Public Policy 529
Linear Regression
Part 4

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 29, 2023

Recap: Key OLS Assumptions

1. Conditional mean of u is zero: $\mathrm{E}[u_i|x_i] = 0$.
2. Data are independent and identically distributed.
3. Large outliers are unlikely.
4. The error term (u) is homoskedastic.

It is important to understand these assumptions, the consequences when they are violated, and what can be done to address the violations.

Outline

Outline

Example: Population Density and Assaults

- The dependent variable is the number of assaults per 100,000 people in each of the 50 states plus DC in 2018.
  - It ranges from 60 to 648.9 assaults per 100,000 people.
- The independent variable is the population density of each state.
  - It ranges from 1.29 to 10,588.7 persons per square mile.
- Is greater population density associated with more/fewer assaults? We start with the premise that there is no effect.
- We test whether the variation across states in the assault rate is, in part, explained by variation in population density.

Regression of assault_rate on popdensity

```
. reg assault_rate popdensity
```

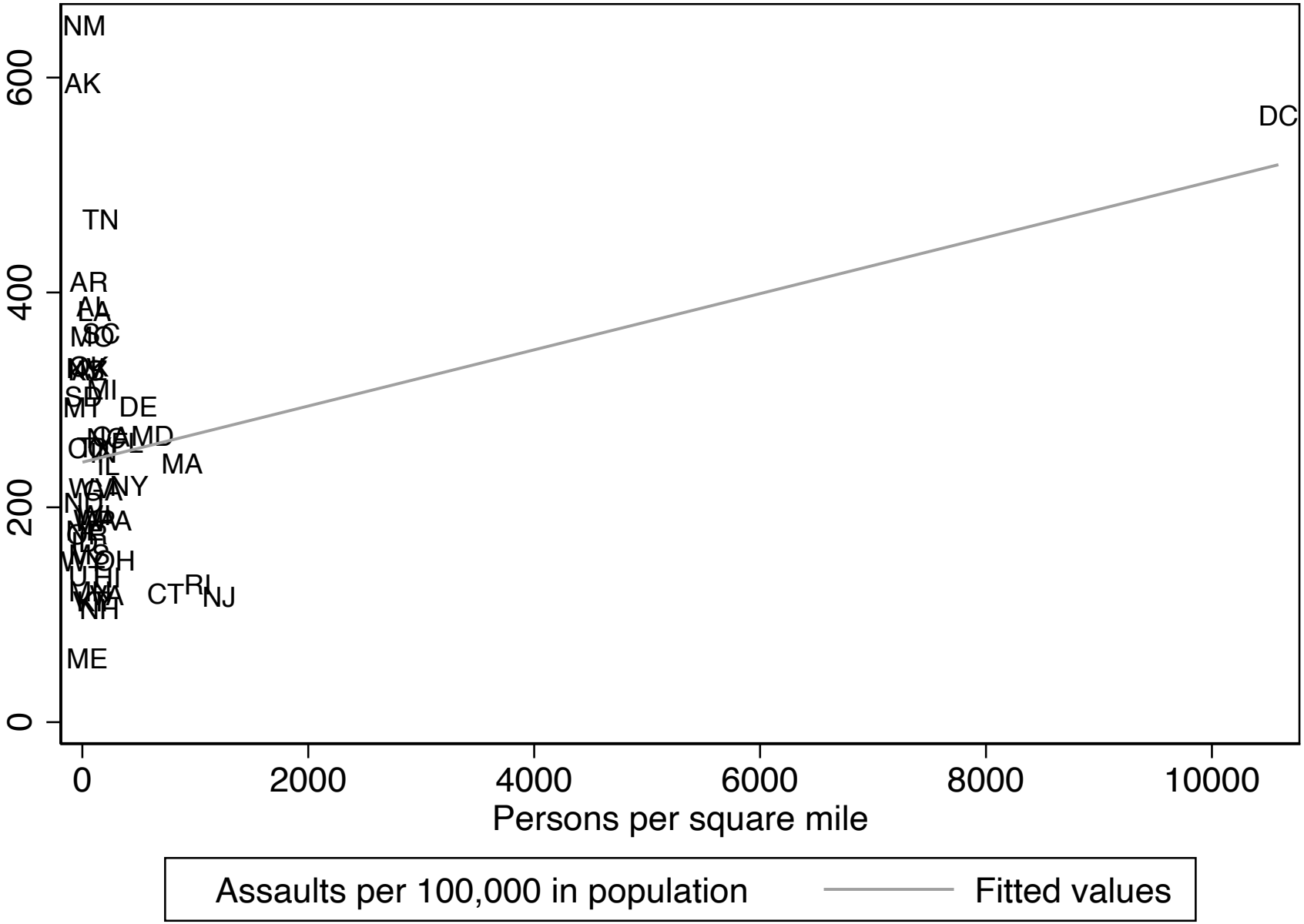
Source	SS	df	MS		
Model	74733.3796	1	74733.3796	Prob > F	0.0310
Residual	742577.548	49	15154.6438	R-squared	0.0914
				Adj R-squared	0.0729
Total	817310.927	50	16346.2185	Root MSE	123.1

Number of obs	=	51
F(1, 49)	=	4.93

assault_rate	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
popdensity	.0261557	.0117783	2.22	0.031	.0024864	.049825
_cons	241.9643	17.88912	13.53	0.000	206.0148	277.9139

The estimated relationship is positive. Each additional person per square mile is associated with .026 more assaults per 100,000.

