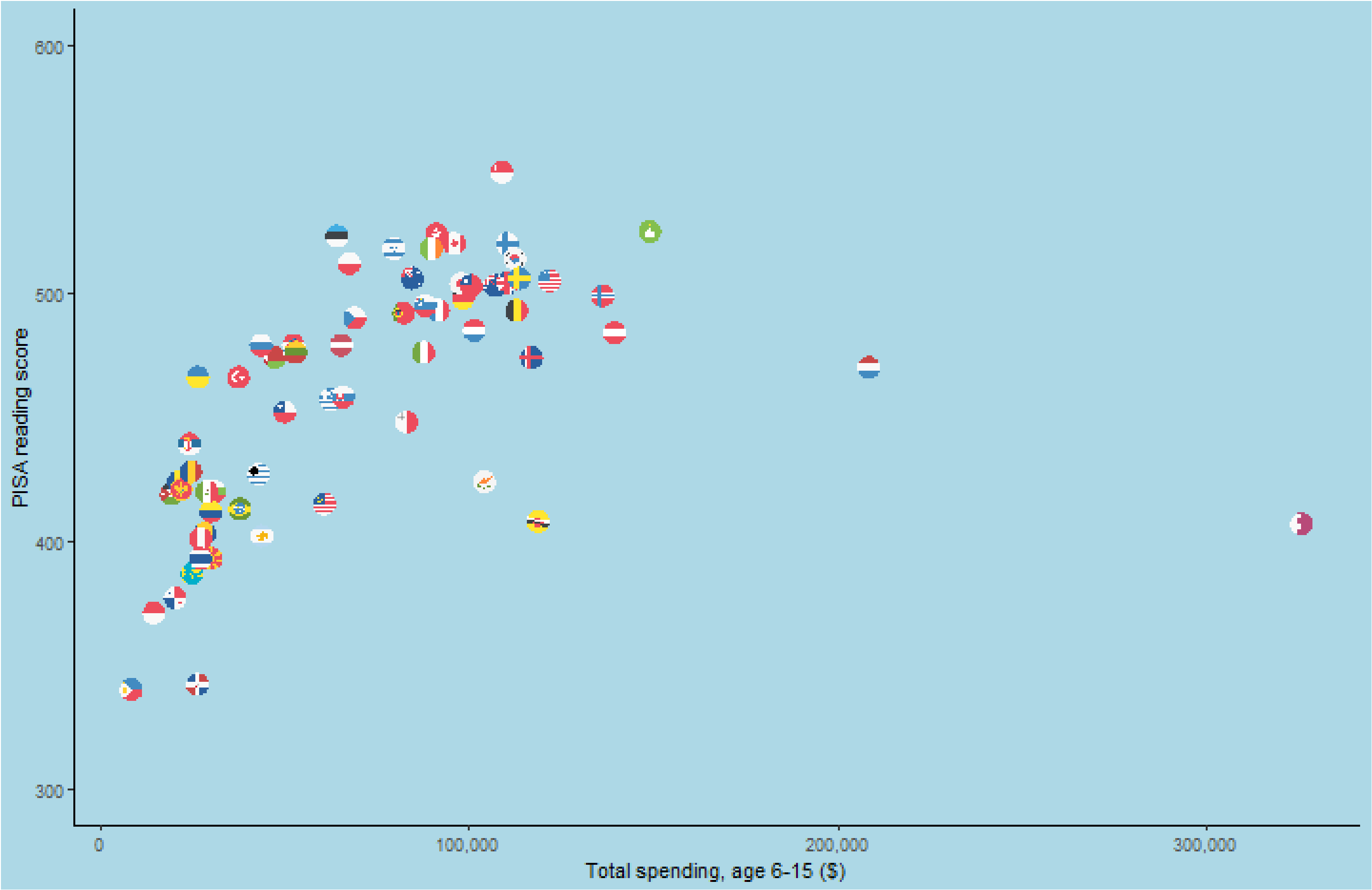

Effects of a predictor can differ by that predictor:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1 * X_1) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

