

4.3 – Nonlinearity & Transformation

ECON 480 • Econometrics • Fall 2022

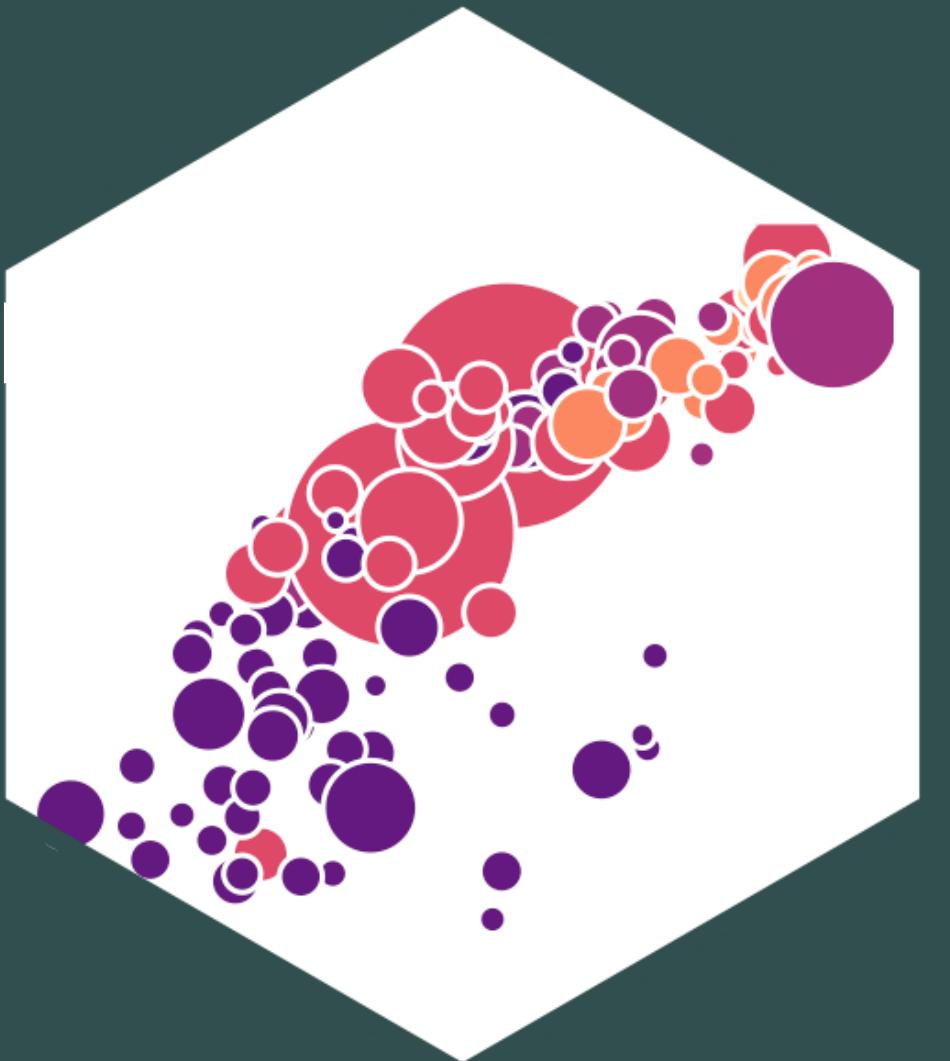
Dr. Ryan Safner

Associate Professor of Economics

 safner@hood.edu

ryansafner/metricsF22

 metricsF22.classes.ryansafner.com



Contents

Nonlinear Effects

Polynomial Models

Quadratic Model

Logarithmic Models

Linear-Log Model

Log-Linear Model

Log-Log Model

Standardizing & Comparing Across Units

Joint Hypothesis Testing

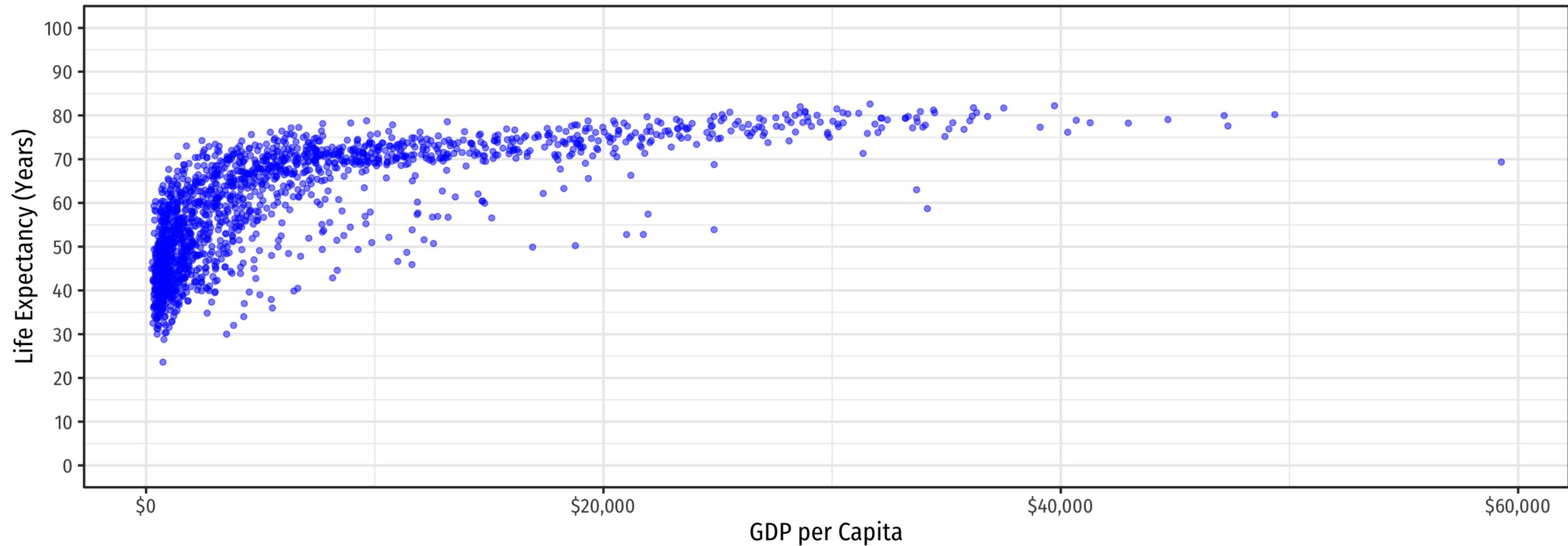
Nonlinear Effects

Linear Regression

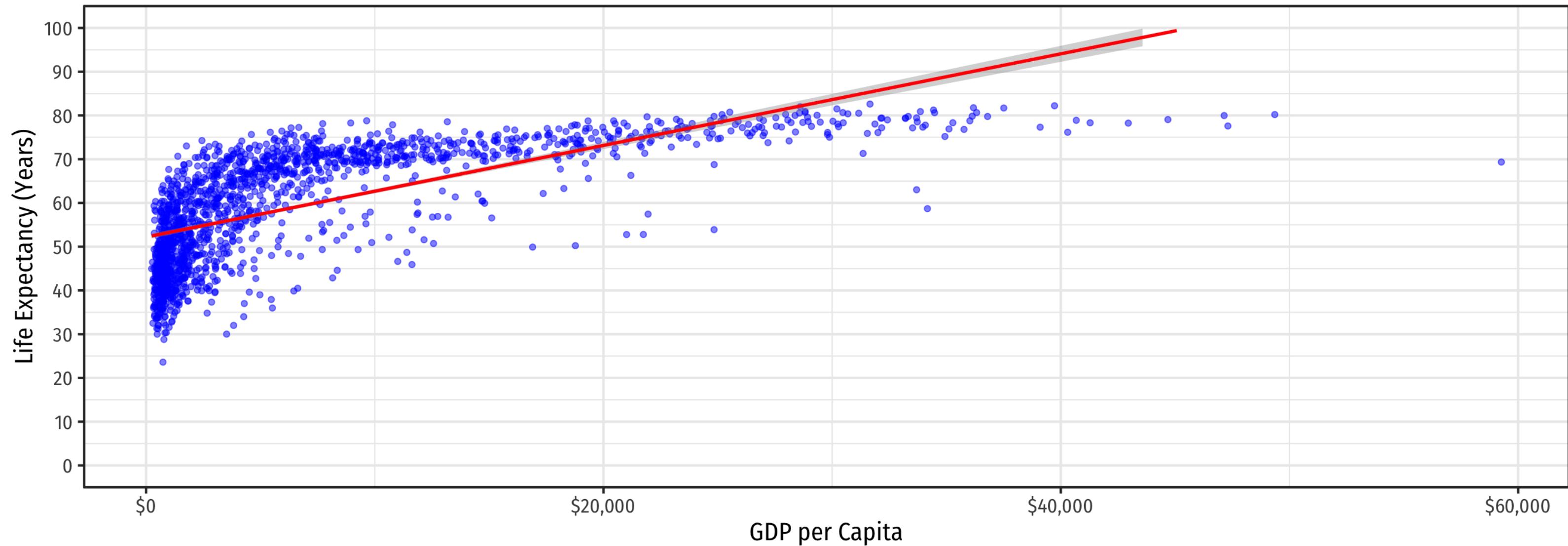
- OLS is commonly known as “**linear regression**” as it fits a **straight line** to data points
- Often, data and relationships between variables may *not* be linear



Linear Regression



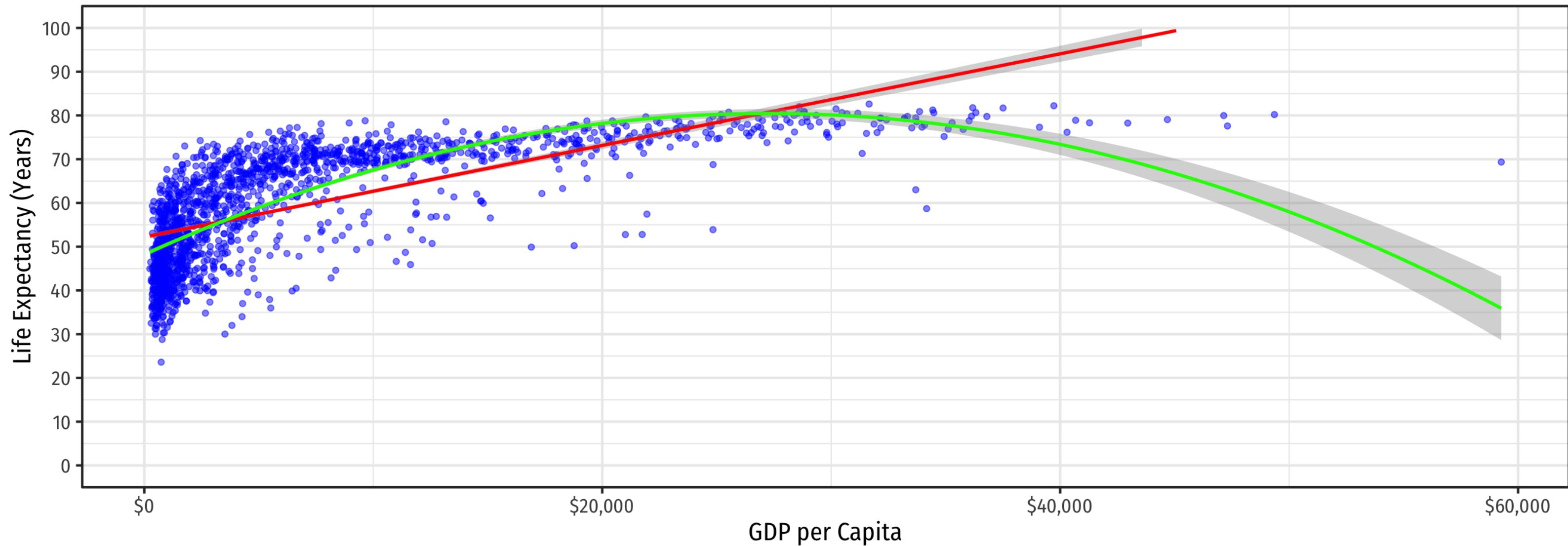
Linear Regression



$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i$$



Linear Regression

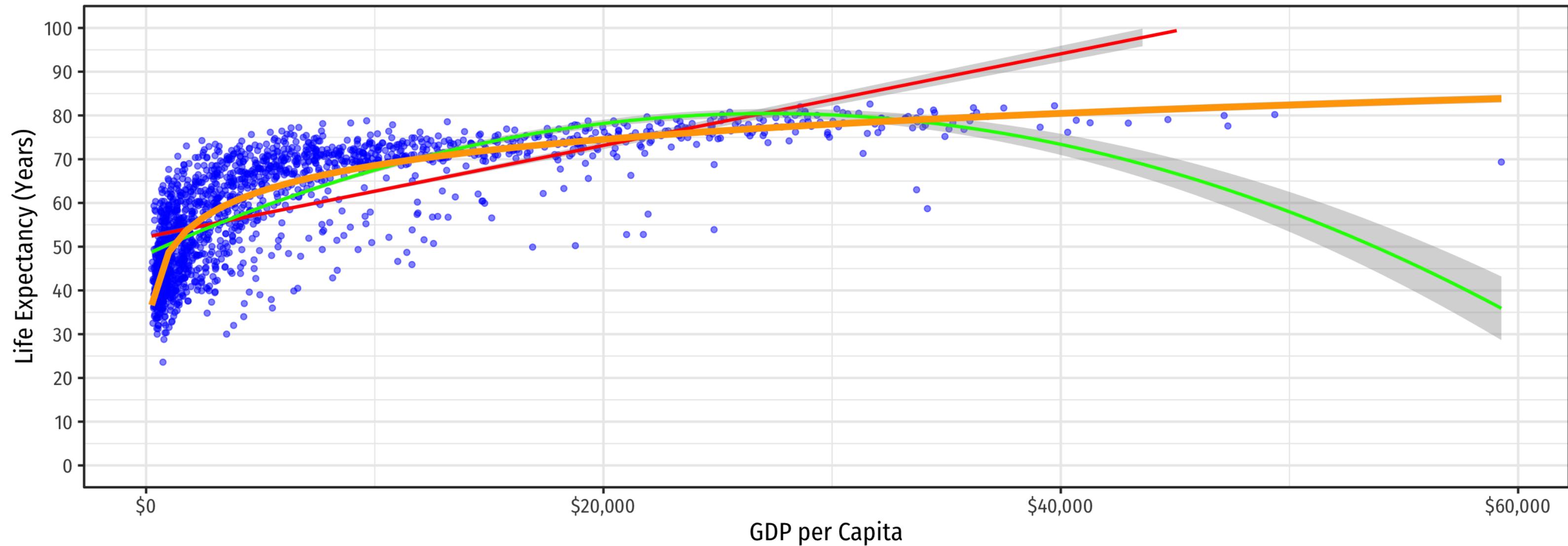


$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i$$

$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i + \hat{\beta}_2 \text{GDP}_i^2$$