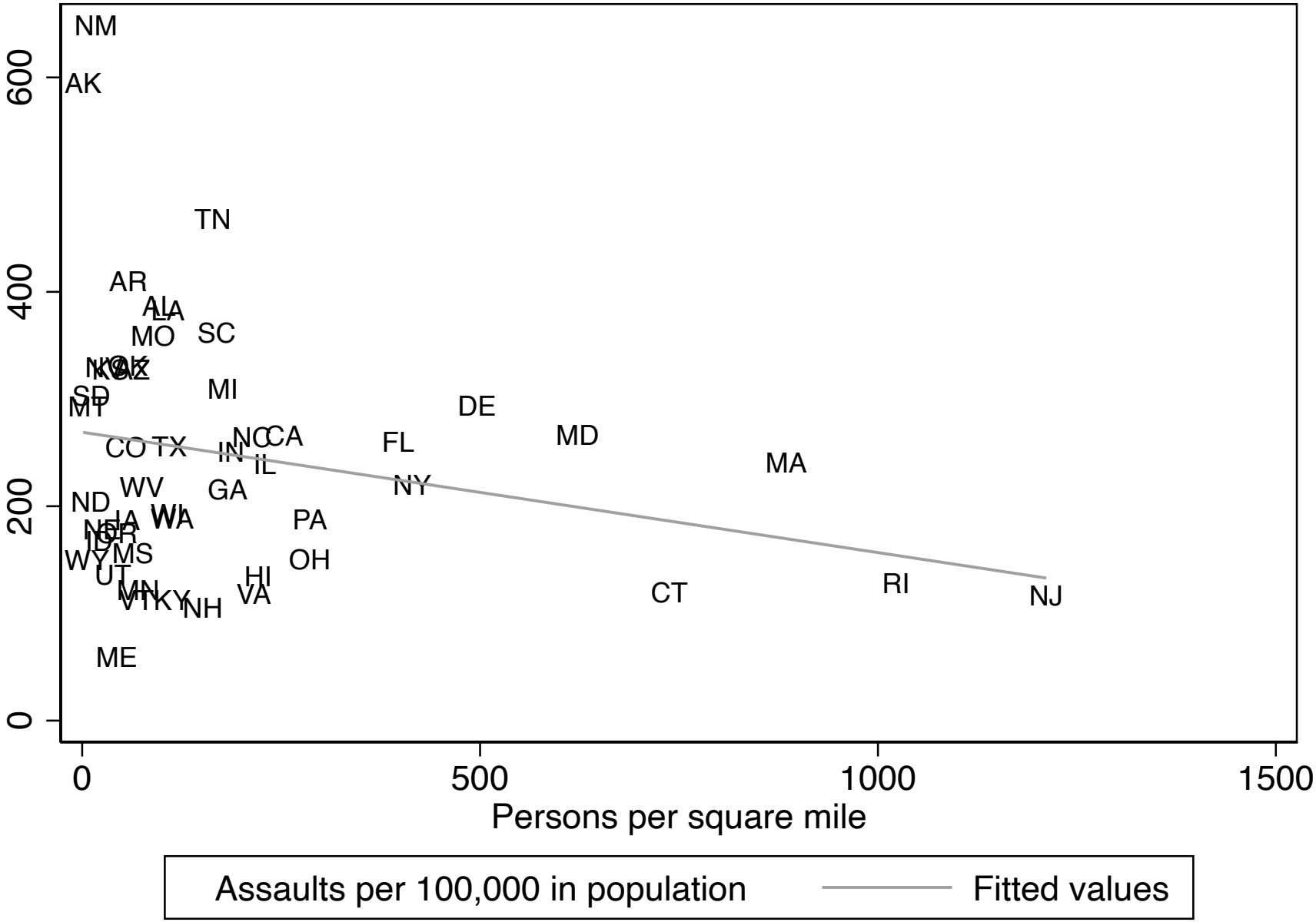
But wait... we have an outlier



What should we do?

- An outlier is a violation of OLS assumptions, and it could cause our coefficients to be very misleading.
- But, we cannot just drop a case because it is inconvenient!
- Think about the situation theoretically. Are we missing variables that might help us explain the unusual case(s)?
- Is there measurement error?
- We MAY drop the case if we decide that it does not belong to the same population as our cases of interest.

Same Plot without DC



Heteroskedasticity

- The scatterplot shows evidence of heteroskedasticity.
- Specifically, the prediction misses appear much wider when `popdensity` is low.
- We can do a test for heteroskedasticity with our software. Specifically, we will implement the Breusch-Pagan test.
- The null hypothesis is homoskedasticity.
- If we reject the null hypothesis, we should take measures to address the bias in our standard errors.

Testing for Heteroskedasticity

We will use a version of the test that performs well without requiring an assumption that the disturbances are normally distributed.

- In Stata, after you run the model, use this command:

  ```
  estat hettest, iid
  ```

- In R, if you have the `lmtest` library loaded, use this command:

  ```
  bptest(model)
  ```

The tests give you a test-statistic and p-value. If $p < \alpha$, we reject the null hypothesis of homoskedasticity.