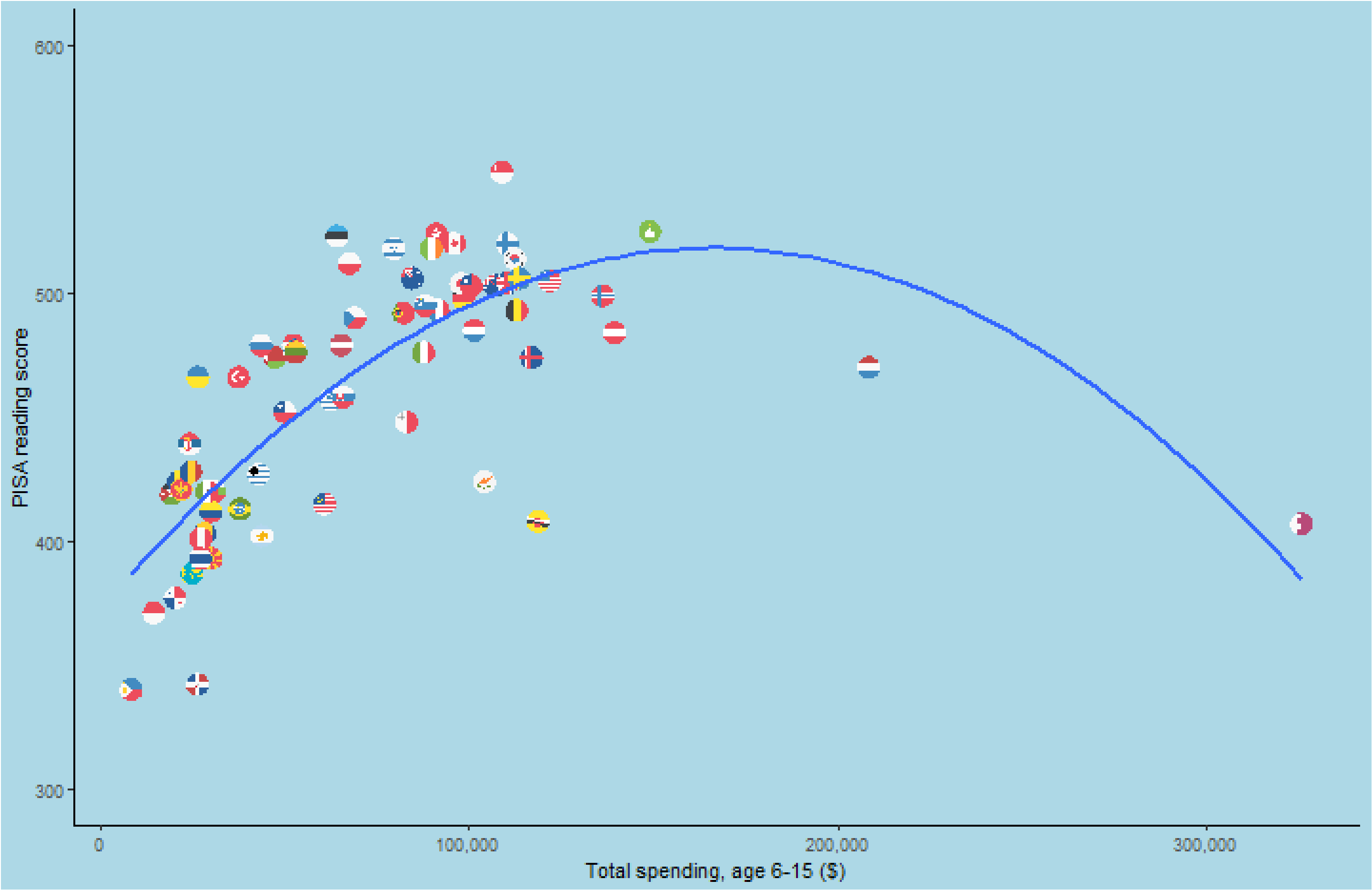
Can point upwards or downwards, but **all quadratic relationships are non-monotic; the relationship both rises and falls (or falls and rises)**

