Today's Agenda: Dataset 1

Extending the OLS Regression

1. Week 9: Dichotomous and categorical predictors
2. Today: Transforming the variables
3. Thursday: Transforming the model

Justin Leinaweaver (Spring 2022)

Let's Practice with Categorical Predictors

Regress GDP (millions) on the three population level categories in pop_category

Let's Practice with Categorical Predictors

Regress GDP (millions) on the three population level categories in pop_category

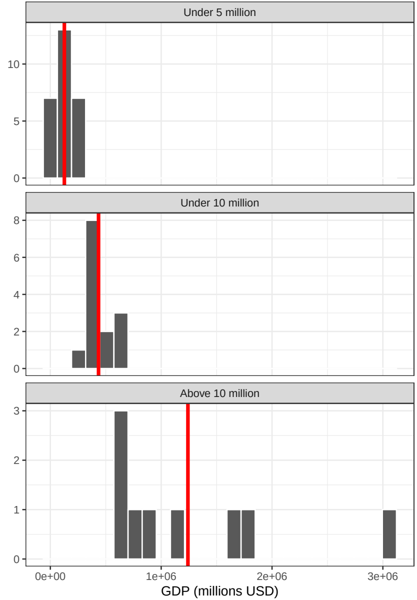
Create Dummy Vars:

1. Under 10 million
2. Above 10 million

Group Means

Population	GDP (millions)
Under 5 million	126,486
Under 10 million	434,881
Above 10 million	1,241,183

	GDP (millions)
Pop (5-10million)	308,395.20* (115,178.90)
Pop (Above 10million)	1,114,697.00* (134,609.30)
Constant	126,485.50 (67,304.65)
Observations	50
Adjusted R^2	0.58
Residual Std. Error	349,725.20 (df = 47)
F Statistic	34.32* (df = 2; 47)
Note:	*p<0.05



	GDP (millions)
Pop (5-10million)	308,395.20*
	(115,178.90)
Pop (Above 10million)	1,114,697.00*
	(134,609.30)
Constant	126,485.50
	(67,304.65)
Observations	50
Adjusted R^2	0.58
Residual Std. Error	349,725.20 (df = 47)
F Statistic	34.32* (df = 2; 47)
Note:	*p<0.05

Improving Model Fit: Transforming Variables

Do states with more educated workforces have larger economies?

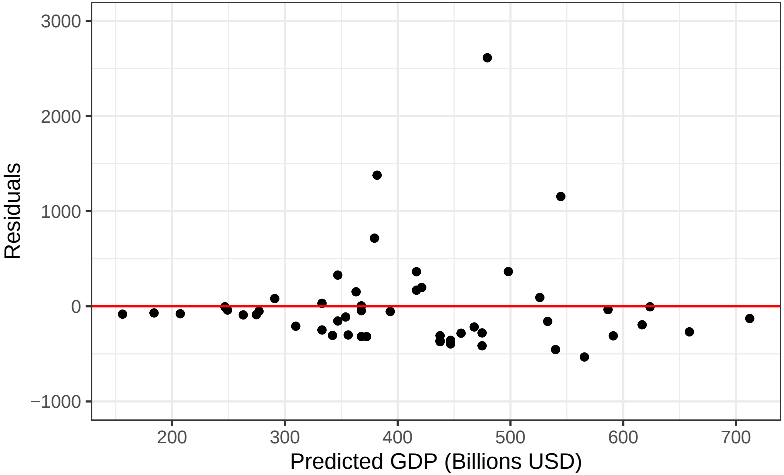
Model 1: Regress GDP (**millions**) on bachelors

Model 2: Regress GDP (**billions**) on bachelors

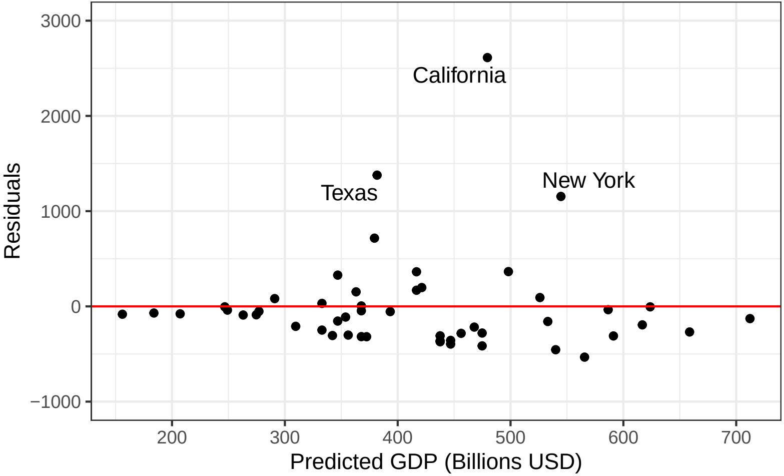
	GDP (millions)	GDP (billions)
	(1)	(2)
Bachelors (%)	23,271.42	23.27
	(14,124.28)	(14.12)
Intercept	−335,020.10	−335.02
	(460,391.60)	(460.39)
Observations	50	50
Adjusted R^2	0.03	0.03
Residual Std. Error (df = 48)	528,114.80	528.11
F Statistic (df = 1; 48)	2.71	2.71
Note:		*p<0.05