Performing the Test

```
. estat hettest, iid

Breusch–Pagan/Cook–Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variable: Fitted values of assault_rate

H0: Constant variance

    chi2(1) =   0.32
Prob > chi2 = 0.5701
```

```
> bptest(model2)

        studentized Breusch-Pagan test

data:  model2
BP = 0.32253, df = 1, p-value = 0.5701
```

We fail to reject the null hypothesis of homoskedasticity. We do not need robust standard errors, but there's no harm in using them.

Regression with Robust Standard Errors

```
. reg assault_rate popdensity if stateid != "DC", vce(hc3)

Linear regression                               Number of obs     =         50
                                                F(1, 48)          =       6.42
                                                Prob > F          =     0.0146
                                                R-squared         =     0.0612
                                                Root MSE          =     118.47

------------------------------------------------------------------------------
             |             Robust HC3
assault_rate | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  popdensity |  -.1124653   .0443824    -2.53   0.015    -.2017021   -.0232286
       _cons |   269.0628   23.54118    11.43   0.000       221.73    316.3955
------------------------------------------------------------------------------
```

The standard error for `popdensity` is now larger than it would be with the regular formula. This is fairly typical.

Interpretation

$$\widehat{\text{assault\_rate}} = 269.1 - 0.1(\text{popdensity})$$

- Each additional person per square mile is associated with a decrease the assault rate by 0.1 assaults per 100,000 in population.
- In a state has 0 population density, we predict that there would be 269.1 assaults per 100,000 people.
- In a state where population density is 75 persons per square mile, the assault rate is predicted to be $269.1 - 0.1(75) = 268.9$ per 100,000.

Outline

Example: the Minimum Wage and Poverty

- The sample consists of the states of the United States in the year 2021.
- We will regress the poverty rate (`poverty_rate` on the minimum wage in the state (`minwage`).
- The poverty rate is measured as the percentage of people in the state that fall below the poverty line.
- The minimum wage is the hourly wage in dollars.

Regression Output

```
. reg poverty_rate minwage
```

Source	SS	df	MS		
				Number of obs	= 51
				F(1, 49)	= 2.47
Model	24.1349275	1	24.1349275	Prob > F	= 0.1227
Residual	479.381935	49	9.7833048	R-squared	= 0.0479
				Adj R-squared	= 0.0285
Total	503.516863	50	10.0703373	Root MSE	= 3.1278

poverty_rate	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
minwage	-.3028011	.1927868	-1.57	0.123	-.6902206	.0846184
_cons	14.07475	1.897752	7.42	0.000	10.26107	17.88843

Questions

- Assess the statistical significance of the coefficients.
- What is the predicted effect of raising the minimum wage by a dollar?
- What is the predicted difference in poverty for two states that are 3 dollars apart on the minimum wage?

Outline

Basics of Multiple Regression

- The principles of bivariate regression generalize to multiple independent variables.
- The function is still linear, and we now estimate a coefficient (i.e. slope) for each independent variable in the analysis.
- These coefficients give us the effect of each independent variable, holding the other independent variables constant.
- Regression still involves minimizing the sum of the squared errors: the squared distances between y_i and $\hat{y}_i$.

A Multiple Regression Function

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- We have three independent variables: x_1, x_2, and x_3.
- The coefficient on each variable gives the effect of a one-unit increase in that variable on y, holding the other variables constant.

  e.g. if we increase x_2 by 1, then y would change by β_2.
- β_0 is still the intercept: the value of y when all of the independent variables are 0.

Multiple Regression = Multi-Dimensional Space

- In bivariate regression, we plot the points on two-dimensional space. Each point is an (x,y) coordinate.
- For each new independent variable, we add another dimension.

  e.g. with two independent variables, x_1 and x_2, our points would be plotted in three-dimensional space: (x_1, x_2, y).
- Instead of a line, we fit the points with a plane, and this plane has slopes specific to the x_1 and x_2 dimensions.
- For a visual display: `http://tinyurl.com/hkuebet`

Example: Infant Mortality

- Previously, we performed a bivariate regression with a country's rate of infant mortality as the dependent variable and the percentage of the population with access to an improved water source as the independent variable.
- We will re-examine those results and then add additional independent variables.
- As we add new variables, we will see the estimated coefficients change. So will R^2 and other statistics.

Regression Output

```
. reg InfMort Water
```

Source	SS	df	MS
Model	152188.787	1	152188.787
Residual	60057.948	170	353.282047
Total	212246.735	171	1241.20897

Number of obs	172
F(1, 170)	430.79
Prob > F	0.0000
R-squared	0.7170
Adj R-squared	0.7154
Root MSE	18.796

InfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Water	-1.537692	.0740864	-20.76	0.000	-1.68394	-1.391444
_cons	167.306	6.222146	26.89	0.000	155.0234	179.5886

Note that the coefficient on Water is -1.5 and the R^2 is .72.

Example: Infant Mortality

- We will add two variables sequentially:

  lnGDPcap: log GDP per capita in 2000. This variable ranges from 4.8 to 11.2.

  PopDensity: measured in thousands of people per square km. The range of this variable is .002 to 15.6.

- Both variables plausibly are good predictors of the level of infant mortality.