Linear Regression



$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i$$

$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i + \hat{\beta}_2 \text{GDP}_i^2$$

$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln \text{GDP}_i$$



Sources of Nonlinearities

- Effect of $X_1 \rightarrow Y$ might be nonlinear if:
 1. $X_1 \rightarrow Y$ is different for different levels of X_1
 - e.g. **diminishing returns**: $\uparrow X_1$ increases Y at a *decreasing* rate
 - e.g. **increasing returns**: $\uparrow X_1$ increases Y at an *increasing* rate
 2. $X_1 \rightarrow Y$ is different for different levels of X_2 • e.g. interaction effects (last lesson)

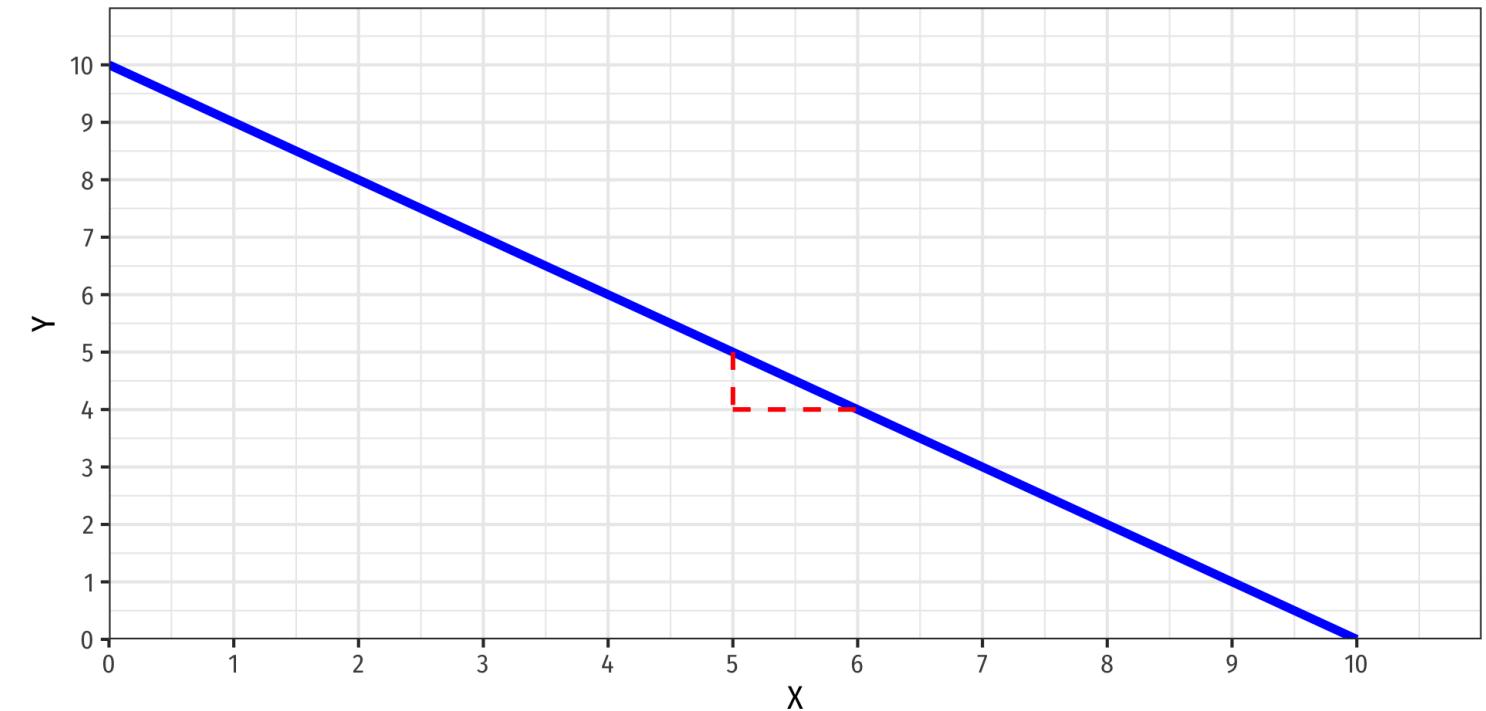


Nonlinearities Alter Marginal Effects

- **Linear:**

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

- marginal effect (slope), $(\hat{\beta}_1) = \frac{\Delta Y}{\Delta X}$ is constant for all X

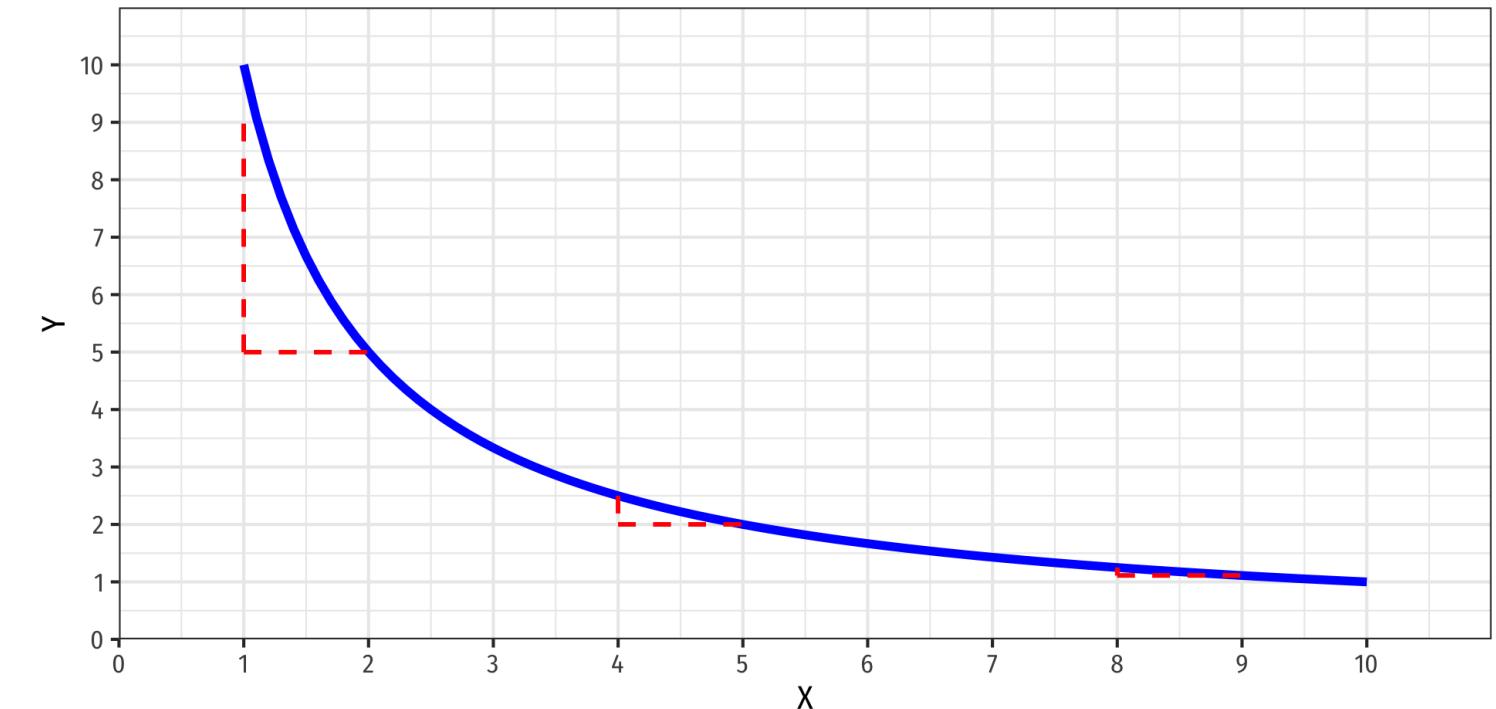


Nonlinearities Alter Marginal Effects

- **Polynomial:**

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

- Marginal effect, “slope” ($\neq \hat{\beta}_1$) depends on the value of X !

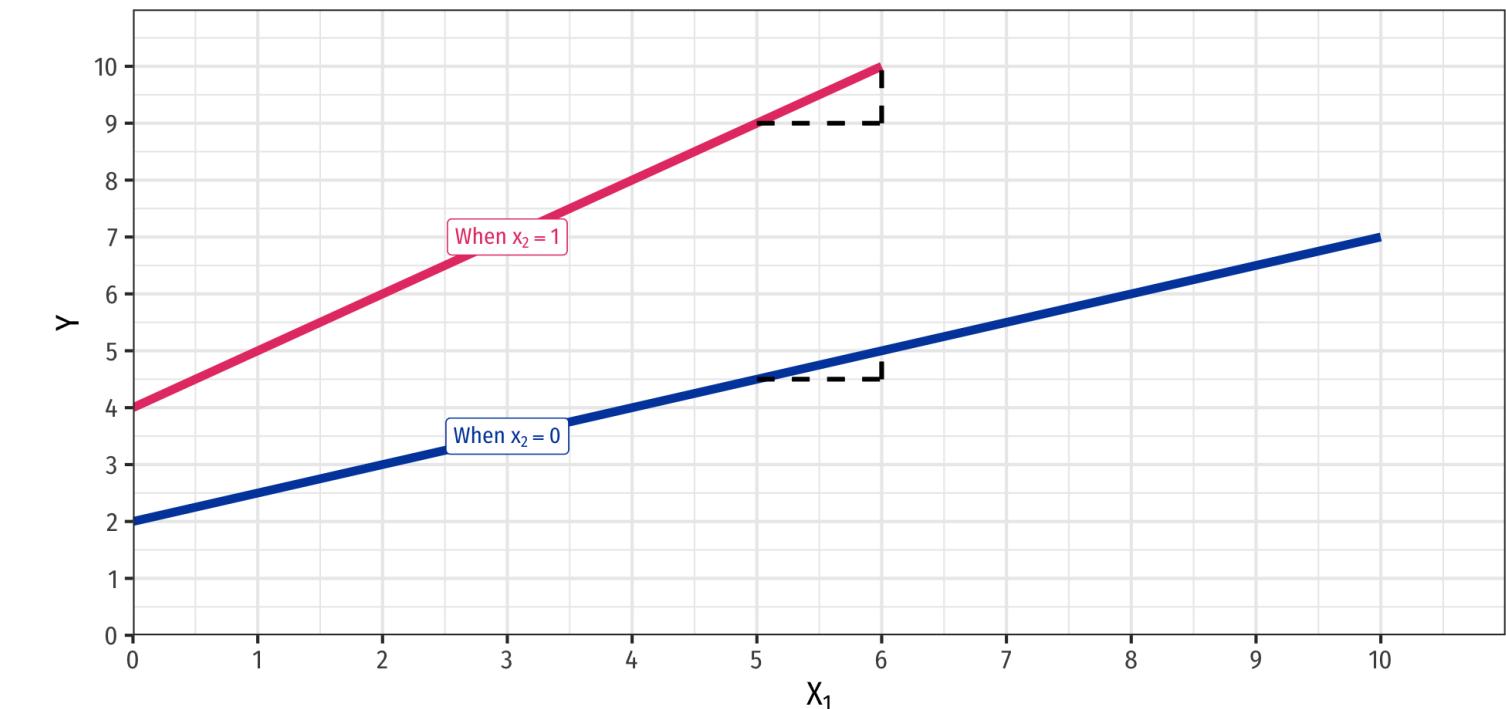


Nonlinearities Alter Marginal Effects

- **Interaction Effect:**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 \times X_2$$

- Marginal effect, “slope” depends on the value of X_2 !
- Easy example: if X_2 is a dummy variable:
 - $X_2 = 0$ (control) vs. $X_2 = 1$ (treatment)