	GDP (millions)	(billions)	(10 billion)	(100 billion)
	(1)	(2)	(3)	(4)
Bachelors (%)	23,271.42	23.27	2.33	0.23
	(14,124.28)	(14.12)	(1.41)	(0.14)
Intercept	−335,020.10	−335.02	−33.50	−3.35
	(460,391.60)	(460.39)	(46.04)	(4.60)
Observations	50	50	50	50
Adjusted R^2	0.03	0.03	0.03	0.03
Residual Std. Error (df = 48)	528,114.80	528.11	52.81	5.28
F Statistic (df = 1; 48)	2.71	2.71	2.71	2.71
Note:				*p<0.05

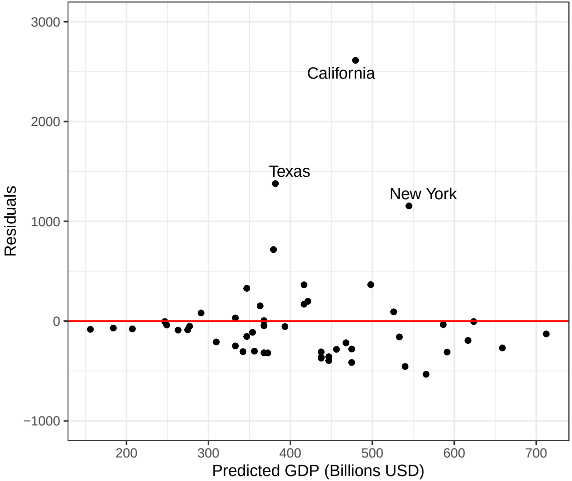
Transformation 1: Shift the Decimal Point



Transformation 1: Shift the Decimal Point



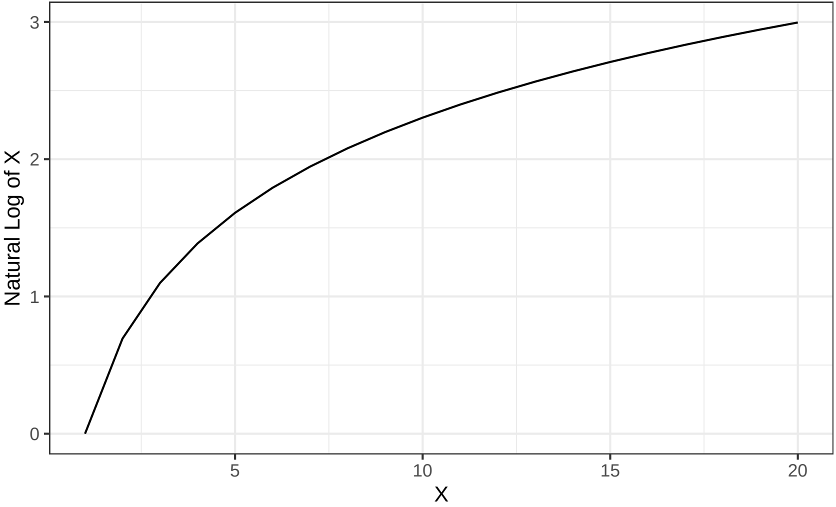
Transformation 1: Shift the Decimal Point



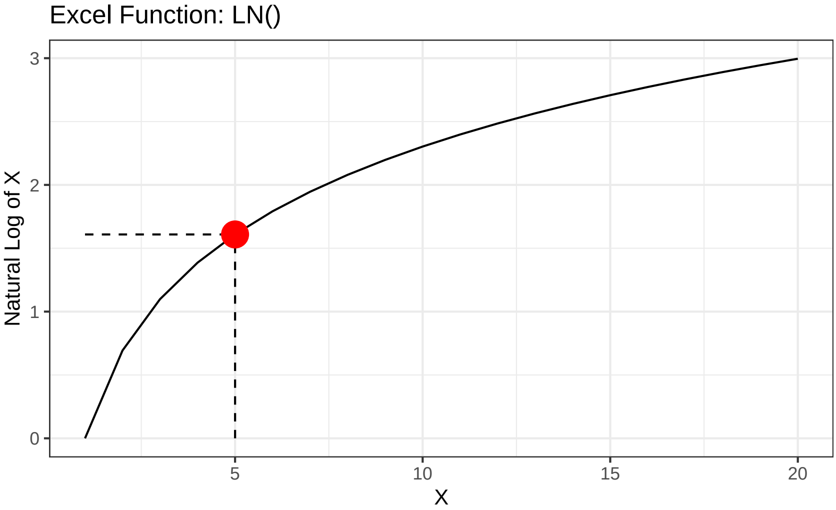
36 / 50 (72 %) are below the zero line.