Regression with Water and lnGDPcap

```
. reg InfMort Water lnGDPcap
```

Source	SS	df	MS
Model	170588.279	2	85294.1393
Residual	40145.3585	165	243.305203
Total	210733.637	167	1261.87807

Number of obs	168
F(2, 165)	350.56
Prob > F	0.0000
R-squared	0.8095
Adj R-squared	0.8072
Root MSE	15.598

InfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Water	-.8675481	.0988716	-8.77	0.000	-1.062765	-.6723316
lnGDPcap	-12.43565	1.410922	-8.81	0.000	-15.22144	-9.649864
_cons	216.7662	7.594608	28.54	0.000	201.7711	231.7613

Note: magnitude of coefficient on Water is now lower: -.8 vs. -1.5. Controlling for wealth reduced the estimated effect of access to water. Also, the R^2 has gone up, and n decreased (which can also affect coefficients). The "Root MSE" is smaller.

Interpretation

$$\widehat{\text{InfMort}} = 216.8 - .9(\text{Water}) - 12.4(\text{lnGDPcap})$$

- Each percentage point increase in access to improved water is associated with .9 fewer infant deaths out of every 1,000 births.
- For each one-unit increase in log GDP per capita, the infant mortality rate is predicted to decline by 12.4 deaths.
  - Note: a one-unit increase in log value is 271.8% growth.
- In a country where 50% of the population has access to improved water and GDP per capita is \$2,000 (i.e. lnGDPcap=7.6), the predicted rate of infant morality is:

$$216.8 - .9(50) - 12.4(7.6) = 76.8$$

Interpretation

$$\widehat{\text{InfMort}} = 216.8 - .9(\text{Water}) - 12.4(\text{lnGDPcap})$$

- Suppose two countries have the same GDP per capita but access to water is 10 percentage points higher in one country. What is the predicted difference in their infant mortality rates?

$$-.9 * 10 = -9$$

- Suppose that a country experienced growth in GDP per capita from \$2,500 to \$3,000. What is predicted to be the change in its infant mortality rate?

$$-12.4[ln(3000) - ln(2500)] = -12.4[8.0 - 7.8] = -2.5.$$

Regression with Water, lnGDPcap, and PopDensity

```
. reg InfMort Water lnGDPcap PopDensity
```

Source	SS	df	MS
Model	**169933.067**	**3**	**56644.3558**
Residual	**39937.2197**	**163**	**245.013618**
Total	**209870.287**	**166**	**1264.27884**

Number of obs =	**167**
F(3, 163) =	**231.19**
Prob > F =	**0.0000**
R-squared =	**0.8097**
Adj R-squared =	**0.8062**
Root MSE =	**15.653**

InfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Water	**-.8603657**	**.0997774**	**-8.62**	**0.000**	**-1.057389**	**-.6633428**
lnGDPcap	**-12.55082**	**1.422457**	**-8.82**	**0.000**	**-15.35964**	**-9.742001**
PopDensity	**.728273**	**2.52563**	**0.29**	**0.773**	**-4.258899**	**5.715445**
_cons	**217.1206**	**7.659664**	**28.35**	**0.000**	**201.9957**	**232.2456**

Note: adding PopDensity brings little change to the model. Its coefficient is not statistically significant.

Interpretation

$$\widehat{\text{InfMort}} = 217.1 - .9(\text{Water}) - 12.6(\text{lnGDPcap}) + .7(\text{PopDensity})$$

- Each increase in population density by 1,000 persons per square km is predicted to raise the infant mortality rate by .7 deaths, but this prediction is not statistically significant.
- In a country where 50% of the population has access to improved water, GDP per capita is \$2,000 (i.e. lnGDPcap=7.6), and population density is 2,000 persons per square km, the predicted rate of infant morality is:

$$217.1 - .9(50) - 12.6(7.6) + .7(2) = 77.7$$

Outline

Non-Linearities

- Linear regression, naturally, finds the best linear fit between x and y.
- The true relationship may not be linear, however, so we need to be careful.
- Sometimes, we may think we have found a linear relationship where one does not exist.
- Other times, we may fail to identify a relationship because we are looking only for a linear one.
- Examination of scatterplots is very useful!