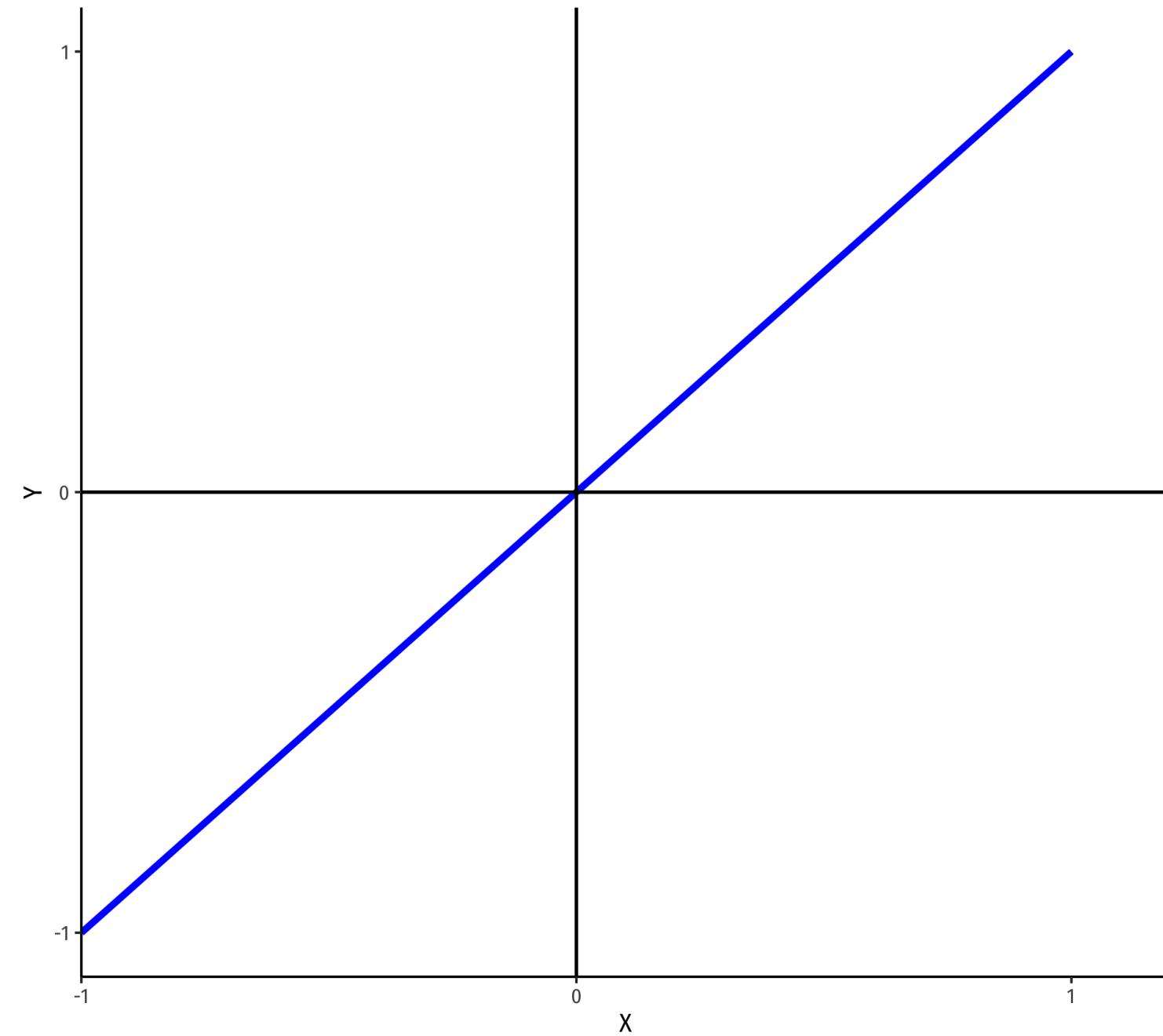


Polynomial Models

Polynomial Functions of X I

- Linear

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$



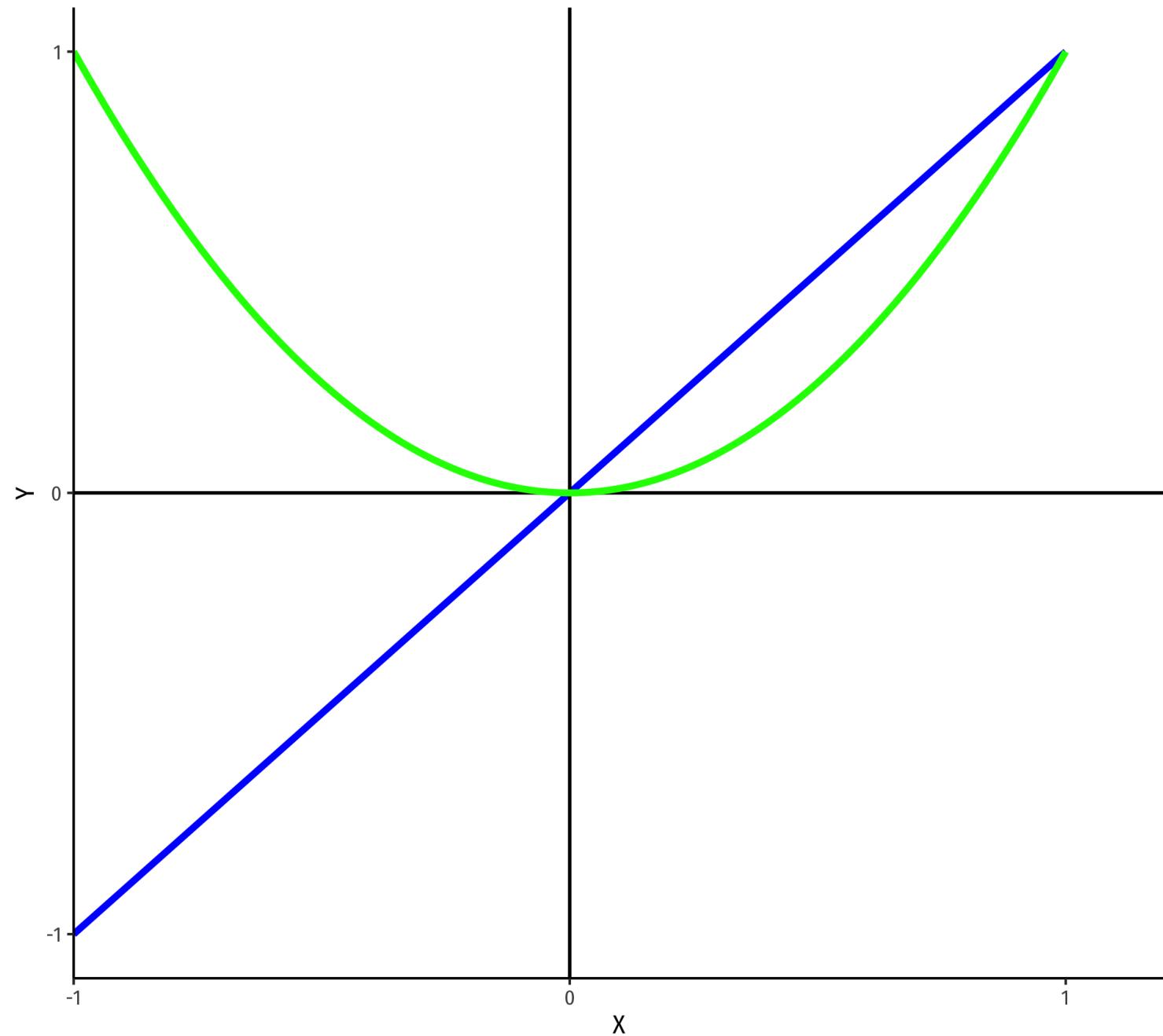
Polynomial Functions of X I

- Linear

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Quadratic

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$



Polynomial Functions of X I

- Linear

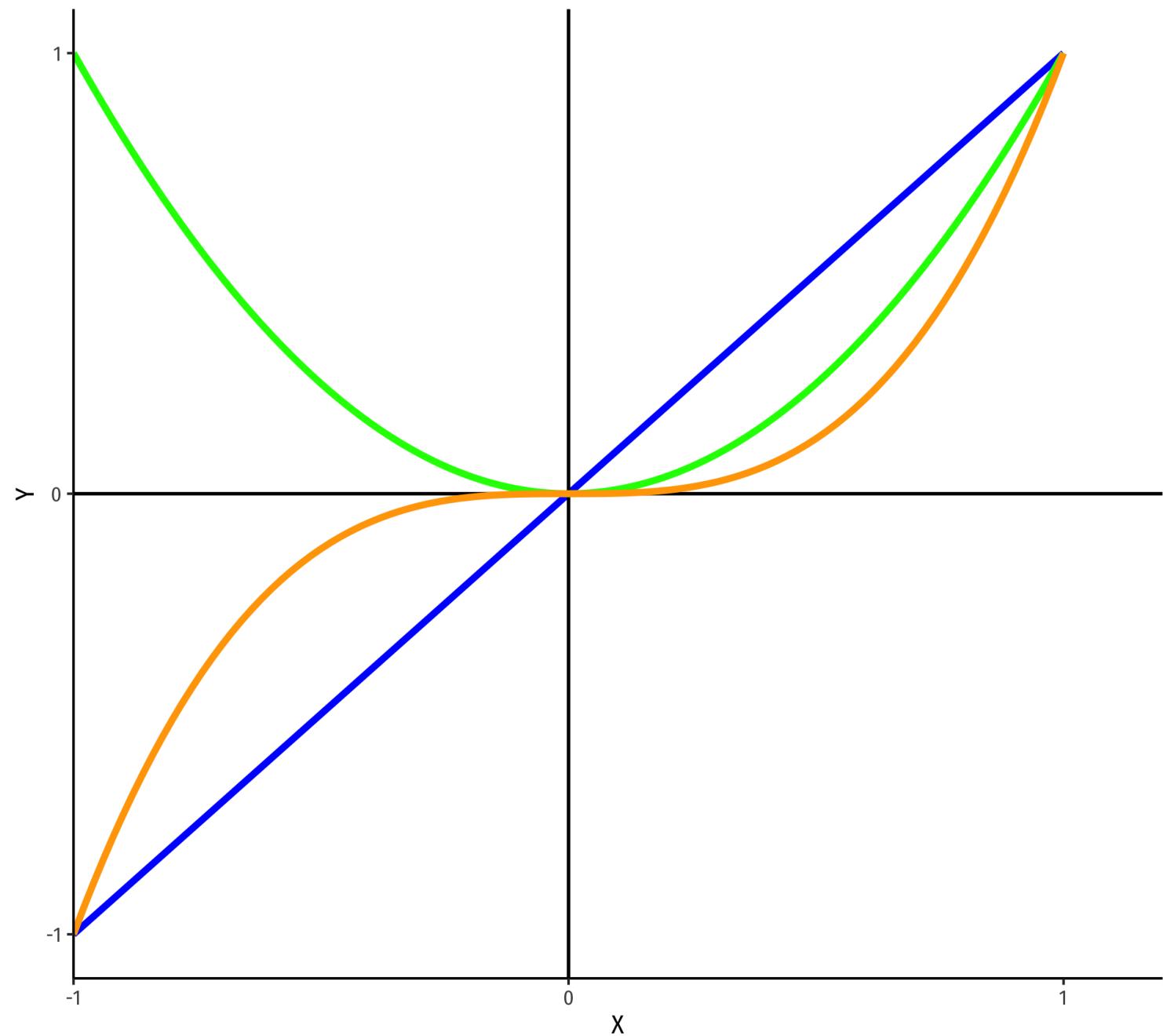
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Quadratic

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

- Cubic

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3$$



Polynomial Functions of X I

- Linear

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Quadratic

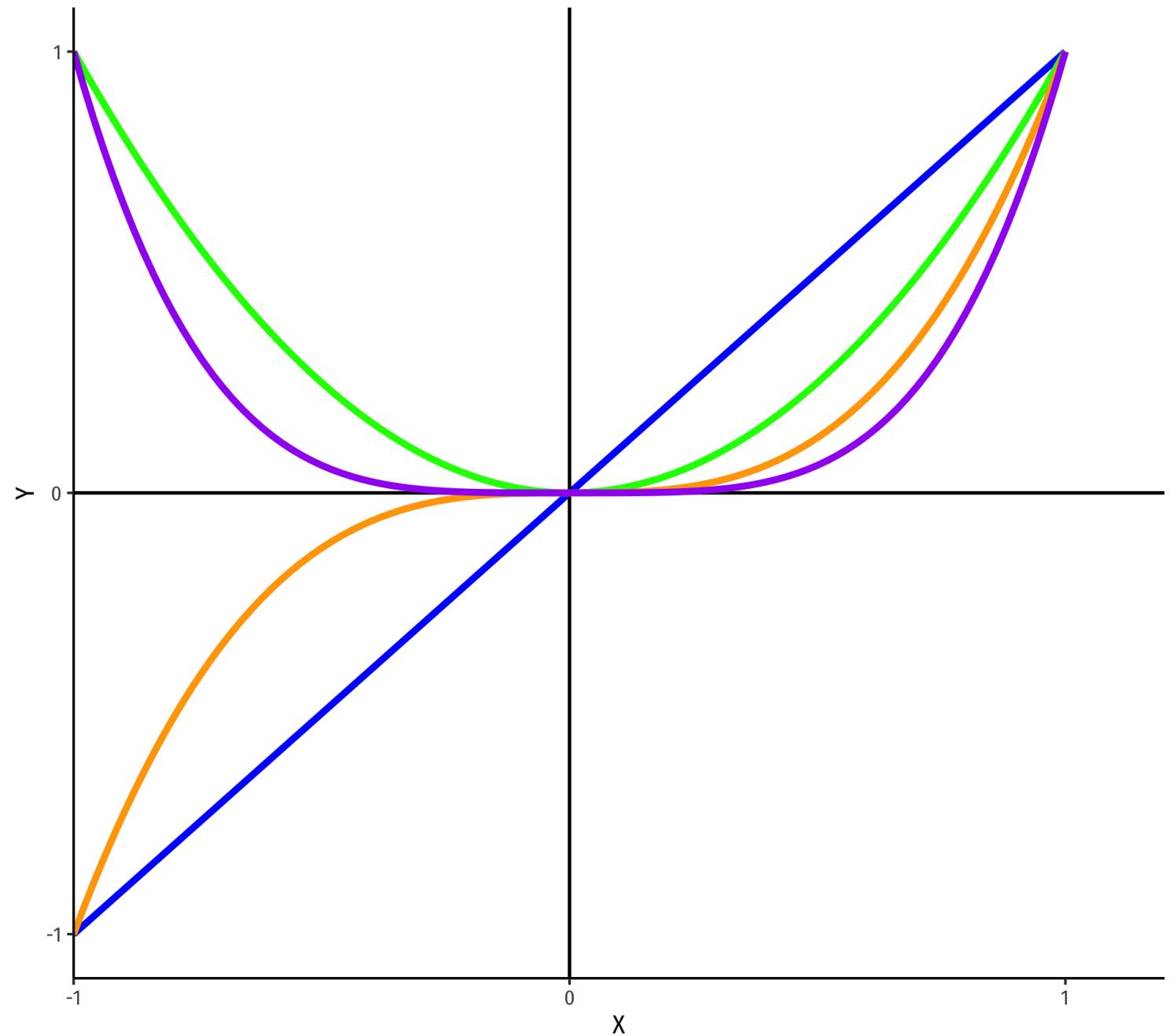
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

- Cubic

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3$$

- Quartic

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3 + \hat{\beta}_4 X^4$$



Polynomial Functions of X II

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \cdots + \hat{\beta}_r X_i^r + u_i$$

- Where r is the highest power X_i is raised to
 - quadratic $r = 2$
 - cubic $r = 3$
- The graph of an r^{th} -degree polynomial function has $(r - 1)$ bends
- Just another multivariate OLS regression model!