Aim is 50%-50%

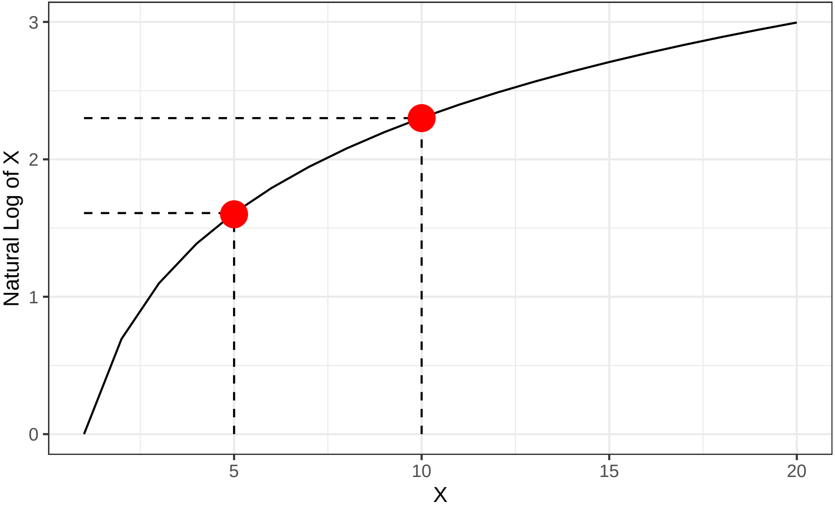
Transformation 2: Natural Logarithms



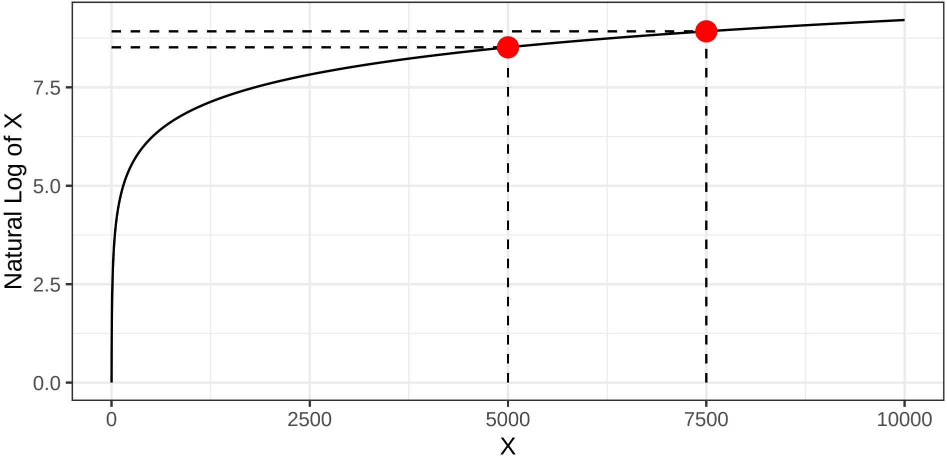
Transformation 2: Natural Logarithms



Transformation 2: Natural Logarithms



Transformation 2: Natural Logarithms



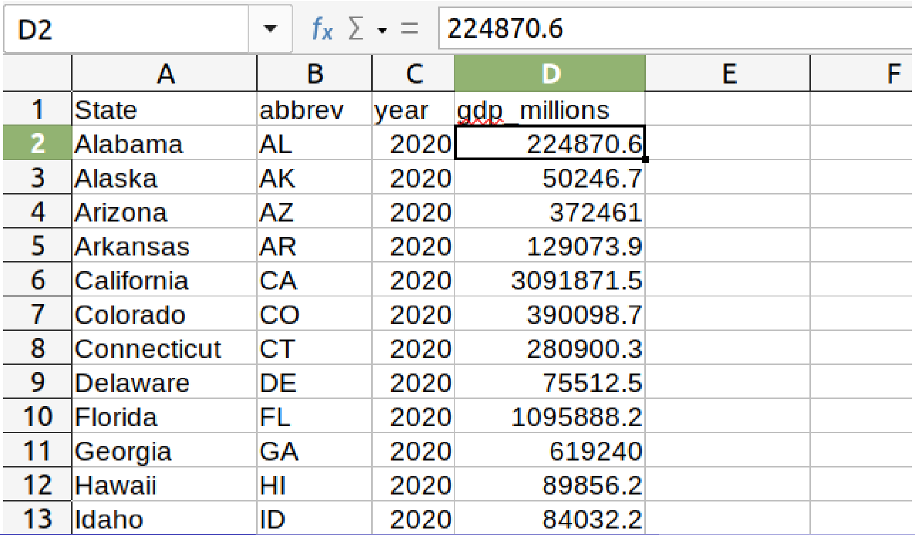
Transformation 2: Natural Logarithms

The natural log scale = multiplying by e

- e is Euler's Number (2.718282...)
- Typically written as log_e X or ln X

Transform back to linear scale using e^X

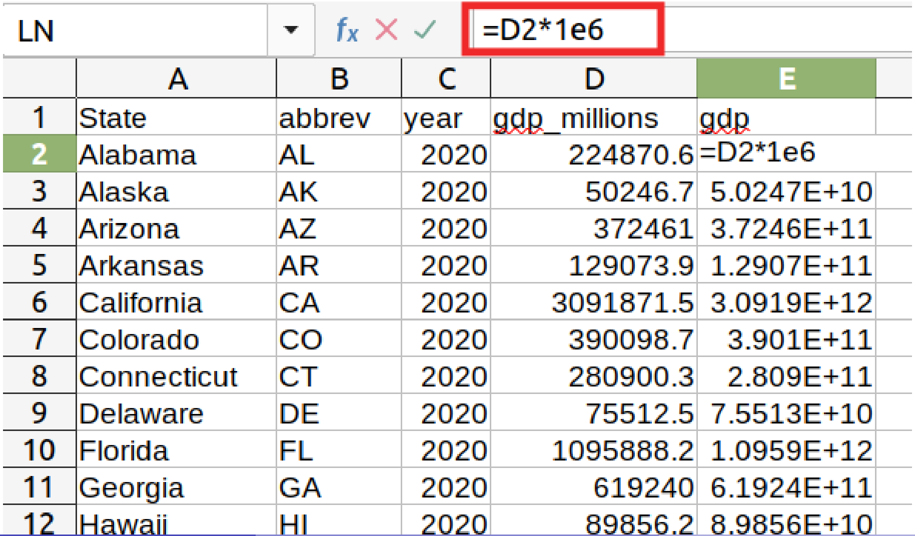
Transformation 2: Natural Logarithms



D2 224870.6

	A	B	C	D	E	F
1	State	abbrev	year	gdp_millions		
2	Alabama	AL	2020	224870.6		
3	Alaska	AK	2020	50246.7		
4	Arizona	AZ	2020	372461		
5	Arkansas	AR	2020	129073.9		
6	California	CA	2020	3091871.5		