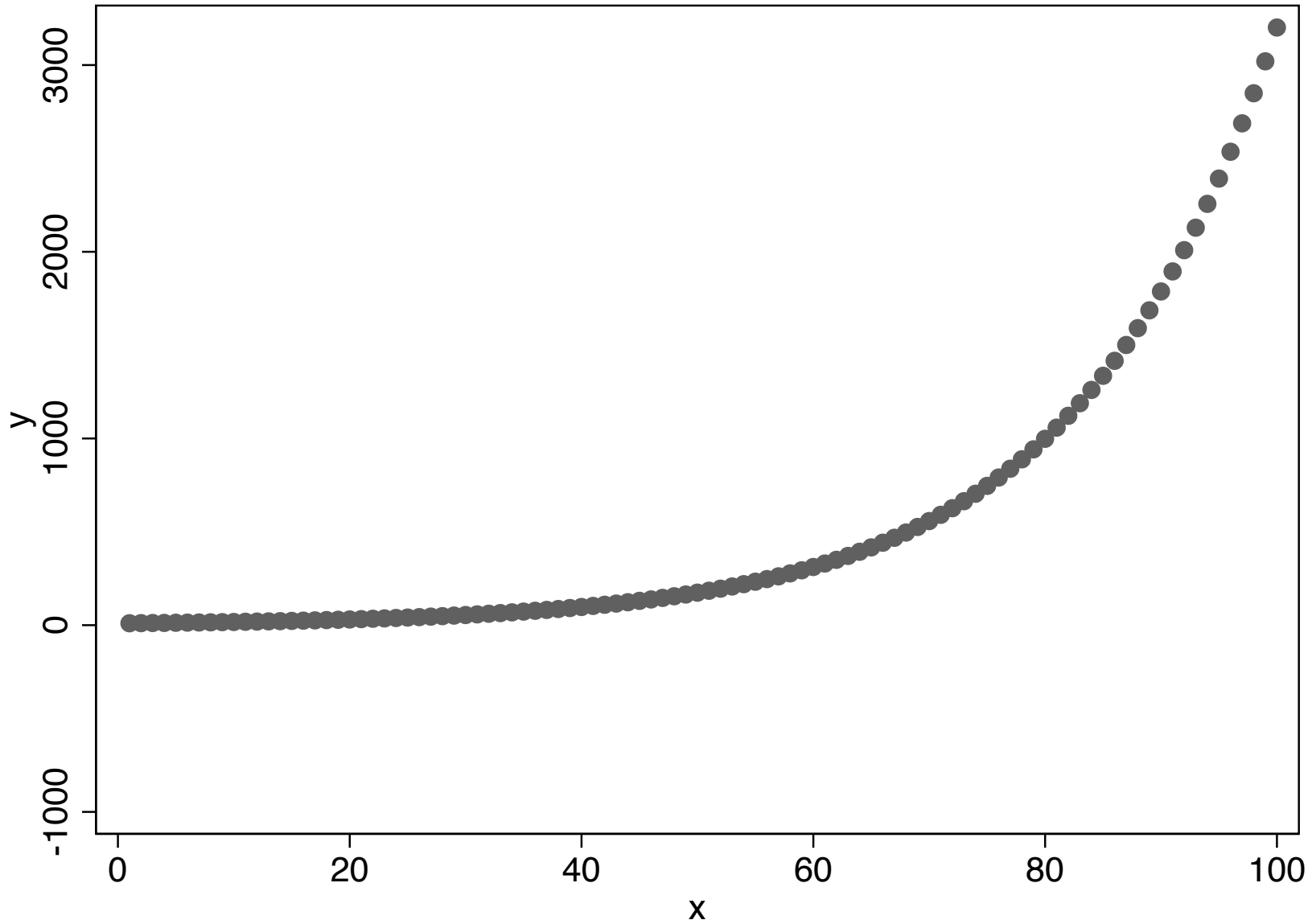
  Stata: `twoway (scatter depvar indvar)`