Quadratic Model

Quadratic Model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$$

- **Quadratic model** has X and X^2 variables in it (yes, need both!)
- How to interpret coefficients (betas)?
 - β_0 as “intercept” and β_1 as “slope” makes no sense 🤔
 - β_1 as effect $X_i \rightarrow Y_i$ holding X_i^2 constant??¹
- **Estimate marginal effects** by calculating predicted \hat{Y}_i for different levels of X_i



Quadratic Model: Calculating Marginal Effects

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$$

- What is the **marginal effect** of $\Delta X_i \rightarrow \Delta Y_i$?
- Take the **derivative** of Y_i with respect to X_i :

$$\frac{\partial Y_i}{\partial X_i} = \hat{\beta}_1 + 2\hat{\beta}_2 X_i$$

- **Marginal effect** of a 1 unit change in X_i is a $(\hat{\beta}_1 + 2\hat{\beta}_2 X_i)$ unit change in Y



Quadratic Model: Example I



Example

$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP per capita}_i + \hat{\beta}_2 \text{GDP per capita}_i^2$$

- Use `gapminder` package and data

```
1 library(gapminder)
```



Quadratic Model: Example II

- These coefficients will be very large, so let's transform gdpPercap to be in \$1,000's

```

1 gapminder <- gapminder %>%
2   mutate(GDP_t = gdpPercap/1000)
3
4 gapminder %>% head() # look at it

```

country	continent	year
<fct>	<fct>	<int>
Afghanistan	Asia	1952
Afghanistan	Asia	1957
Afghanistan	Asia	1962
Afghanistan	Asia	1967
Afghanistan	Asia	1972
Afghanistan	Asia	1977

6 rows | 1-3 of 7 columns



Quadratic Model: Example II

- Let's also create a squared term, gdp_sq

```

1 gapminder <- gapminder %>%
2   mutate(GDP_sq = GDP_t^2)
3
4 gapminder %>% head() # look at it

```

country	continent	year
<fct>	<fct>	<int>
Afghanistan	Asia	1952
Afghanistan	Asia	1957
Afghanistan	Asia	1962
Afghanistan	Asia	1967
Afghanistan	Asia	1972
Afghanistan	Asia	1977

6 rows | 1-3 of 8 columns



Quadratic Model: Example IV

- Can “manually” run a multivariate regression with `GDP_t` and `GDP_sq`

```

1 library(broom)
2 reg1 <- lm(lifeExp ~ GDP_t + GDP_sq, data = gapminder)
3
4 reg1 %>% tidy()

```

term	estimate
<chr>	<dbl>
(Intercept)	50.52400578
GDP_t	1.55099112
GDP_sq	-0.01501927
3 rows 1-2 of 5 columns	



Quadratic Model: Example IV

- OR use `gdp_t` and add the `I()` operator to transform the variable in the regression,
`I(gdp_t^2)`¹

```
1 reg1_alt <- lm(lifeExp ~ GDP_t + I(GDP_t^2), data = gapminder)
2
3 reg1_alt %>% tidy()
```

term	estimate
<chr>	<dbl>
(Intercept)	50.52400578
GDP_t	1.55099112
I(GDP_t^2)	-0.01501927
3 rows 1-2 of 5 columns	

1. Here is a decent explanation of what `I()` does. An alternative is to use `poly(GDP_t, 2)` to make the squared term, but this has some issues.



Quadratic Model: Example V

term	estimate
	<dbl>
(Intercept)	50.52400578
GDP_t	1.55099112
GDP_sq	-0.01501927

3 rows | 1-2 of 5 columns

$$\widehat{\text{Life Expectancy}}_i = 50.52 + 1.55 \text{GDP}_i - 0.02 \text{GDP}_i^2$$

- Positive effect ($\hat{\beta}_1 > 0$), with diminishing returns ($\hat{\beta}_2 < 0$)
- Marginal effect of GDP on Life Expectancy **depends on initial value of GDP!**



Quadratic Model: Example VI

term	estimate
	<dbl>
(Intercept)	50.52400578
GDP_t	1.55099112
GDP_sq	-0.01501927

3 rows | 1-2 of 5 columns

- **Marginal effect** of GDP on Life Expectancy:

$$\frac{\partial Y}{\partial X} = \hat{\beta}_1 + 2\hat{\beta}_2 X_i$$

$$\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} \approx 1.55 + 2(-0.02) \text{ GDP}$$

$$\approx 1.55 - 0.04 \text{ GDP}$$



Quadratic Model: Example VII

$$\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} = 1.55 - 0.04 \text{ GDP}$$

Marginal effect of GDP if GDP = 5 (\$ thousand):

$$\begin{aligned}\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} &= 1.55 - 0.04\text{GDP} \\ &= 1.55 - 0.04(5) \\ &= 1.55 - 0.20 \\ &= 1.35\end{aligned}$$

- i.e. for every addition \$1 (thousand) in GDP per capita, average life expectancy increases by 1.35 years



Quadratic Model: Example VIII

$$\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} = 1.55 - 0.04 \text{ GDP}$$

Marginal effect of GDP if GDP = 25 (\$ thousand):

$$\begin{aligned}\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} &= 1.55 - 0.04\text{GDP} \\ &= 1.55 - 0.04(25) \\ &= 1.55 - 1.00 \\ &= 0.55\end{aligned}$$

- i.e. for every addition \$1 (thousand) in GDP per capita, average life expectancy increases by 0.55 years



Quadratic Model: Example X

$$\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} = 1.55 - 0.04 \text{ GDP}$$

Marginal effect of GDP if GDP = 50 (\$ thousand):

$$\begin{aligned}\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} &= 1.55 - 0.04\text{GDP} \\ &= 1.55 - 0.04(50) \\ &= 1.55 - 2.00 \\ &= -0.45\end{aligned}$$

- i.e. for every addition \$1 (thousand) in GDP per capita, average life expectancy *decreases* by 0.45 years



Quadratic Model: Example XI

$$\widehat{\text{Life Expectancy}}_i = 50.52 + 1.55 \text{ GDP per capita}_i - 0.02 \text{ GDP per capita}_i^2$$

$$\frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} = 1.55 - 0.04\text{GDP}$$

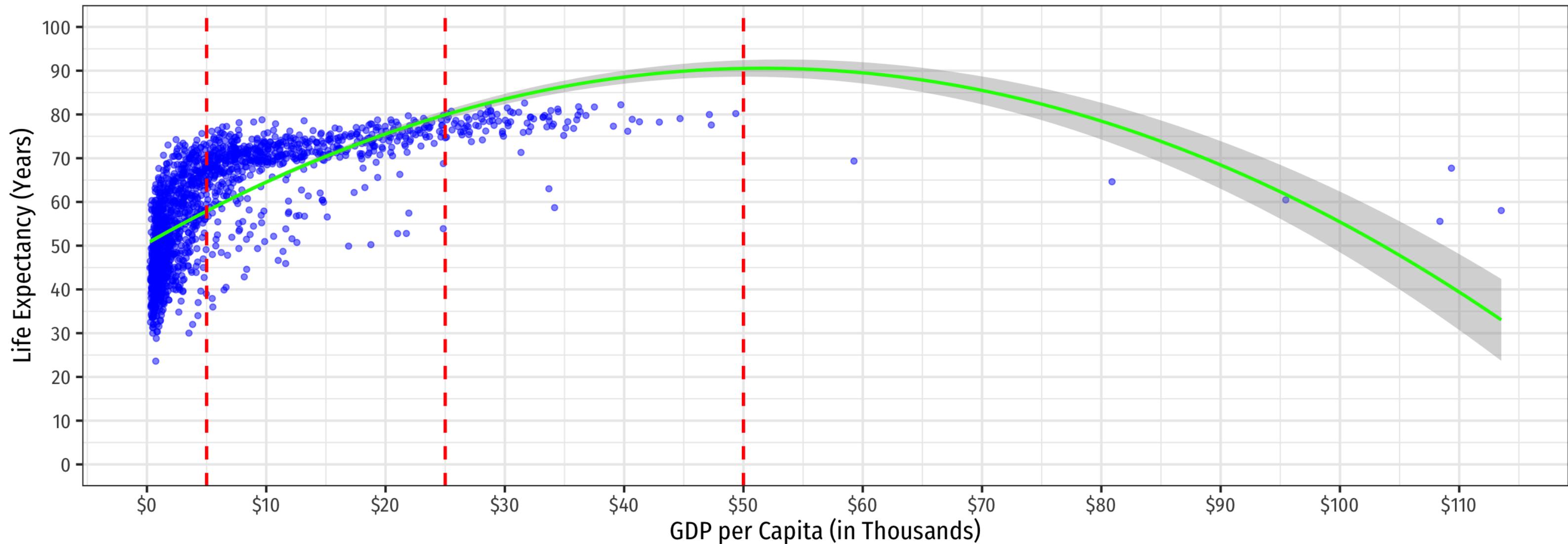
Initial GDP per capita	Marginal Effect¹
\$5,000	1.35 years
\$25,000	0.55 years
\$50,000	-0.45 years

¹ Of +\$1,000 GDP/capita on Life Expectancy



Quadratic Model: Example XII

► Code



Quadratic Model: Maxima and Minima I

- For a polynomial model, we can also find the predicted **maximum or minimum** of \hat{Y}_i
- A quadratic model has a single global maximum or minimum (1 bend)
- By calculus, a minimum or maximum occurs where:

$$\frac{\partial Y_i}{\partial X_i} = 0$$

$$\beta_1 + 2\beta_2 X_i = 0$$

$$2\beta_2 X_i = -\beta_1$$

$$X_i^* = -\frac{\beta_1}{2\beta_2}$$



Quadratic Model: Maxima and Minima II

term	estimate
	<dbl>
(Intercept)	50.52400578
GDP_t	1.55099112
GDP_sq	-0.01501927
3 rows 1-2 of 5 columns	

$$GDP_i^* = -\frac{\beta_1}{2\beta_2}$$

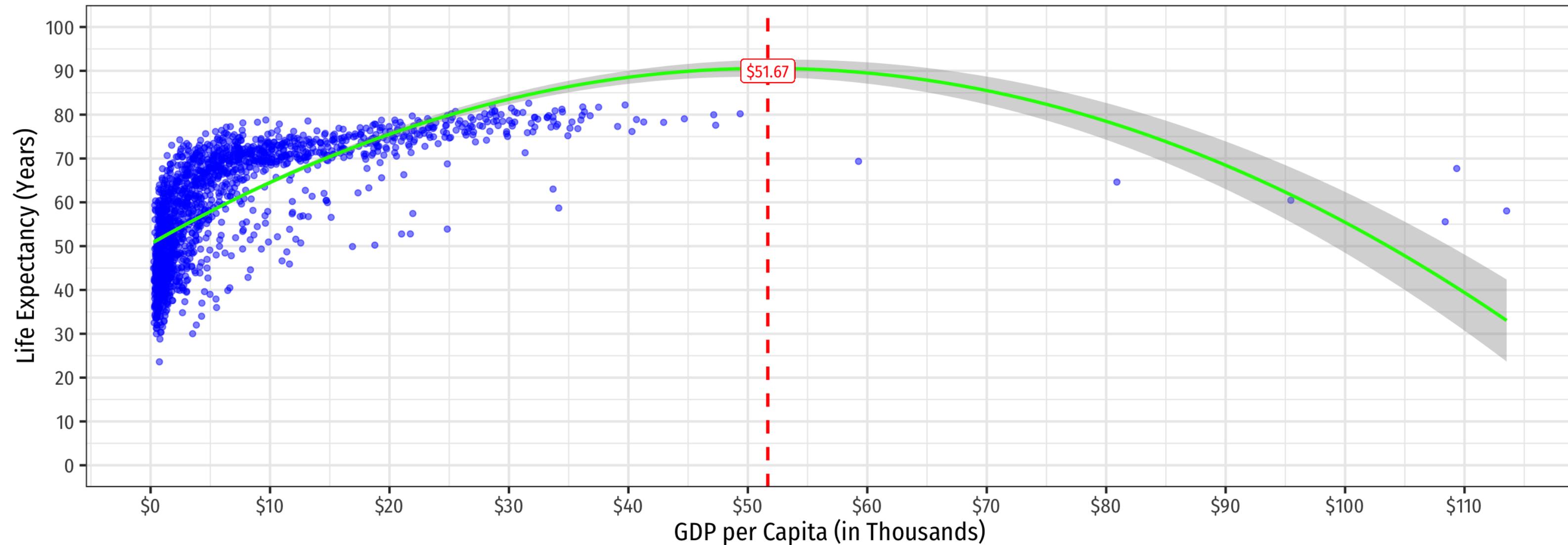
$$GDP_i^* = -\frac{(1.55)}{2(-0.015)}$$

$$GDP_i^* \approx 51.67$$



Quadratic Model: Maxima and Minima III

► Code



Determining If Polynomials Are Necessary I

term	estimate
	<dbl>
(Intercept)	50.52400578
GDP_t	1.55099112
GDP_sq	-0.01501927

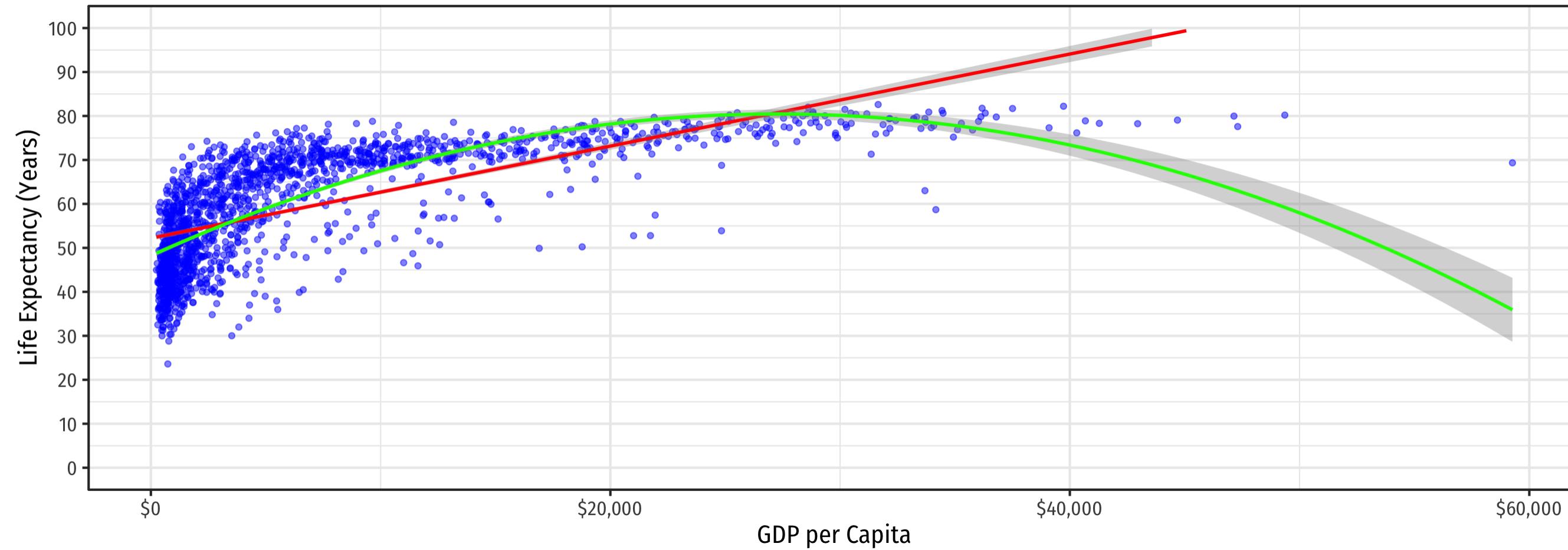
3 rows | 1-2 of 5 columns

- Is the quadratic term necessary?
- Determine if $\hat{\beta}_2$ (on X_i^2) is statistically significant:
 - $H_0 : \hat{\beta}_2 = 0$
 - $H_a : \hat{\beta}_2 \neq 0$