7	Colorado	CO	2020	390098.7		
8	Connecticut	CT	2020	280900.3		
9	Delaware	DE	2020	75512.5		
10	Florida	FL	2020	1095888.2		
11	Georgia	GA	2020	619240		
12	Hawaii	HI	2020	89856.2		
13	Idaho	ID	2020	84032.2		

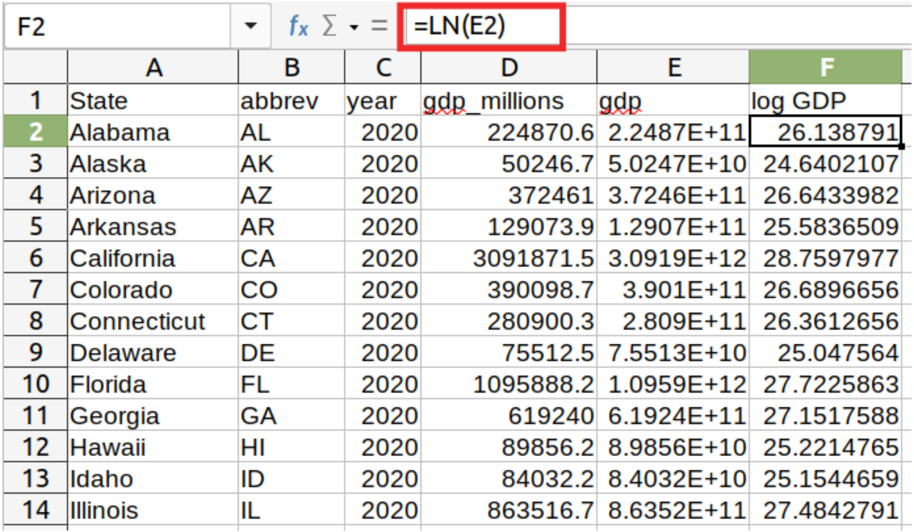
Transformation 2: Natural Logarithms



=D2*1e6

	A	B	C	D	E
1	State	abbrev	year	gdp_millions	gdp
2	Alabama	AL	2020	224870.6	=D2*1e6
3	Alaska	AK	2020	50246.7	5.0247E+10
4	Arizona	AZ	2020	372461	3.7246E+11
5	Arkansas	AR	2020	129073.9	1.2907E+11
6	California	CA	2020	3091871.5	3.0919E+12
7	Colorado	CO	2020	390098.7	3.901E+11
8	Connecticut	CT	2020	280900.3	2.809E+11
9	Delaware	DE	2020	75512.5	7.5513E+10
10	Florida	FL	2020	1095888.2	1.0959E+12
11	Georgia	GA	2020	619240	6.1924E+11
12	Hawaii	HI	2020	89856.2	8.9856E+10