  R: `plot(data$yvar, data$xvar)`

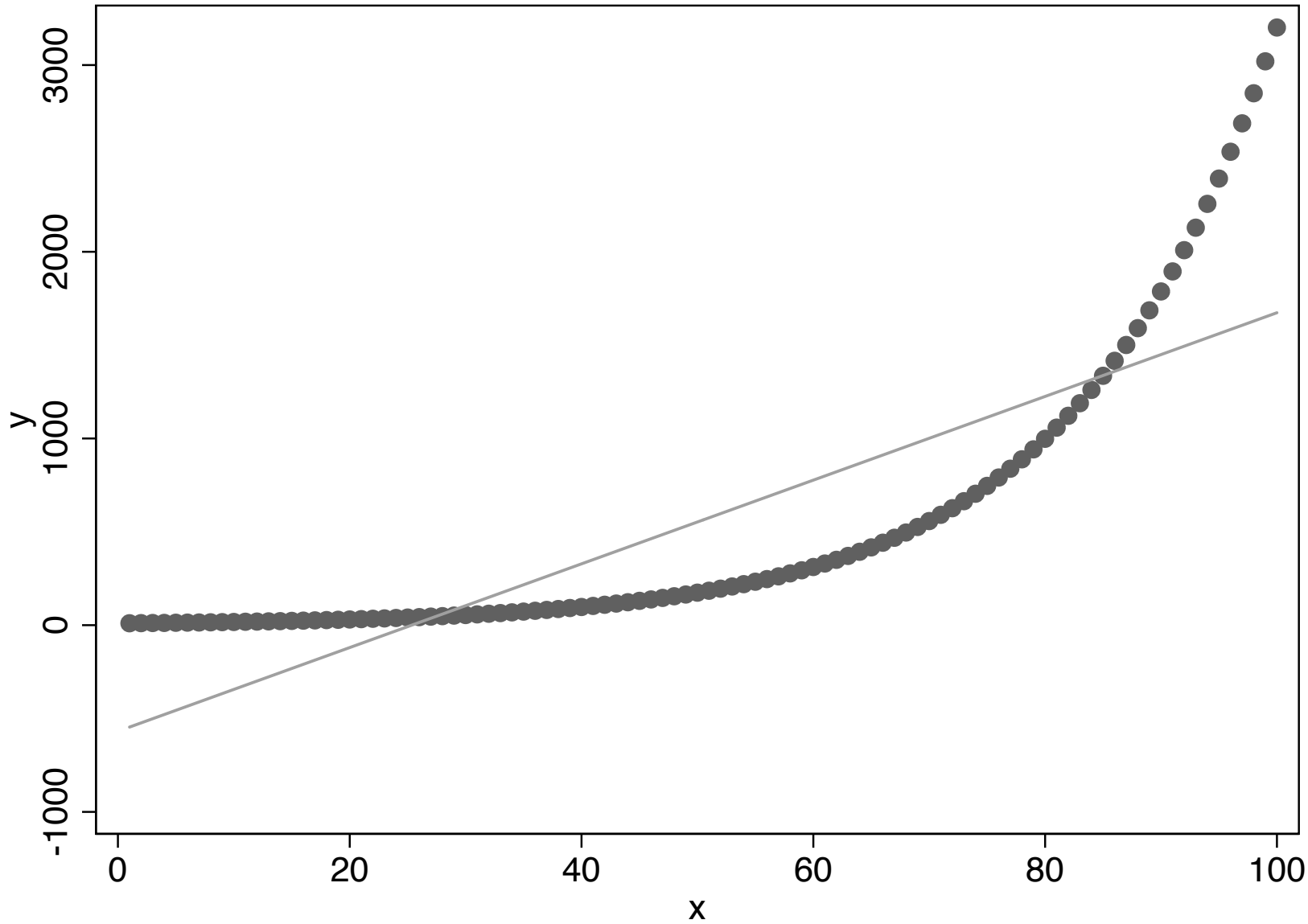
Making Non-Linear Transformations

- We can transform our variables mathematically to make them have a more linear relationship with each other.
- We use these transformed variables in the regression.
- Everything is still the same, but we have to be careful with interpretation. Most likely, we should convert values back to the original scale for interpretation.

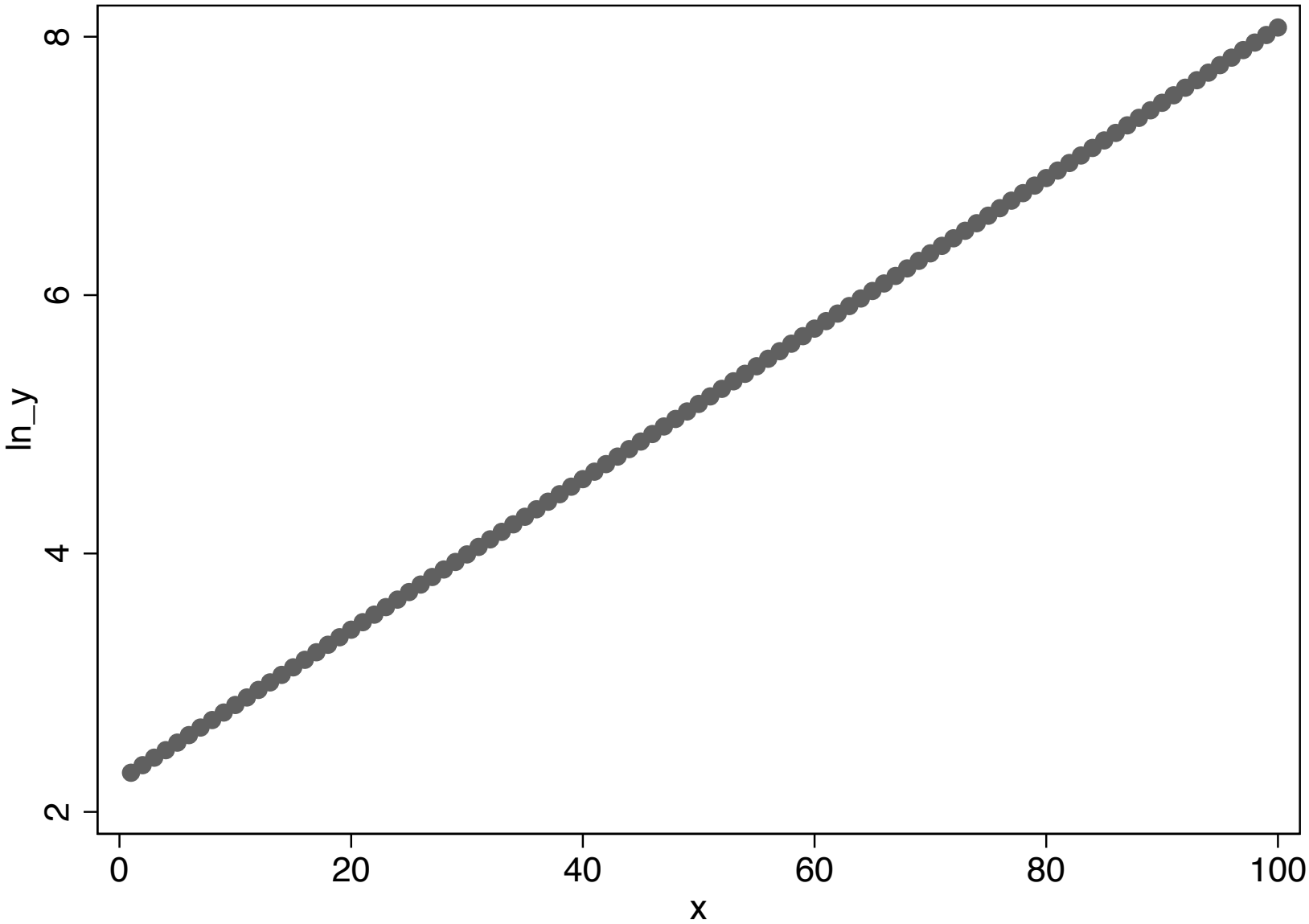
An Exponential Relationship



An Exponential Relationship



Take the Log Value of y



Example: GDP per capita and Infant Mortality

Let's test the idea that, across countries, higher GDP per capita is associated with lower infant mortality.