# A quadratic relationship

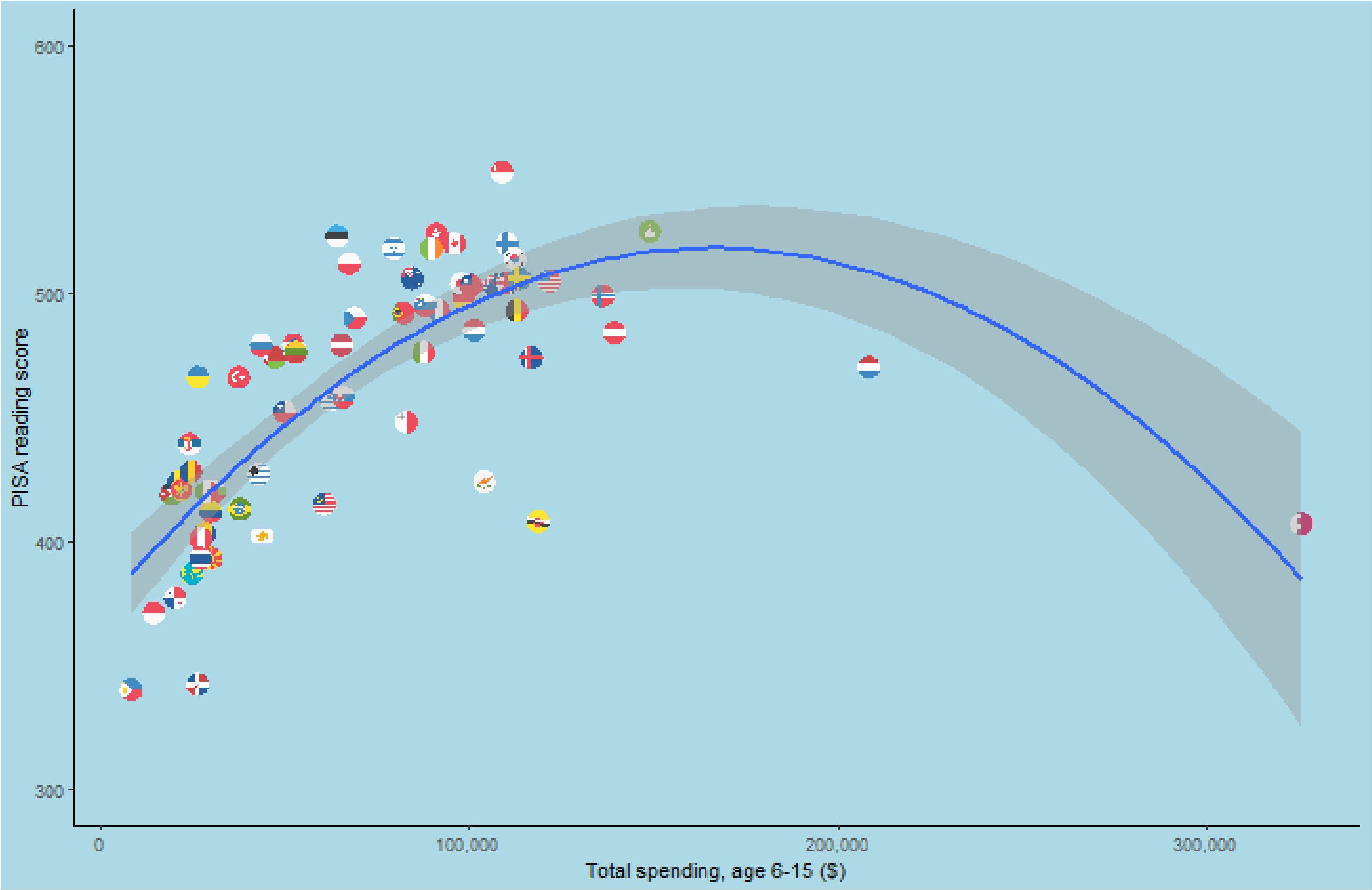



**Which direction will the quadratic line of best fit point?**

# A quadratic relationship



# A quadratic relationship




We can represent quadratic fits mathematically in generic form: $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