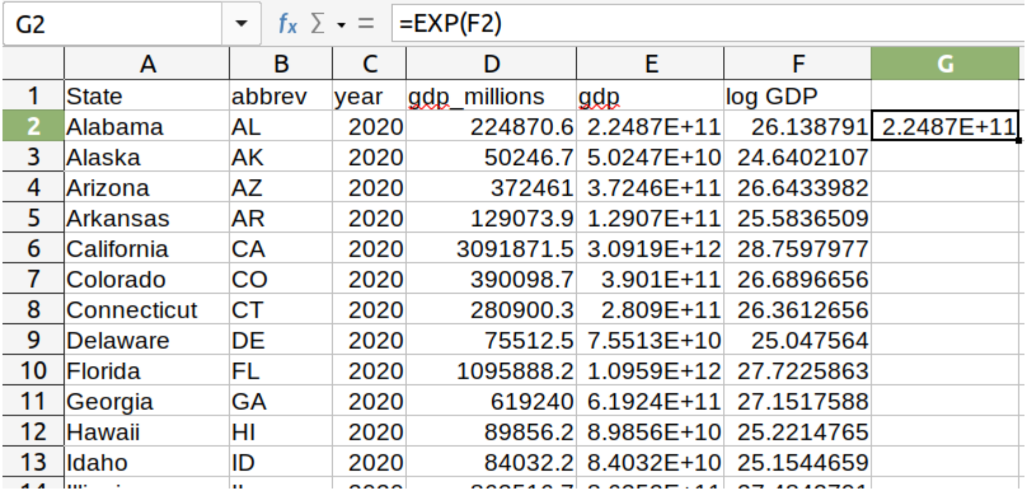
Transformation 2: Natural Logarithms



=LN(E2)

	A	B	C	D	E	F
1	State	abbrev	year	gdp_millions	gdp	log GDP
2	Alabama	AL	2020	224870.6	2.2487E+11	26.138791
3	Alaska	AK	2020	50246.7	5.0247E+10	24.6402107
4	Arizona	AZ	2020	372461	3.7246E+11	26.6433982
5	Arkansas	AR	2020	129073.9	1.2907E+11	25.5836509
6	California	CA	2020	3091871.5	3.0919E+12	28.7597977
7	Colorado	CO	2020	390098.7	3.901E+11	26.6896656
8	Connecticut	CT	2020	280900.3	2.809E+11	26.3612656
9	Delaware	DE	2020	75512.5	7.5513E+10	25.047564
10	Florida	FL	2020	1095888.2	1.0959E+12	27.7225863
11	Georgia	GA	2020	619240	6.1924E+11	27.1517588
12	Hawaii	HI	2020	89856.2	8.9856E+10	25.2214765
13	Idaho	ID	2020	84032.2	8.4032E+10	25.1544659
14	Illinois	IL	2020	863516.7	8.6352E+11	27.4842791

Transformation 2: Natural Logarithms



G2 | =EXP(F2)

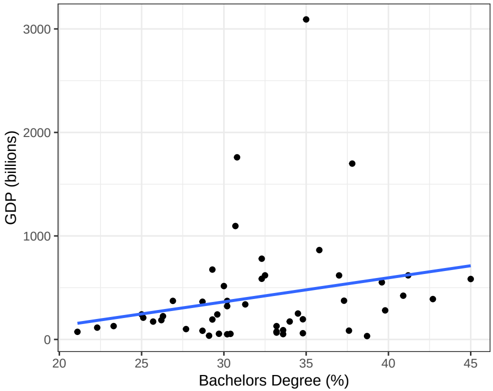
	A	B	C	D	E	F	G
1	State	abbrev	year	gdp_millions	gdp	log GDP	
2	Alabama	AL	2020	224870.6	2.2487E+11	26.138791	2.2487E+11
3	Alaska	AK	2020	50246.7	5.0247E+10	24.6402107	
4	Arizona	AZ	2020	372461	3.7246E+11	26.6433982	
5	Arkansas	AR	2020	129073.9	1.2907E+11	25.5836509	
6	California	CA	2020	3091871.5	3.0919E+12	28.7597977	
7	Colorado	CO	2020	390098.7	3.901E+11	26.6896656	
8	Connecticut	CT	2020	280900.3	2.809E+11	26.3612656	
9	Delaware	DE	2020	75512.5	7.5513E+10	25.047564	
10	Florida	FL	2020	1095888.2	1.0959E+12	27.7225863	
11	Georgia	GA	2020	619240	6.1924E+11	27.1517588	
12	Hawaii	HI	2020	89856.2	8.9856E+10	25.2214765	
13	Idaho	ID	2020	84032.2	8.4032E+10	25.1544659	