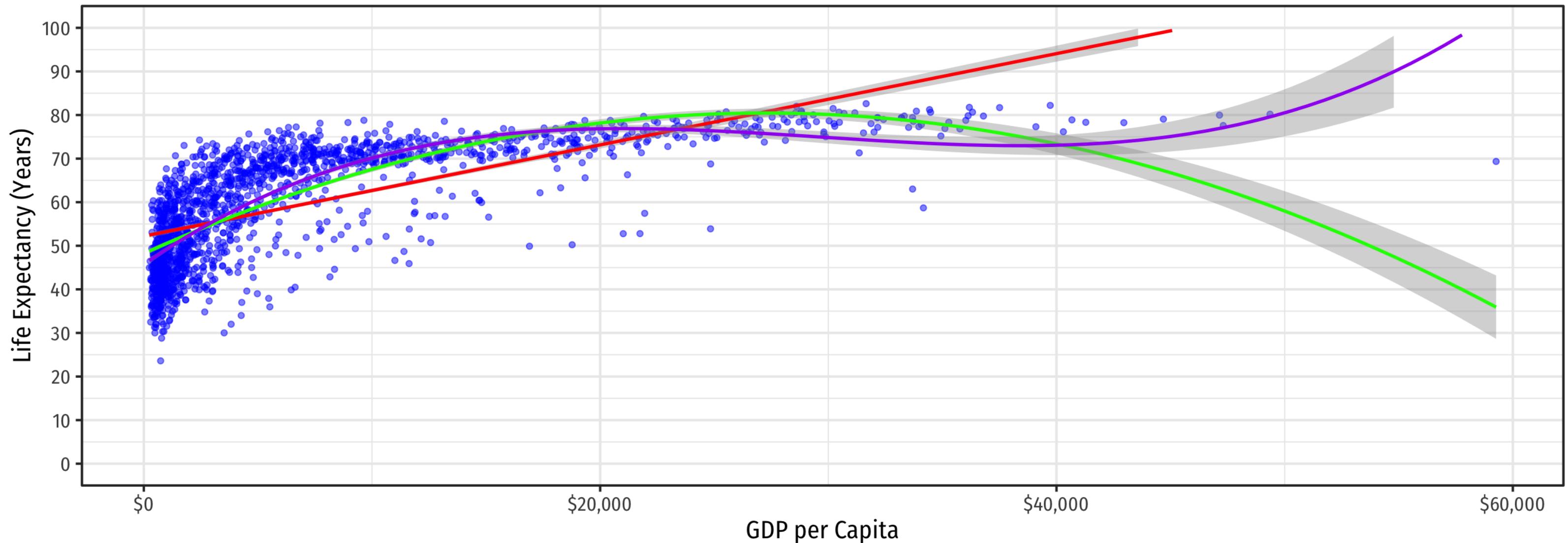
Determining Polynomials are Necessary II

- Should we keep going up in polynomials?



Determining Polynomials are Necessary II

- Should we keep going up in polynomials?

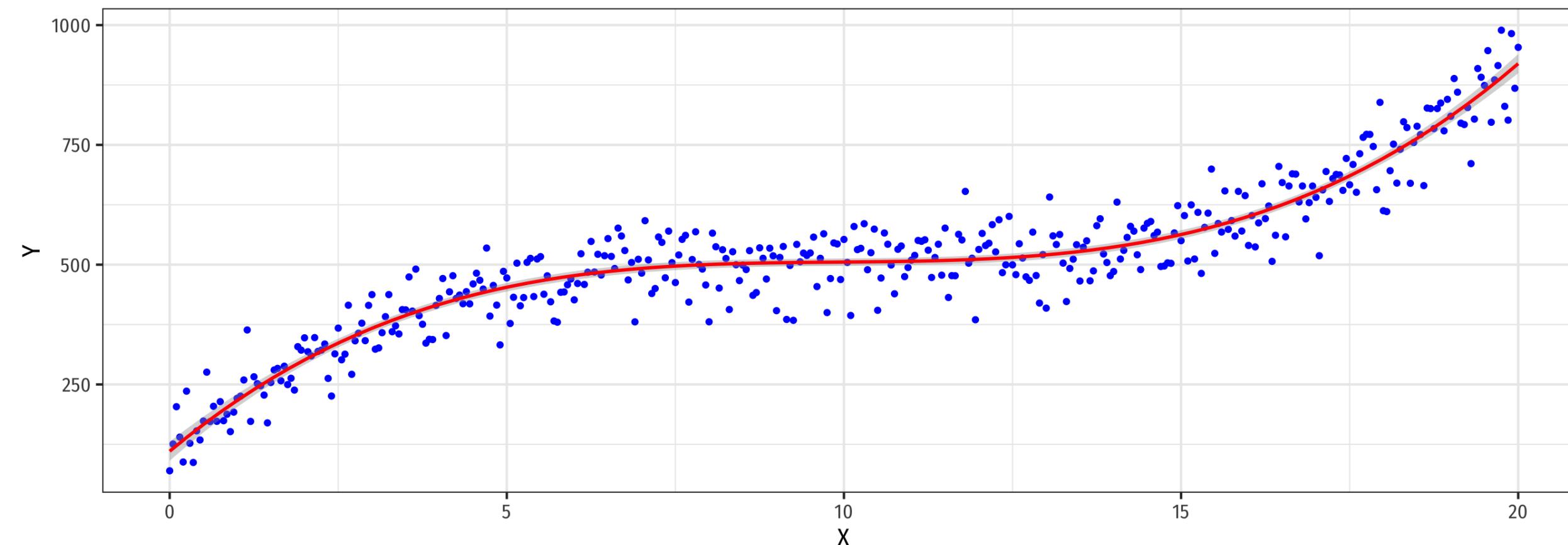


$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 GDP_i + \hat{\beta}_2 GDP_i^2 + \hat{\beta}_3 GDP_i^3$$



Determining Polynomials are Necessary III

- In general, you should have a **compelling theoretical reason** why data or relationships should **“change direction”** multiple times
- Or clear data patterns that have multiple “bends”
- Recall, **we care more** about accurately measuring the causal effect of $X \rightarrow Y$, rather than getting the most accurate prediction possible for \hat{Y}



Determining Polynomials are Necessary IV

term	estimate
<chr>	<dbl>
(Intercept)	47.4755069510
GDP_t	2.7226370698
I(GDP_t^2)	-0.0681545071
I(GDP_t^3)	0.0004093149

4 rows | 1-2 of 5 columns

- $\hat{\beta}_3$ is statistically significant...
- ...but can we really think of a good reason to complicate the model?



If You Kept Going...

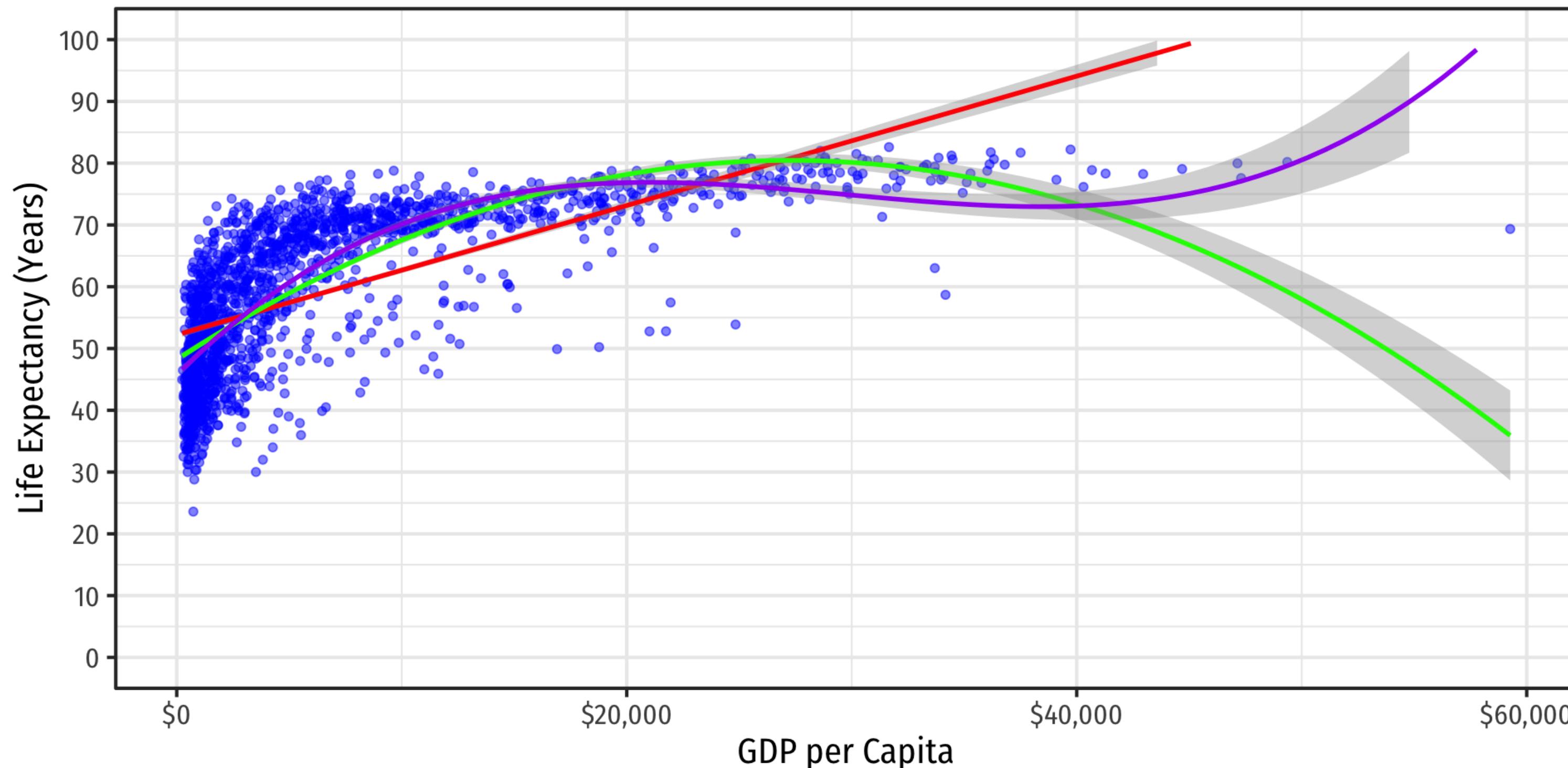
term	estimate	std.error
<chr>	<dbl>	<dbl>
(Intercept)	4.003294e+01	5.846282e-01
GDP_t	8.722968e+00	5.290582e-01
I(GDP_t^2)	-1.081312e+00	1.294759e-01
I(GDP_t^3)	7.190930e-02	1.334295e-02
I(GDP_t^4)	-2.705563e-03	7.010624e-04
I(GDP_t^5)	6.063170e-05	2.056983e-05
I(GDP_t^6)	-8.254873e-07	3.495442e-07
I(GDP_t^7)	6.685309e-09	3.408241e-09
I(GDP_t^8)	-2.956581e-11	1.766287e-11
I(GDP_t^9)	5.490732e-14	3.765889e-14

1-10 of 10 rows | 1-3 of 5 columns

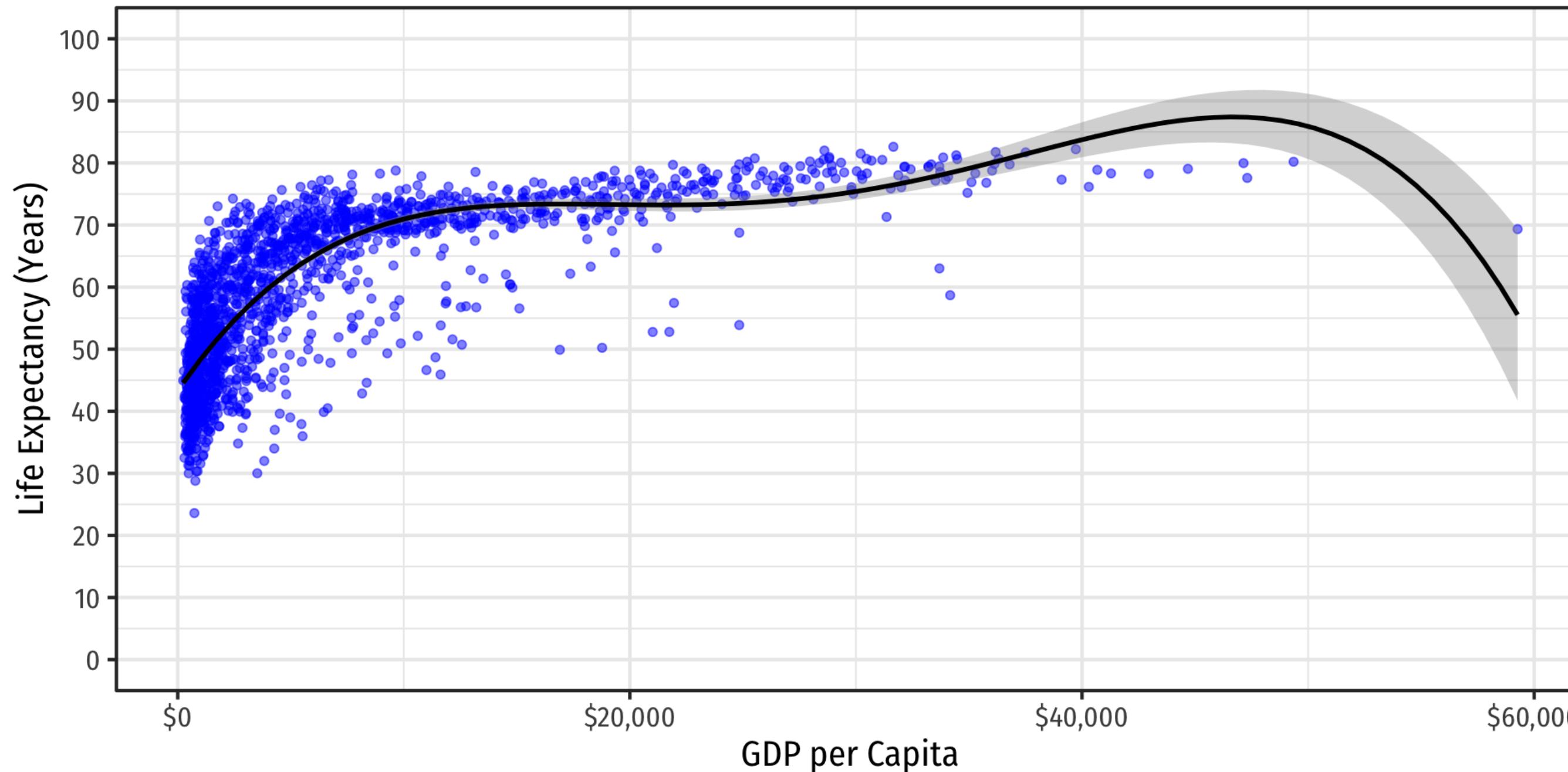
- It takes until a 9th-degree polynomial for one of the terms to become insignificant...
- ...but does this make the model *better? more interpretable?*
- A famous problem of **overfitting**