**Challenge: what signs will each of the three coefficients take for the above relationship?**

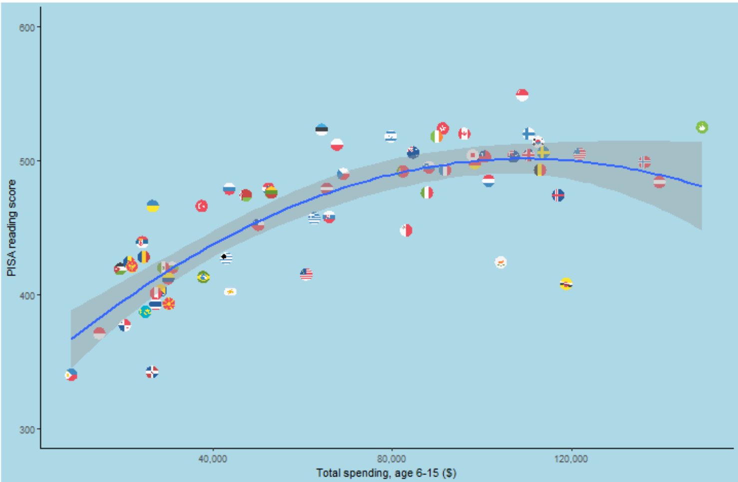
# Fitting the quadratic

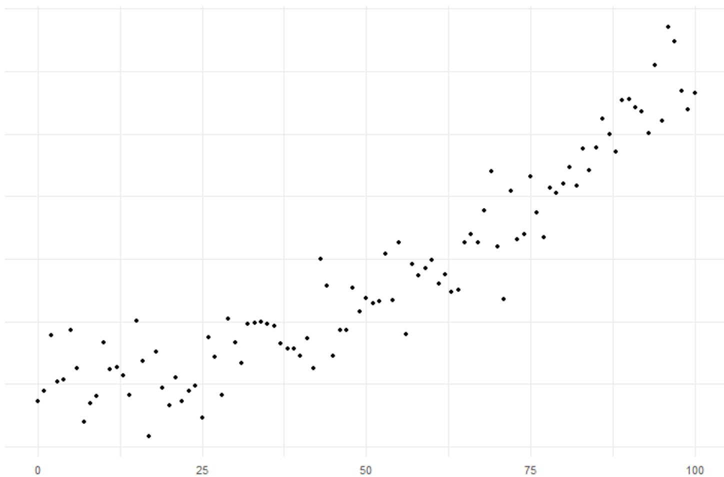
```
summary(lm(read_score ~ total_spending + I(total_spending^2), pisa))
```

```
...
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -98.511 -15.722   3.806  22.651  59.394 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          3.728e+02  9.665e+00  38.574  < 2e-16 ***
## total_spending       1.750e-03  1.798e-04   9.732 4.22e-14 ***
## I(total_spending^2) -5.260e-09  6.498e-10  -8.096 2.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 31.34 on 62 degrees of freedom
## Multiple R-squared:  0.6117,    Adjusted R-squared:  0.5992 
## F-statistic: 48.84 on 2 and 62 DF,  p-value: 1.834e-13
...
```

Fitted equation: $\hat{read} = 372.8 + 0.00175 * spend - 0.00000000526 * spend^2$. **How do our model fit statistics compare to the linear version?**

# The "right" fit to data




- A declining relationship between spending and performance doesn't make much substantive sense, so we would probably not use a quadratic fit for our full data
- However, without Qatar and Luxembourg, a quadratic describes the relationship quite nicely

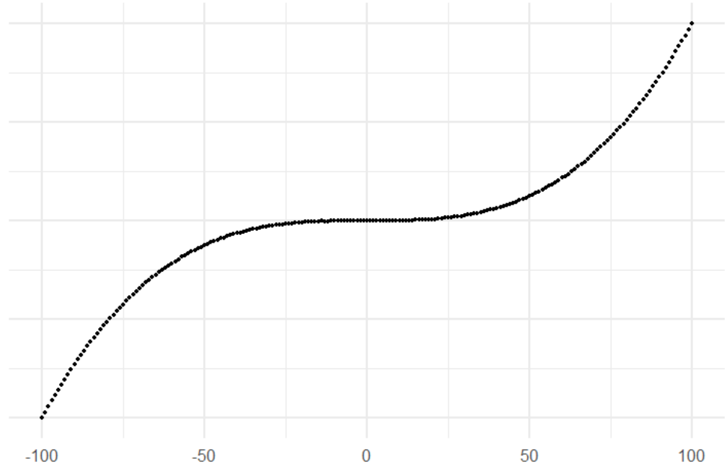
- Don't extrapolate the shape of the parabola to the left of the y-axis
- Shouldn't assume the y values will be higher to the left of the y-axis