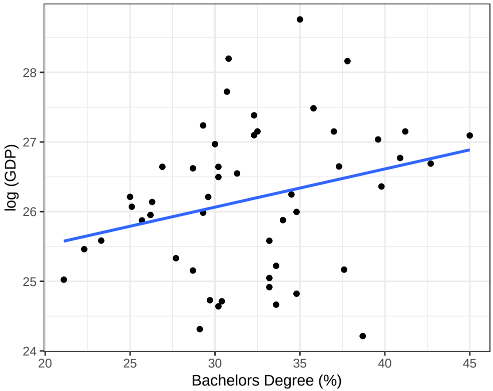
Let's Practice with Log Transformations

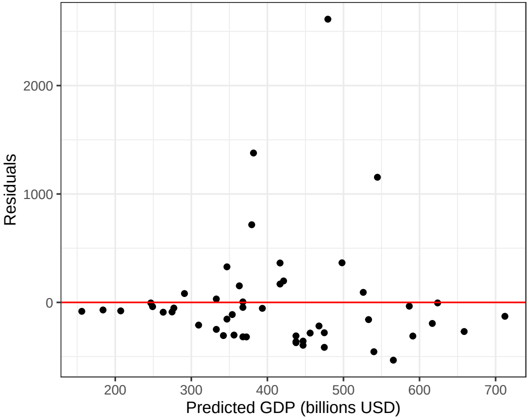
Do states with more educated workforces have larger economies?

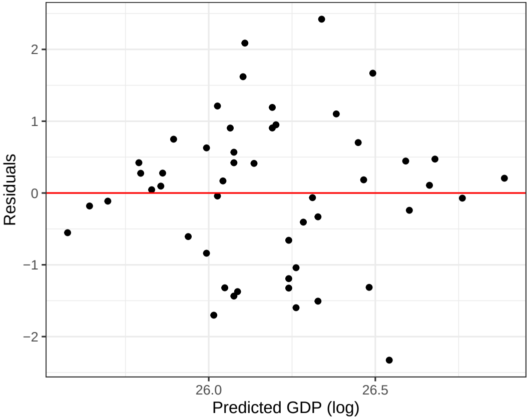
Model 3: Regress GDP (**log**) on bachelors

	GDP (billions)	GDP (log)
	(1)	(2)
Bachelors (%)	23.27	0.05*
	(14.12)	(0.03)
Intercept	−335.02	24.42*
	(460.39)	(0.91)
Observations	50	50
Adjusted R^2	0.03	0.06
Residual Std. Error (df = 48)	528.11	1.04
F Statistic (df = 1; 48)	2.71	3.86*
Note:		*p<0.056


GDP (billions)
3000
2000
1000
0
20
25
30
35
40
45
Bachelors Degree (%)


log (GDP)
28
27
26
25
24
20
25
30
35
40
45
Bachelors Degree (%)





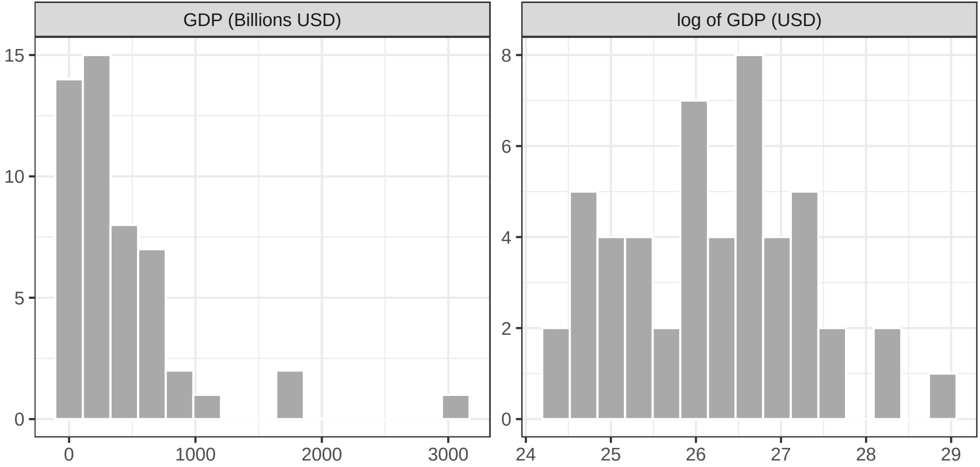
Above the line = 14 (28%)

Below the line = 36 (72%)