InfMort (dependent variable): the number of babies, out of every 1,000 born, that die before age 1. Measured in the year 2000.

GDPcap (independent variable): GDP per capita in international dollars. Measured in the year 2000.

$$\text{InfMort}_i = \beta_0 + \beta_1 \text{GDPcap}_i + \epsilon_i$$

Regression Output

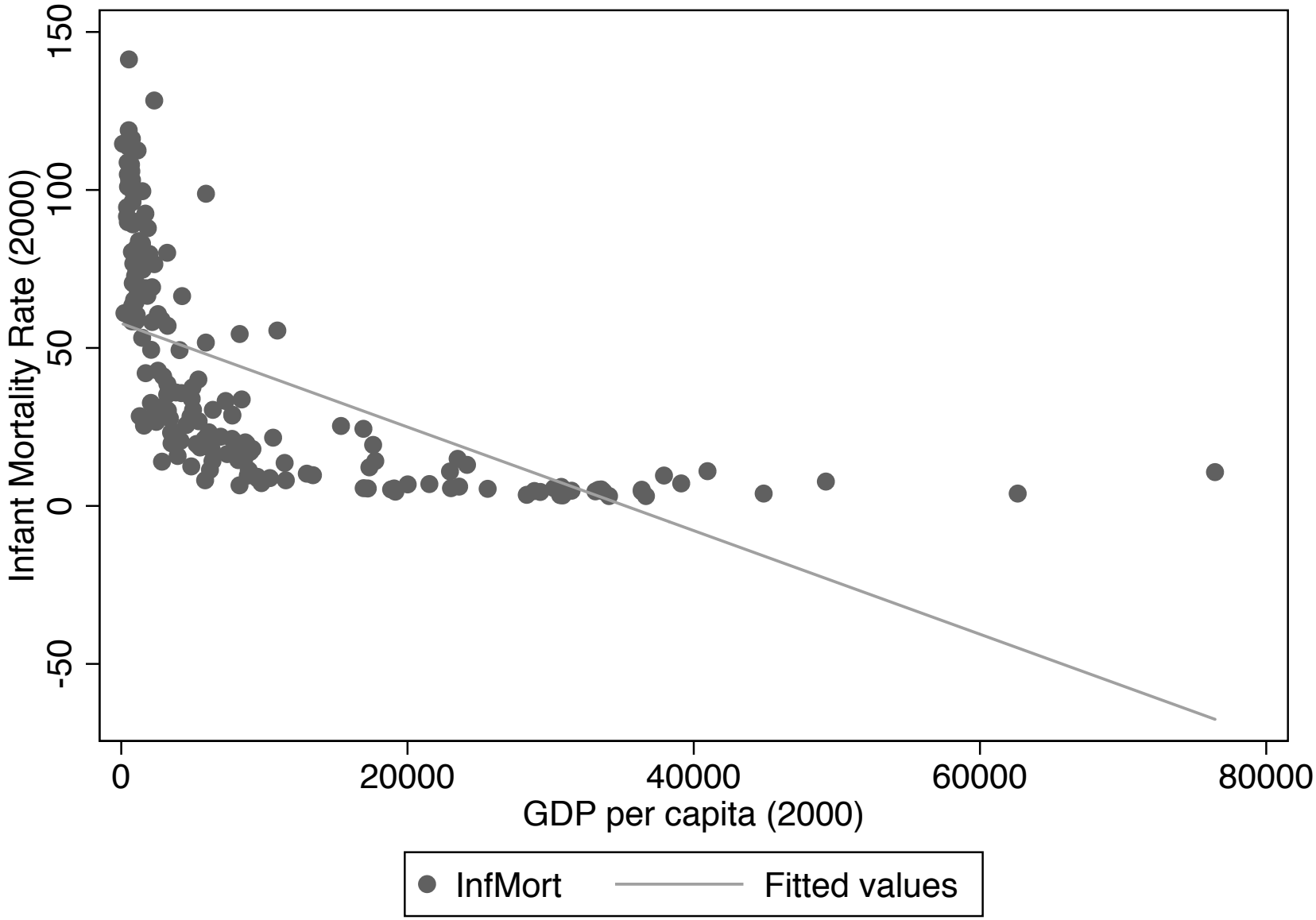
```
. reg InfMort GDPcap
```

Source	SS	df	MS
Model	78619.2541	1	78619.2541
Residual	136617.26	171	798.931343
Total	215236.514	172	1251.37508

Number of obs	= 173
F(1, 171)	= 98.41
Prob > F	= 0.0000
R-squared	= 0.3653
Adj R-squared	= 0.3616
Root MSE	= 28.265

InfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPcap	-.0016394	.0001653	-9.92	0.000	-.0019656	-.0013132
_cons	57.75891	2.738296	21.09	0.000	52.3537	63.16413