# Higher-order polynomials

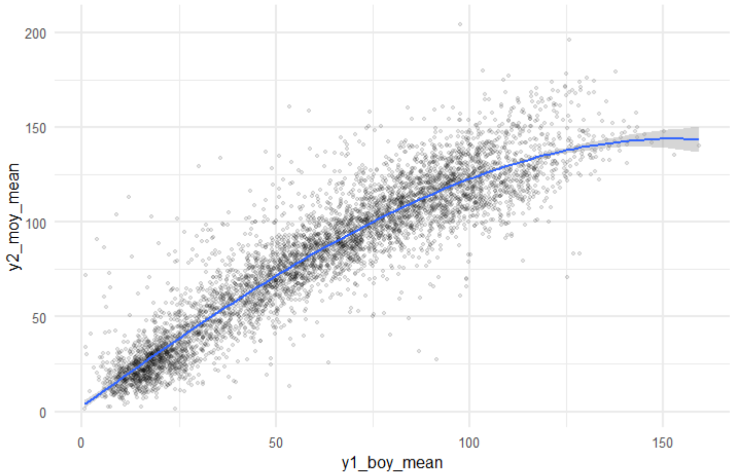
# Cubics

We needn't restrict ourselves to transformations to normality to only quadratic relationships. Many relationships, for example are cubic (third-power) in nature. Particularly true when there are measurement issues in the tails and/or floor/ceiling effects.

### Strong cubic




### Our DIBELS data




$$W20\hat{\_}ORF = 2.81 + 1.47 * F19\_ORF - 0.0010 * F19\_ORF^2 - 0.000017 * F19\_ORF^3$$

# Other approaches

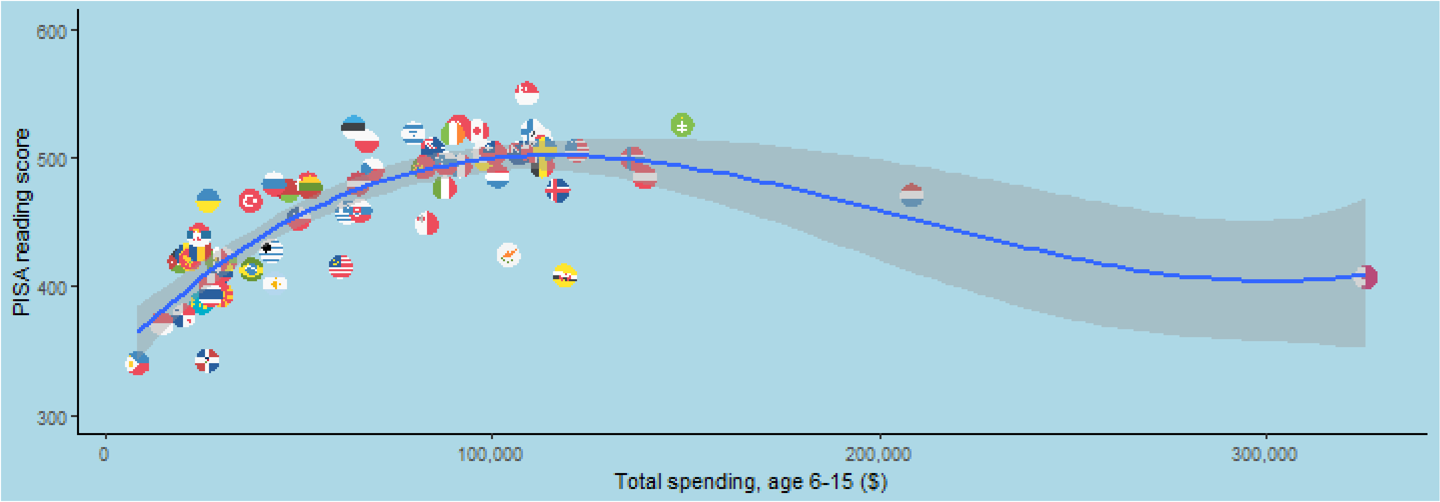
There are an infinite number of potentially effective transformations:

- Squares, cubes, quartic, quintics, …
- Square roots, cube roots, fourth roots, …
- Logarithms (of any base), antilogarithms
- Inverses
- Trigonometric functions
- Hyperbolic functions
- Combinations of above…

Approaches to achieve local linearity:

- Splines
- Local estimated scatterplot smoothing (LOESS)

**Some emerging issues:**



# Synthesis and wrap-up

# Different approaches

**Empirical approach**

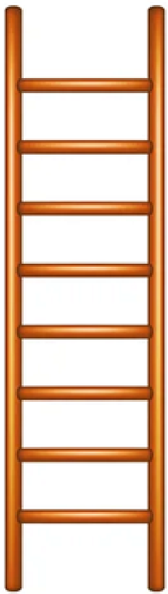
- Notice presence of non-linearity in relationship
- Find an *ad-hoc* transformation of either the predictor, the outcome, or both that renders the relationship linear
- Use OLS in the transformed world, and conduct inference there
- De-transform fitted model to produce sensible plots