If You Kept Going...Visually



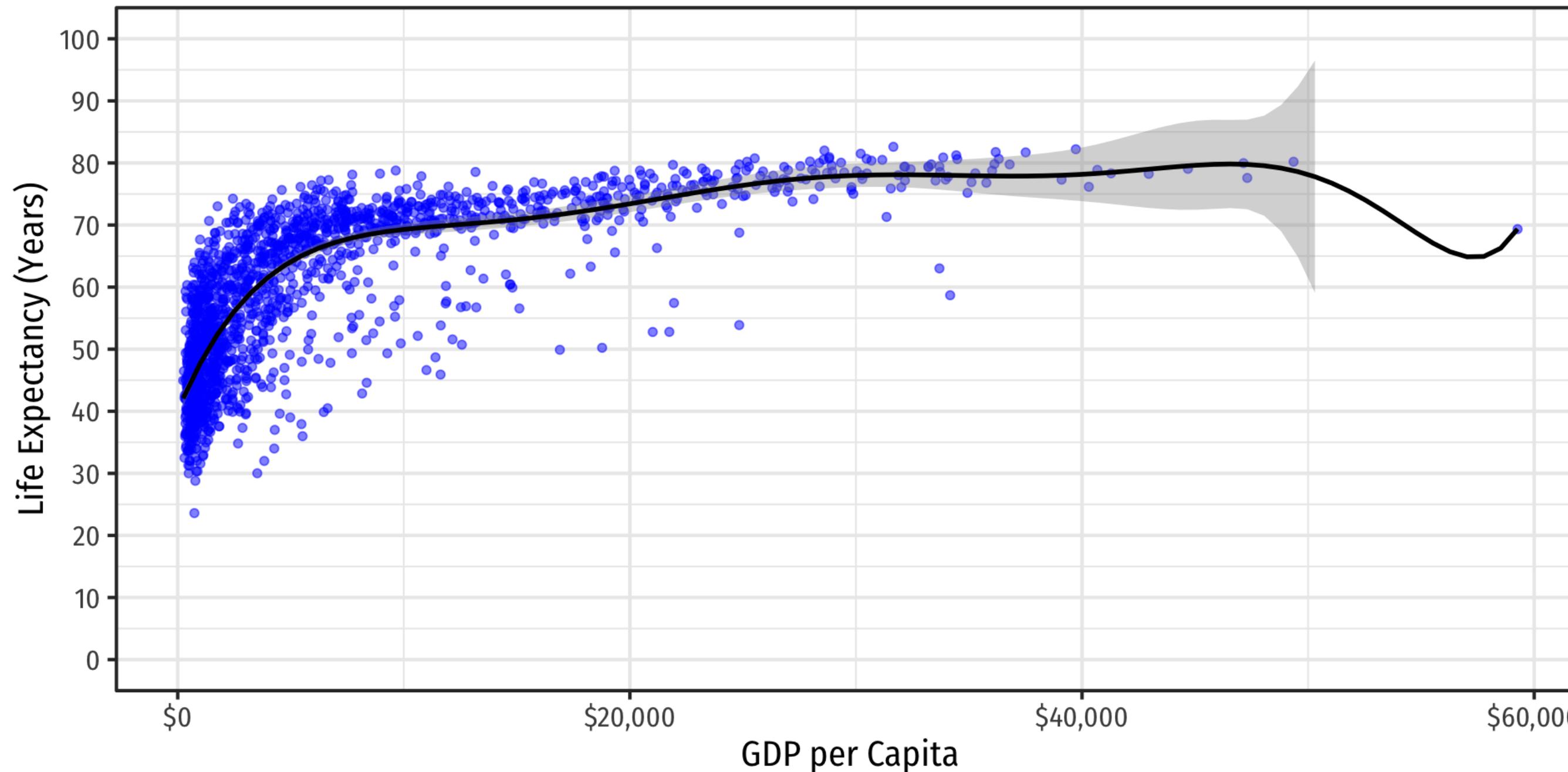
If You Kept Going...Visually



A 4th-degree polynomial



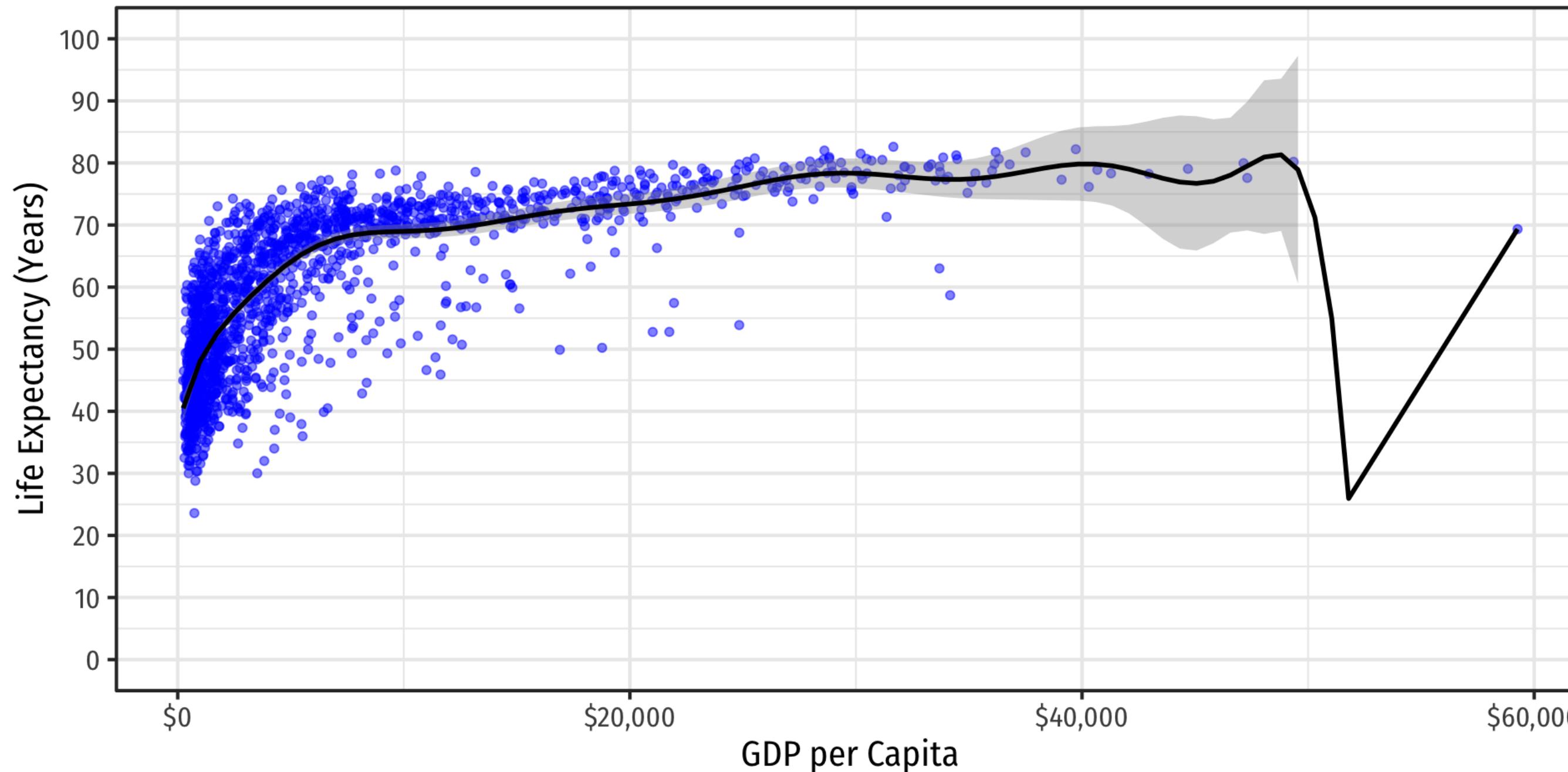
If You Kept Going...Visually



A 9th-degree polynomial



If You Kept Going...Visually



A 14th-degree polynomial



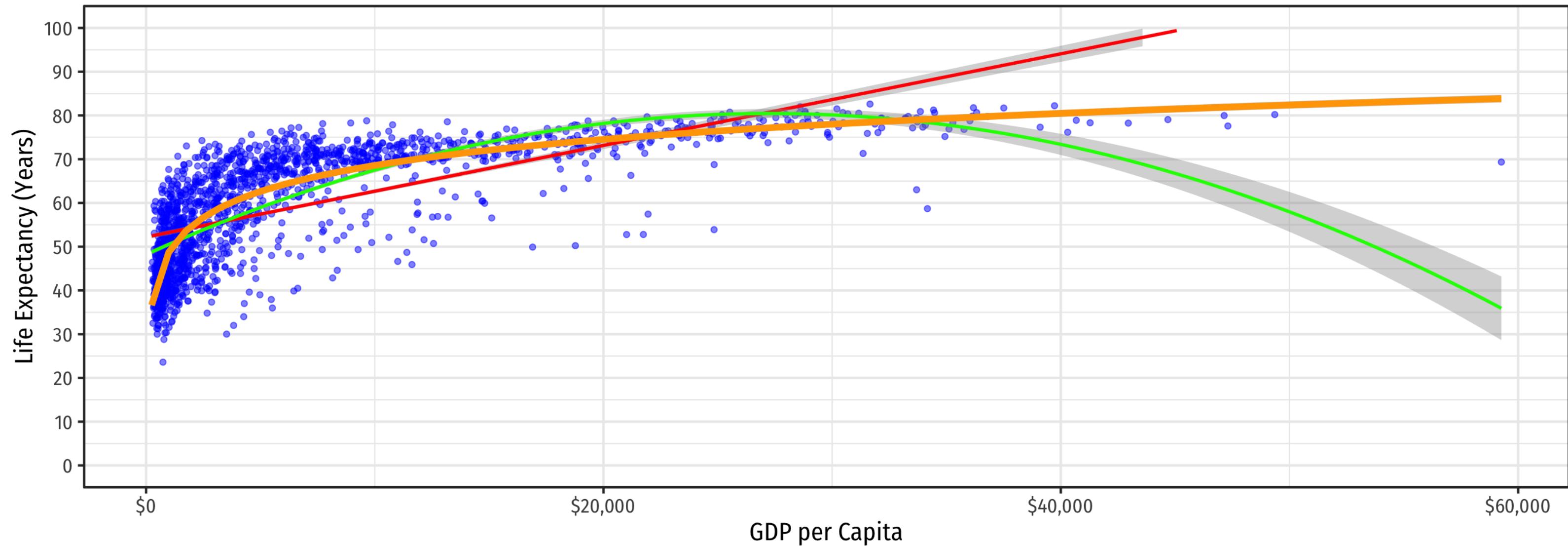
Strategy for Polynomial Model Specification

1. Are there good theoretical reasons for relationships changing (e.g. increasing/decreasing returns)?
2. Plot your data: does a straight line fit well enough?
3. Specify a polynomial function of a higher power (start with 2) and estimate OLS regression
4. Use t -test to determine if higher-power term is significant
5. Interpret effect of change in X on Y
6. Repeat steps 3-5 as necessary (if there are good theoretical reasons)



Logarithmic Models

Linear Regression



$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i$$

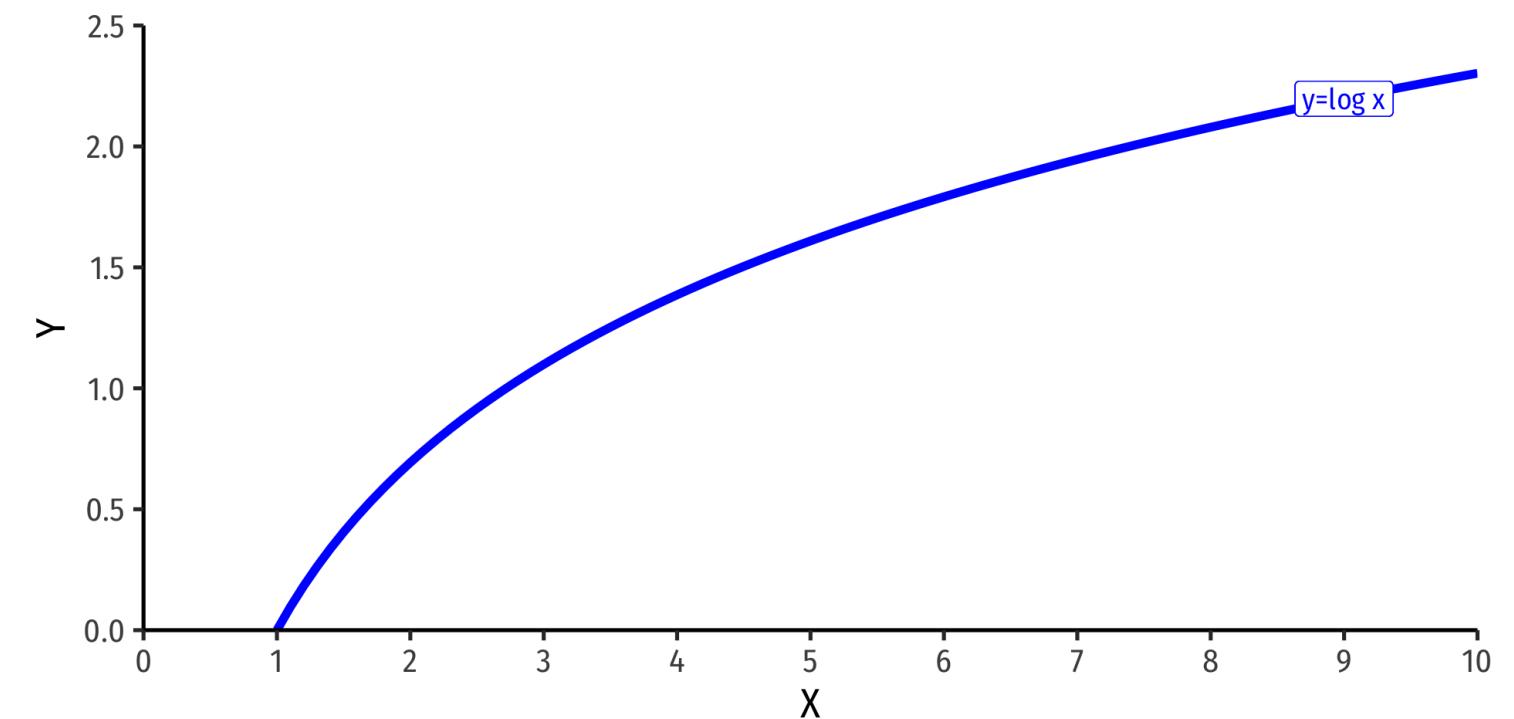
$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{GDP}_i + \hat{\beta}_2 \text{GDP}_i^2$$

$$\text{Life Expectancy}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln \text{GDP}_i$$