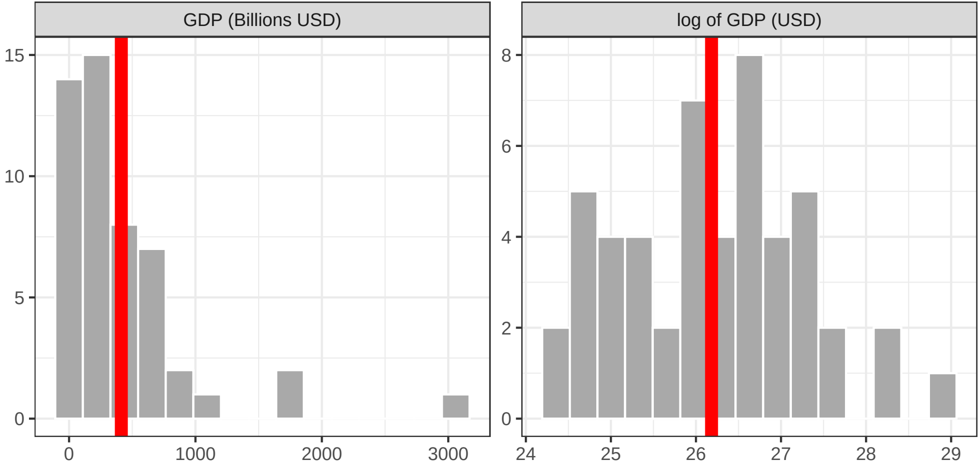
Above the line = 27 (54%)

Below the line = 23 (46%)