**Theory-driven approach**

- Use theory or knowledge from prior research to postulate a non-linear model
- Use non-linear regression (`nls` or other estimation packages) (part of the Generalized Linear Model family) to fit the postulated trend in the real world and conduct inference there
- Interpret parameter estimates directly
- **We are not learning how to do this, but worth exploring yourself**

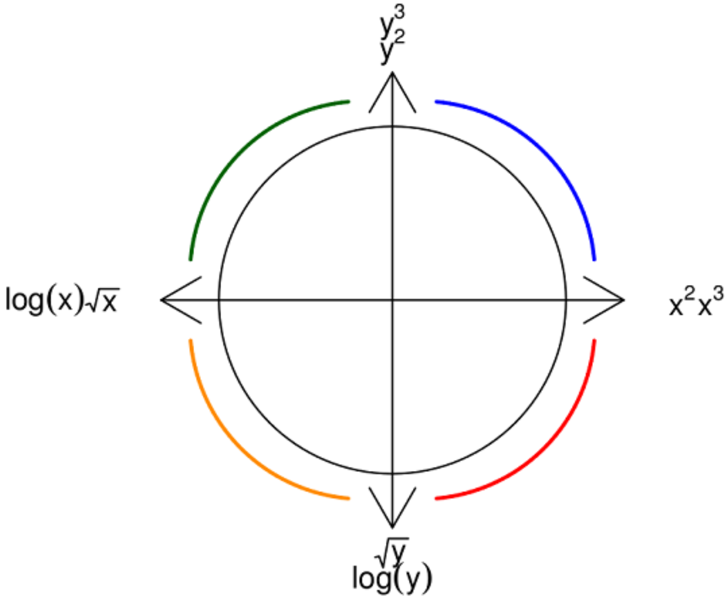
# The Ladder and the Bulge

## Tukey's Ladder



| Power | Transformation | Name |
|---|---|---|
| 3 | $x^3$ | Cubic |
| 2 | $x^2$ | Quadratic |
| 1 | $x$ | Untransformed |
| ½ | $x^{1/2} = \sqrt{x}$ | Square root |
| "0" | $\log(x)$ | Logarithm |
| -1/2 | $x^{-1/2} = 1/\sqrt{x}$ | Reciprocal root |
| -1 | $x^{-1} = 1/x$ | Reciprocal |
| -2 | $x^{-2} = 1/x^2$ | Reciprocal square |
| -3 | $x^{-3} = 1/x^3$ | Reciprocal cubic |

## Tukey's Bulge



# Putting non-linearity together

- **Remember to check your linearity assumption**
  - Use bivariate scatter plots
  - Use residual and Q-Q plots to diagnose
- **Make sensible transformations**
  - Logarithmic, inverse, root and other functions can allow a return to a world of linearity and permit you to use the GLM tools of OLS to estimate non-linear relationships
  - Best to use transformations that are the most straightforward to interpret
  - Use Tukey's Bulge to guide what kind of transformation you will attempt
  - There is no one "right" transformation for a given data shape
  - Start with transforming x before y
  - Generally, do **not** use a "start" to log transform data that includes 0s
  - Inspect scatter plots post-transformation to check for success in linearizing
    - With large data, can be hard to see; consider binscatter options (by hand or `binsreg`)
- **Predictors can interact with themselves**
  - Quadratic and cubic models provide a flexible strategy for fitting non-linear models, especially those that cannot be linearized by logarithms
  - Be careful about overfitting and model instability with polynomials of order >3!
  - Quadratics and logs will often produce similar fitted lines; quadratic allows direct statistical test for non-linearity, logarithm may fit with theory better and/or can be more readily interpretable

# Goals of the unit

- Describe in writing and verbally the assumptions we violate when we fit a non-linear relationship with a linear model
- Transform non-linear relationships into linear ones by using logarithmic scales
- Estimate regression models using logarithmic scales and interpret the results
- Estimate and interpret models with quadratic and higher-order polynomial terms (special kinds of interactions)
- Select between transformation options

# To-Dos

## Assignment 5:

- Due March 8, 11:59p

## Final

- Due March 23, 12:01p

## Re- (late) submissions

- Everything due March 17, 9:00a (no exceptions)
- Assignments with scores <10.8 only
- Earn up to 10.8

# Log vs. quadratic