Logarithmic Models

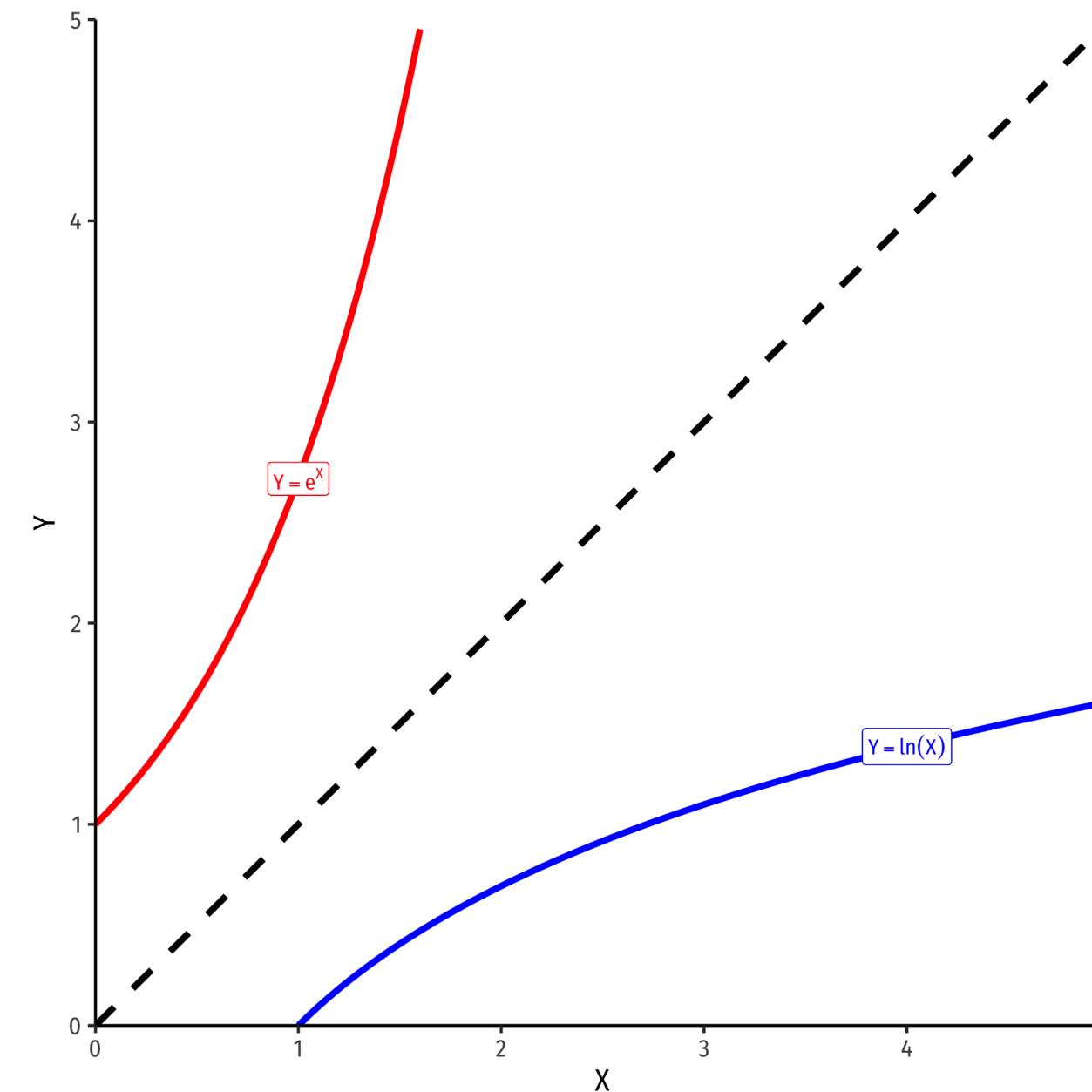
- Another useful model for nonlinear data is the **logarithmic model**¹
 - We transform either X , Y , or both by taking the **(natural) logarithm**
- Logarithmic model has two additional advantages
 1. We can easily interpret coefficients as **percentage changes** or **elasticities**
 2. Useful economic shape: diminishing returns (production functions, utility functions, etc)



¹ Don't confuse this with a **logistic (logit) model** for dependent dummy variables



The Natural Logarithm



- The **exponential function**, $Y = e^X$ or $Y = \exp(X)$, where base $e = 2.71828\dots$
- **Natural logarithm** is the inverse, $Y = \ln(X)$



The Natural Logarithm: Review I

- **Exponents** are defined as

$$b^n = \underbrace{b \times b \times \cdots \times b}_{n \text{ times}}$$

- where base b is multiplied by itself n times
- **Example:** $2^3 = \underbrace{2 \times 2 \times 2}_{n=3} = 8$
- **Logarithms** are the inverse, defined as the exponents in the expressions above

If $b^n = y$, then $\log_b(y) = n$

- n is the number you must raise b to in order to get y
- **Example:** $\log_2(8) = 3$



The Natural Logarithm: Review II

- Logarithms can have any base, but common to use the **natural logarithm** (\ln) with base $e = 2.71828\dots$

If $e^n = y$, then $\ln(y) = n$



The Natural Logarithm: Properties

- Natural logs have a lot of useful properties:

$$1. \ln\left(\frac{1}{x}\right) = -\ln(x)$$

$$2. \ln(ab) = \ln(a) + \ln(b)$$

$$3. \ln\left(\frac{x}{a}\right) = \ln(x) - \ln(a)$$

$$4. \ln(x^a) = a \ln(x)$$

$$5. \frac{d \ln x}{d x} = \frac{1}{x}$$



The Natural Logarithm: Example

- Most useful property: for small change in x , Δx :

$$\underbrace{\ln(x + \Delta x) - \ln(x)}_{\text{Difference in logs}} \approx \underbrace{\frac{\Delta x}{x}}_{\text{Relative change}}$$



Let $x = 100$ and $\Delta x = 1$, relative change is:

$$\frac{\Delta x}{x} = \frac{(101 - 100)}{100} = 0.01 \text{ or } 1\%$$

- The logged difference:

$$\ln(101) - \ln(100) = 0.00995 \approx 1\%$$

- This allows us to very easily interpret coefficients as **percent changes** or **elasticities**



Elasticity

- An **elasticity** between any two variables, $\epsilon_{Y,X}$ describes the **responsiveness** (in %) of one variable (Y) to a change in another (X)

$$\epsilon_{Y,X} = \frac{\% \Delta Y}{\% \Delta X} = \frac{\left(\frac{\Delta Y}{Y} \right)}{\left(\frac{\Delta X}{X} \right)}$$

- Numerator is relative change in Y , Denominator is relative change in X
- **Interpretation:** a 1% change in X will cause a $\epsilon_{Y,X}\%$ change in Y



Math FYI: Cobb Douglas Functions and Logs

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^\alpha K^\beta$$

- Taking logs, relationship becomes linear:

$$\ln(Y) = \ln(A) + \alpha \ln(L) + \beta \ln(K)$$

- With data on (Y, L, K) and linear regression, can estimate α and β
 - α : elasticity of Y with respect to L
 - A 1% change in L will lead to an $\alpha\%$ change in Y
 - β : elasticity of Y with respect to K
 - A 1% change in K will lead to a $\beta\%$ change in Y



Math FYI: Cobb Douglas Functions and Logs



Example

$$Y = 2L^{0.75}K^{0.25}$$

- Taking logs:

$$\ln Y = \ln 2 + 0.75 \ln L + 0.25 \ln K$$

- A 1% change in L will yield a 0.75% change in output Y
- A 1% change in K will yield a 0.25% change in output Y



Logarithms in R |

- The `log()` function can easily take the logarithm

```

1 gapminder <- gapminder %>%
2   mutate(loggdp = log(gdpPercap)) # log GDP per capita
3
4 gapminder %>% head() # look at it

```

country	continent	year
<fct>	<fct>	<int>
Afghanistan	Asia	1952
Afghanistan	Asia	1957
Afghanistan	Asia	1962
Afghanistan	Asia	1967
Afghanistan	Asia	1972
Afghanistan	Asia	1977

6 rows | 1-3 of 9 columns



Logarithms in R II

- Note, `log()` by default is the **natural logarithm** $\ln()$, i.e. base e
 - Can change base with e.g. `log(x, base = 5)`
 - Some common built-in logs: `log10`, `log2`

```
1 log10(100)
```

```
[1] 2
```

```
1 log2(16)
```

```
[1] 4
```

```
1 log(19683, base=3)
```

```
[1] 9
```



Logarithms in R III

- Note when running a regression, you can pre-transform the data into logs (as I did above), or just add `log()` around a variable in the regression

term	estimate	std.error
	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674
loggdp	8.405085	0.148762
2 rows 1-3 of 5 columns		



Types of Logarithmic Models

- Three types of log regression models, depending on which variables we log

1. **Linear-log model:** $Y_i = \beta_0 + \beta_1 \ln X_i$

2. **Log-linear model:** $\ln Y_i = \beta_0 + \beta_1 X_i$

3. **Log-log model:** $\ln Y_i = \beta_0 + \beta_1 \ln X_i$



Linear-Log Model

Linear-Log Model: Interpretation

- **Linear-log model** has an independent variable (X) that is logged

$$Y = \beta_0 + \beta_1 \ln X_i$$

$$\beta_1 = \frac{\Delta Y}{\left(\frac{\Delta X}{X} \right)}$$

- **Marginal effect of $X \rightarrow Y$:** a 1% change in $X \rightarrow$ a $\frac{\beta_1}{100}$ unit change in Y



Linear-Log Model in R

term	estimate	std.error	statistic
<chr>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117
loggdp	8.405085	0.148762	56.500206

2 rows | 1-4 of 5 columns

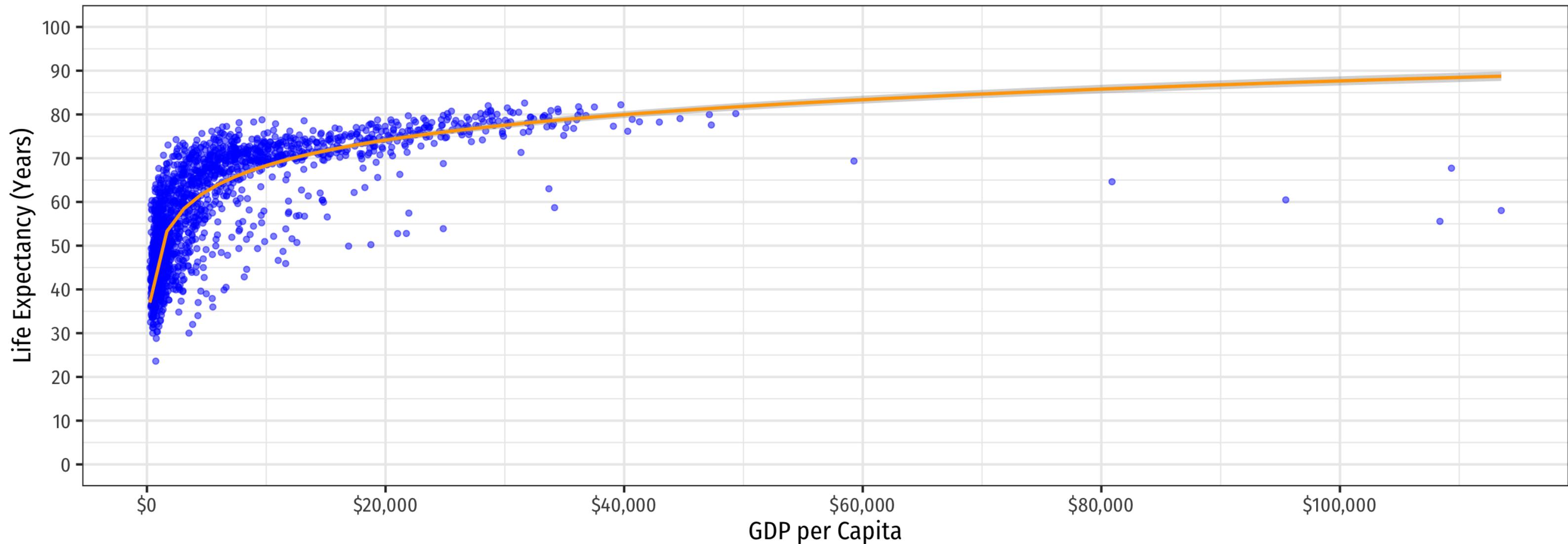
$$\widehat{\text{Life Expectancy}_i} = -9.10 + 8.41 \ln \text{GDP}_i$$

- A **1% change in GDP** → a $\frac{9.41}{100} = 0.0841$ **year increase** in Life Expectancy
- A **25% fall in GDP** → a $(-25 \times 0.0841) = 2.1025$ **year decrease** in Life Expectancy
- A **100% rise in GDP** → a $(100 \times 0.0841) = 8.4100$ **year increase** in Life Expectancy