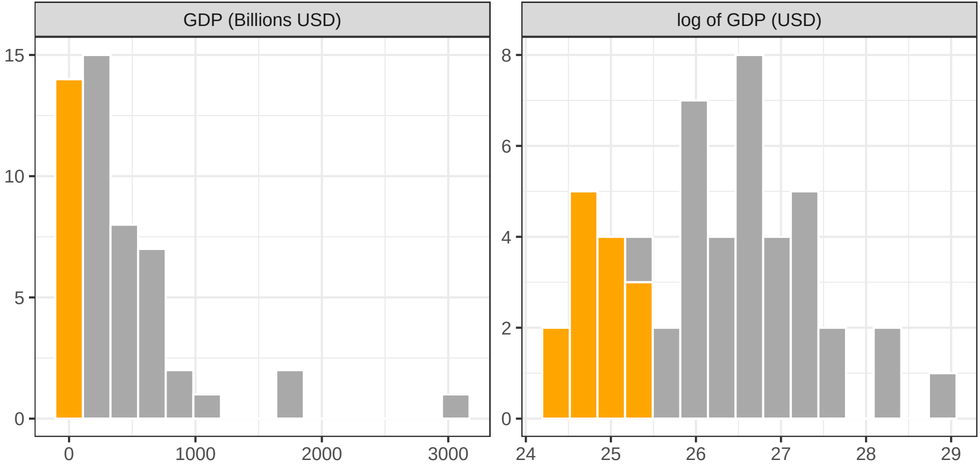
Transformation 2: Natural Logarithms



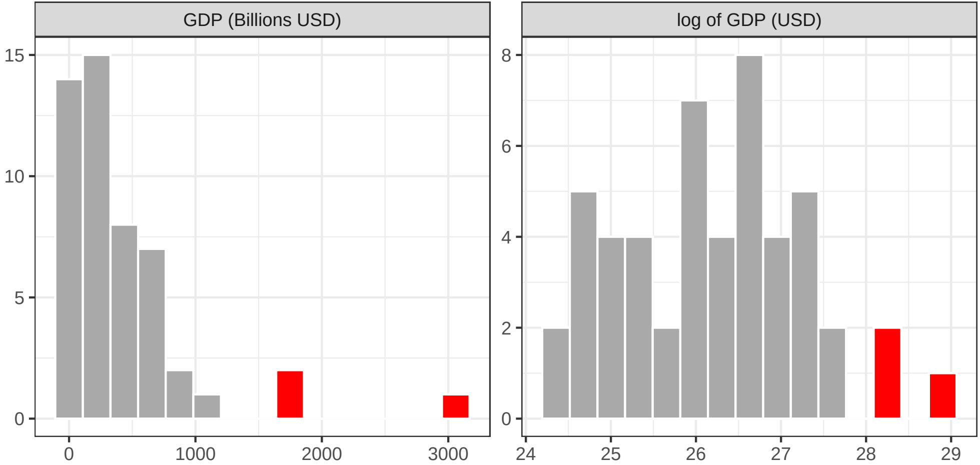
Transformation 2: Natural Logarithms



Transformation 2: Natural Logarithms



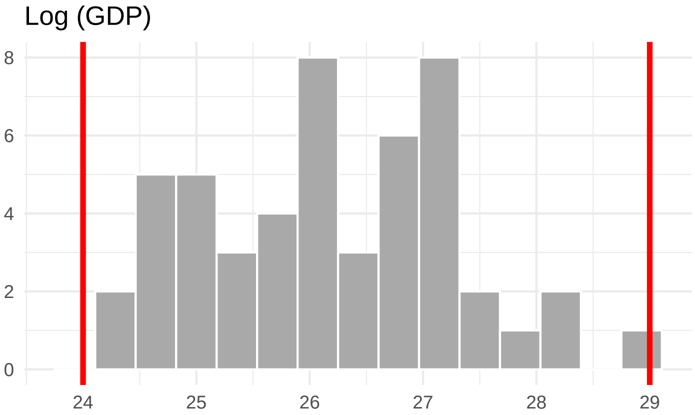
Transformation 2: Natural Logarithms



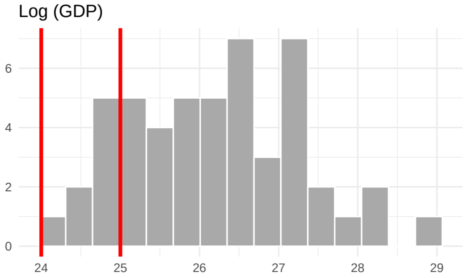
Transformation 2: Natural Logarithms

GDP for State X $=$ \$5,000 or 8.5 (ln)

- $log_e\ 5{,}000 \approx 8.5$
- $e^{8.5} \approx 5{,}000$



- e^{24} is approximately \$26 billion
- e^{29} is approximately \$3.9 trillion



- $e^{24} \approx \$26$ billion
- $e^{25} \approx \$72$ billion

One unit on the LN scale:

- value x e

General rule of thumb:

- value x 3

Making Point Estimates

	(log GDP)
Bachelors (%)	0.05*
	(0.03)
Constant	24.42*
	(0.91)
Observations	50
Adjusted R^2	0.06
Residual Std. Error	1.04 (df = 48)
F Statistic	3.86* (df = 1; 48)
Note:	*p<0.06

ln (Outcome) = Intercept + Coefficient * (Predictor)

Making Point Estimates

	(log GDP)
Bachelors (%)	0.05^{*}
	(0.03)
Constant	24.42^{*}
	(0.91)
Observations	50
Adjusted R^2	0.06
Residual Std. Error	1.04 (df = 48)
F Statistic	3.86^{*} (df = 1; 48)
Note:	*p<0.06

$$\ln (GDP) = 24.42 + 0.05 * (Bachelors)$$

Making Point Estimates

	(log GDP)
Bachelors (%)	0.05*
	(0.03)
Constant	24.42*
	(0.91)
Observations	50
Adjusted R^2	0.06
Residual Std. Error	1.04 (df = 48)
F Statistic	3.86* (df = 1; 48)
Note:	*p<0.06

$$\ln (\mathrm{GDP}) = 24.42 + 0.05 * 32.16 = 26.03$$

Making Point Estimates

	(log GDP)
Bachelors (%)	0.05^{*}
	(0.03)
Constant	24.42^{*}
	(0.91)
Observations	50
Adjusted R^2	0.06
Residual Std. Error	1.04 (df = 48)
F Statistic	3.86^{*} (df = 1; 48)
Note:	*p<0.06

GDP = 26.03 (ln) = $e^{26.03}$ = 201.7 Billion USD

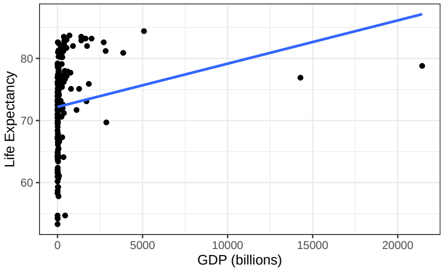
Practice with Dataset 2

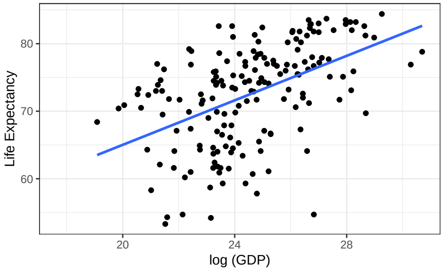
Do wealthier countries live longer?

Model 1: Regress life expectancy on GDP (billions)

Model 2: Regress life expectancy on log(GDP)

	Life Expectancy	
	(1)	(2)
GDP	0.001* (0.0003)	
log(GDP)		1.65* (0.22)
Constant	72.19* (0.57)	32.07* (5.50)
Observations	173	173
Adjusted R^2	0.03	0.24
Residual Std. Error (df = 171)	7.29	6.47
F Statistic (df = 1; 171)	6.53*	54.61*
Note:		*p<0.05





Old

Making Point Estimates

$$\mathrm{ln(Outcome)} = \mathrm{Intercept} + \mathrm{Coefficient} \text{ * } \mathrm{(Predictor)}$$

$$\mathrm{Outcome} = e^{Intercept + Coefficient * (Predictor)}$$