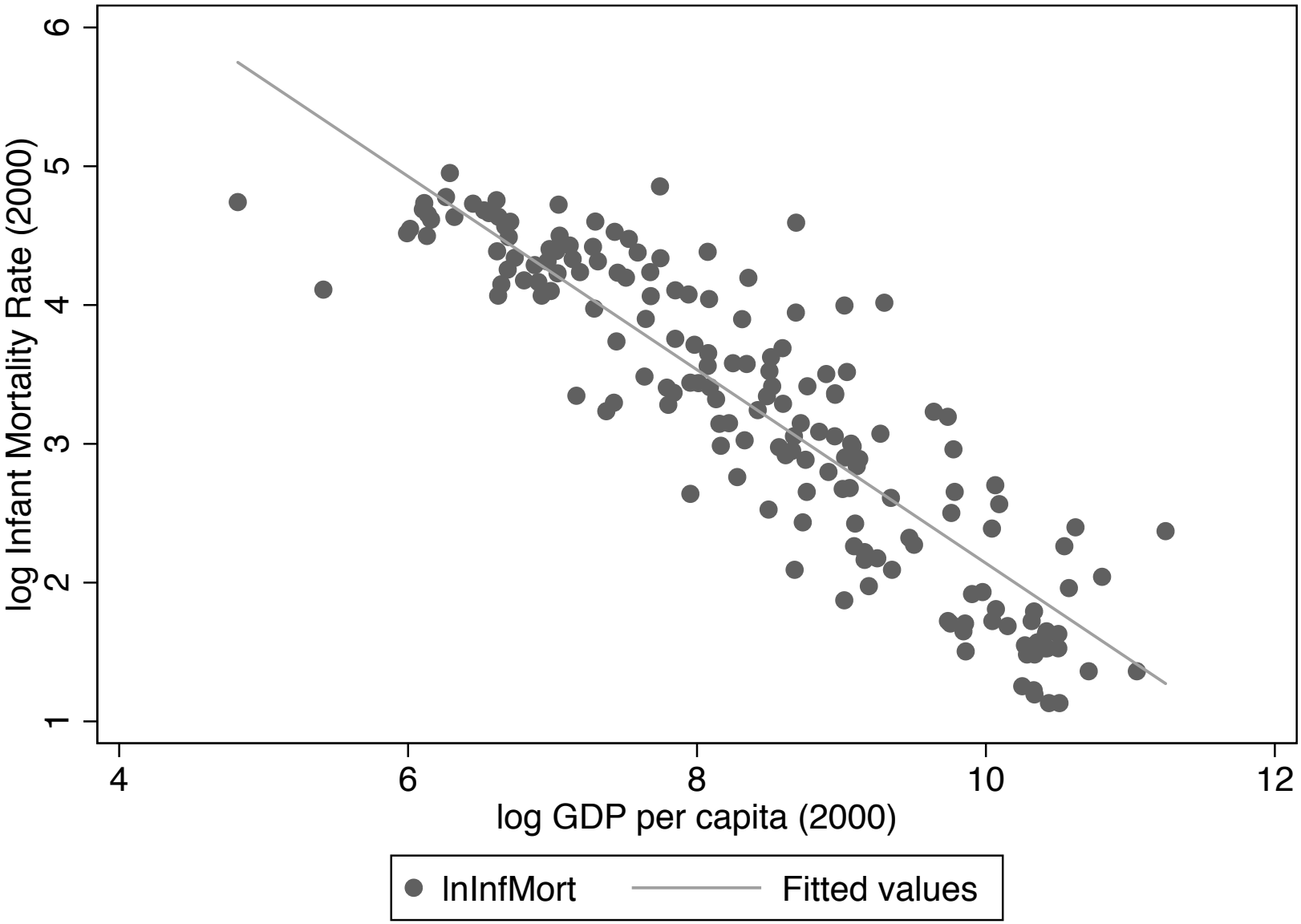
The Relationship is not Linear



Log Transformations of Both Variables

- Taking the natural log of infant mortality is appropriate because it has a natural floor, and it gets more difficult to make improvements near that floor.
- Taking the natural log of GDP per capita is appropriate because its effect is unlikely to be linear.
- Increasing GDP per capita by $1,000 would have a very large effect for low-income countries but only a marginal impact among very wealthy countries.

Linear Regression is Much More Appropriate



Regression Output

```
. reg lnInfMort lnGDPcap
```

Source	SS	df	MS		
Model	157.554735	1	157.554735	Number of obs	= 173
Residual	39.7199906	171	.232280647	F(1, 171)	= 678.29
				Prob > F	= 0.0000
Total	197.274725	172	1.14694608	R-squared	= 0.7987
				Adj R-squared	= 0.7975
				Root MSE	= .48196

lnInfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnGDPcap	-.6968972	.0267584	-26.04	0.000	-.7497164	-.644078
_cons	9.108411	.2284439	39.87	0.000	8.657478	9.559344

The fit is very good. Interpretation is now different, however. Our variables are not in their original scale.

Interpreting Regressions with Logged Variables

Regression of y on x: each one-unit change in x predicts y will change by $\hat{\beta}$ units.

Regression of $\log(y)$ on x: each one-unit change in x predicts a $(100 \cdot \hat{\beta})\%$ change in y.

Regression of y on $\log(x)$: each 1% change in x predicts y will change by $\hat{\beta}\,/100$ units.

Regression of $\log(y)$ on $\log(x)$: each 1% change in x predicts a $\hat{\beta}\,\%$ change in y.