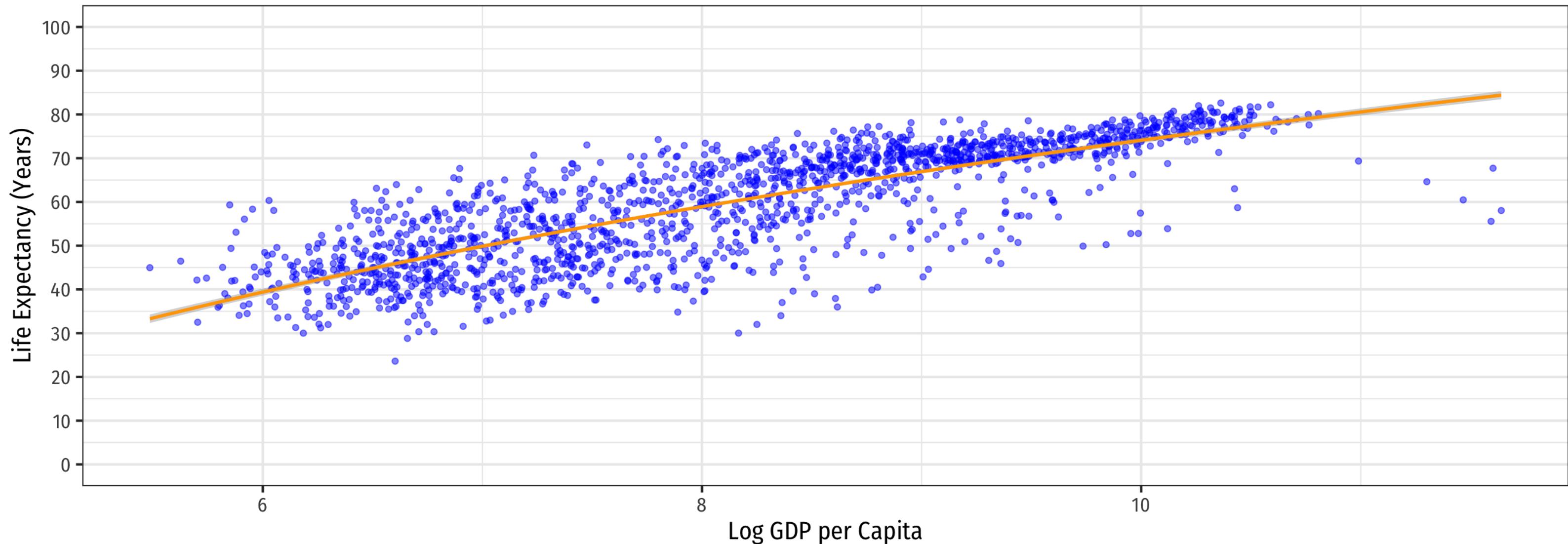
Linear-Log Model Graph (Linear X-Axis)

► Code



Linear-Log Model Graph (Log X-Axis)

► Code



Log-Linear Model

Log-Linear Model: Interpretation

- **Log-linear model** has the dependent variable (Y) logged

$$\ln Y_i = \beta_0 + \beta_1 X$$
$$\beta_1 = \frac{\left(\frac{\Delta Y}{Y} \right)}{\Delta X}$$

- **Marginal effect of $X \rightarrow Y$:** a 1 unit change in $X \rightarrow$ a $\beta_1 \times 100\%$ change in Y



Log-Linear Model in R (Preliminaries)

- We will again have very large/small coefficients if we deal with GDP directly, again let's transform `gdpPerCap` into \$1,000s, call it `gdp_t`
- Then log LifeExp

```

1 gapminder <- gapminder %>%
2   mutate(gdp_t = gdpPerCap/1000, # first make GDP/capita in $1000s
3         loglife = log(lifeExp)) # take the log of LifeExp
4 gapminder %>% head() # look at it

```

country <fct>	continent <fct>	year <int>
Afghanistan	Asia	1952
Afghanistan	Asia	1957
Afghanistan	Asia	1962
Afghanistan	Asia	1967
Afghanistan	Asia	1972



country	continent	year
<fct>	<fct>	<int>
Afghanistan	Asia	1977
6 rows 1-3 of 11 columns		



Log-Linear Model in R

term	estimate	std.error	statistic
<chr>	<dbl>	<dbl>	<dbl>
(Intercept)	3.966639	0.0058345501	679.85339
gdp_t	0.012917	0.0004777072	27.03958

2 rows | 1-4 of 5 columns

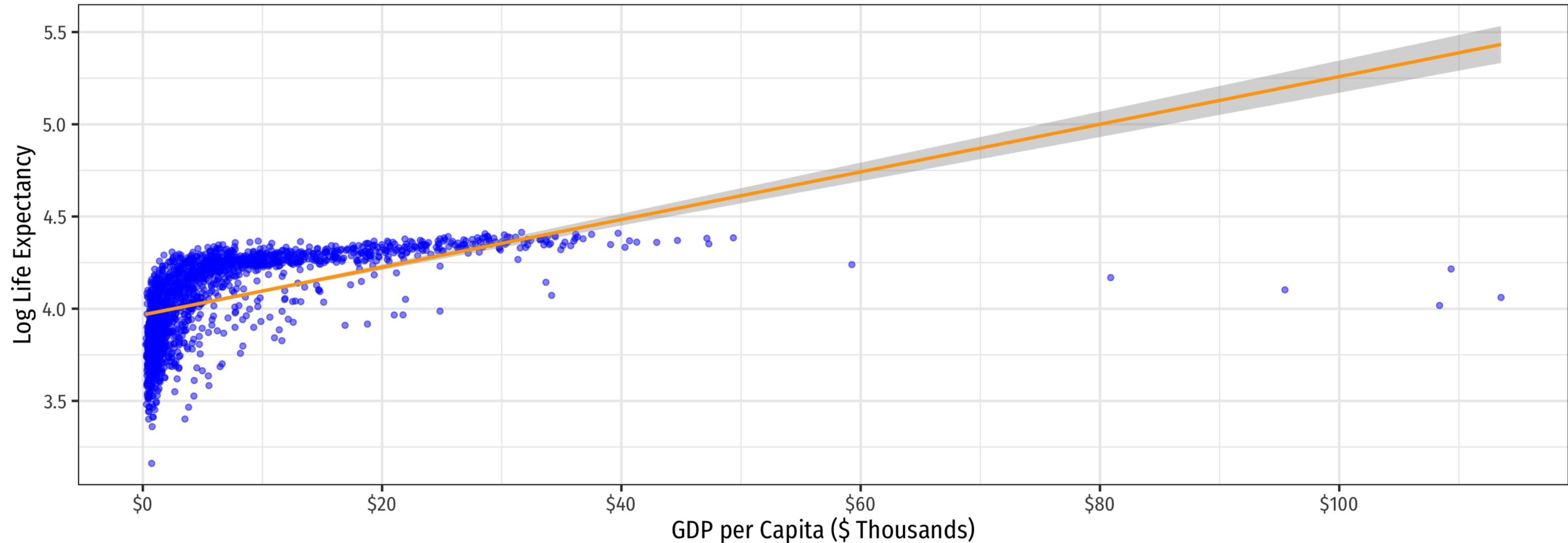
$$\widehat{\ln \text{Life Expectancy}_i} = 3.967 + 0.013 \text{ GDP}_i$$

- A **\$1 (thousand) change in GDP** → a $0.013 \times 100\% = 1.3\%$ **increase** in Life Expectancy
- A **\$25 (thousand) fall in GDP** → a $(-25 \times 1.3\%) = 32.5\%$ **decrease** in Life Expectancy
- A **\$100 (thousand) rise in GDP** → a $(100 \times 1.3\%) = 130\%$ **increase** in Life Expectancy



Linear-Log Model Graph

► Code



Log-Log Model

Log-Log Model

- **Log-log model** has both variables (X and Y) logged

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i$$

$$\beta_1 = \frac{\left(\frac{\Delta Y}{Y} \right)}{\left(\frac{\Delta X}{X} \right)}$$

- **Marginal effect of $X \rightarrow Y$: a 1% change in $X \rightarrow$ a β_1 % change in Y**
- β_1 is the **elasticity** of Y with respect to X !



Log-Log Model in R

term	estimate	std.error	statistic
<chr>	<dbl>	<dbl>	<dbl>
(Intercept)	2.864177	0.02328274	123.01718
loggdp	0.146549	0.00282126	51.94452

2 rows | 1-4 of 5 columns

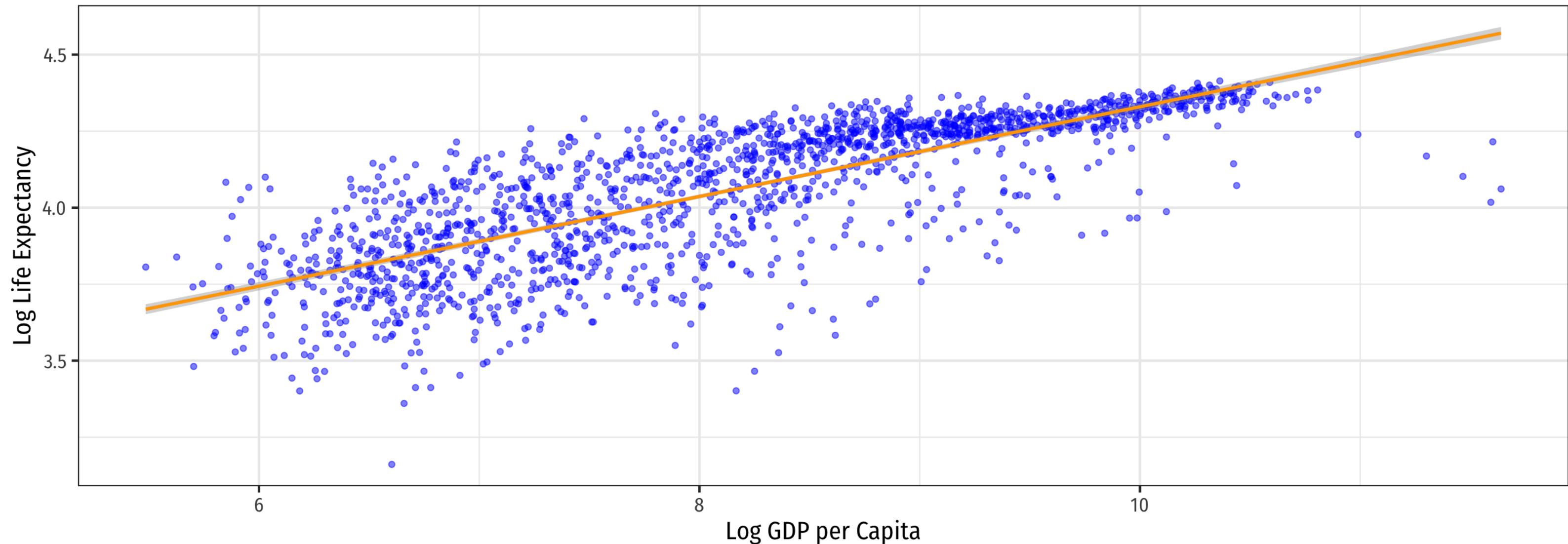
$$\widehat{\ln \text{Life Expectancy}_i} = 2.864 + 0.147 \ln \text{GDP}_i$$

- A **1% change in GDP** → a **0.147% increase** in Life Expectancy
- A **25% fall in GDP** → a $(-25 \times 0.147\%) = 3.675\%$ **decrease** in Life Expectancy
- A **100% rise in GDP** → a $(100 \times 0.147\%) = 14.7\%$ **increase** in Life Expectancy



Log-Log Model Graph

► Code



Comparing Log Models I

Model	Equation	Interpretation
Linear- Log	$Y = \beta_0 + \beta_1 \ln X$	1% change in $X \rightarrow \frac{\hat{\beta}_1}{100}$ unit change in Y
Log -Linear	$\ln Y = \beta_0 + \beta_1 X$	1 unit change in $X \rightarrow \hat{\beta}_1 \times 100\%$ change in Y

Log-Log	$\ln Y = \beta_0 + \beta_1 \ln X$	1% change in $X \rightarrow \hat{\beta}_1\%$ change in Y
----------------	-----------------------------------	--

- Hint: the variable that gets **logged** changes in **percent** terms, the **linear** variable (not logged) changes in **unit** terms
 - Going from units \rightarrow percent: multiply by 100
 - Going from percent \rightarrow units: divide by 100



Comparing Models II

► Code

	Life Exp.	Log Life Exp.	Log Life Exp.
Constant	-9.10*** (1.23)	3.97*** (0.01)	2.86*** (0.02)
Log GDP per Capita	8.41*** (0.15)		0.15*** (0.00)
GDP per capita (\$1,000s)		0.01*** (0.00)	
n	1704	1704	1704
Adj. R ²	0.65	0.30	0.61
SER	7.62	0.19	0.14

* p < 0.1, ** p < 0.05, *** p < 0.01

- Models are very different units, how to choose?

1. Compare intuition
2. Compare R^2 's
3. Compare graphs



Comparing Models III

Linear-Log

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln X_i$$

$$R^2 = 0.65$$

Log-Linear

$$\ln Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$R^2 = 0.30$$

Log-Log

$$\ln Y_i = \hat{\beta}_0 + \hat{\beta}_1 \ln X_i$$

$$R^2 = 0.61$$



When to Log?

- In practice, the following types of variables are usually logged:
 - Variables that must always be **positive** (prices, sales, market values)
 - **Very large** numbers (population, GDP)
 - Variables we want to talk about as **percentage changes or growth rates** (money supply, population, GDP)
 - Variables that have **diminishing returns** (output, utility)
 - Variables that have nonlinear scatterplots
- *Avoid* logs for:
 - Variables that are less than one, decimals, 0, or negative
 - Categorical variables (season, gender, political party)
 - Time variables (year, week, day)



Standardizing & Comparing Across Units

Comparing Coefficients of Different Units I

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- We often want to compare coefficients to see which variable X_1 or X_2 has a bigger effect on Y
- What if X_1 and X_2 are different units?

Example

$$\widehat{\text{Salary}}_i = \beta_0 + \beta_1 \text{Batting average}_i + \beta_2 \text{Home runs}_i$$

$$\widehat{\text{Salary}}_i = -2,869,439.40 + 12,417,629.72 \text{Batting average}_i + 129,627.36 \text{Home runs}_i$$



Comparing Coefficients of Different Units II

- An easy way is to **standardize**¹ the variables (i.e. take the Z-score)

$$X_Z = \frac{X_i - \bar{X}}{sd(X)}$$

- Note doing this will make the constant 0, as both distributions of X and Y are now centered at 0.

¹ Also called “centering” or “scaling”



Comparing Coefficients of Different Units: Example

Variable	Mean	Std. Dev.
Salary	\$2,024,616	\$2,764,512
Batting Average	0.267	0.031
Home Runs	12.11	10.31

$$\widehat{\text{Salary}_i} = -2,869,439.40 + 12,417,629.72 \text{ Batting average}_i + 129,627.36 \text{ Home runs}_i$$

$$\widehat{\text{Salary}_Z} = 0.00 + 0.14 \text{ Batting average}_Z + 0.48 \text{ Home runs}_Z$$

- **Marginal effects** on Y (in standard deviations of Y) from 1 standard deviation change in X :
- $\hat{\beta}_1$: a 1 standard deviation increase in Batting Average increases Salary by 0.14 standard deviations

$$0.14 \times \$2,764,512 = \$387,032$$

- $\hat{\beta}_2$: a 1 standard deviation increase in Home Runs increases Salary by 0.48 standard deviations

$$0.48 \times \$2,764,512 = \$1,326,966$$



Standardizing in R

Variable	Mean	SD
LifeExp	59.47	12.92
gdpPercap	\$7215.32	\$9857.46

- Use the `scale()` command inside `mutate()` function to standardize a variable

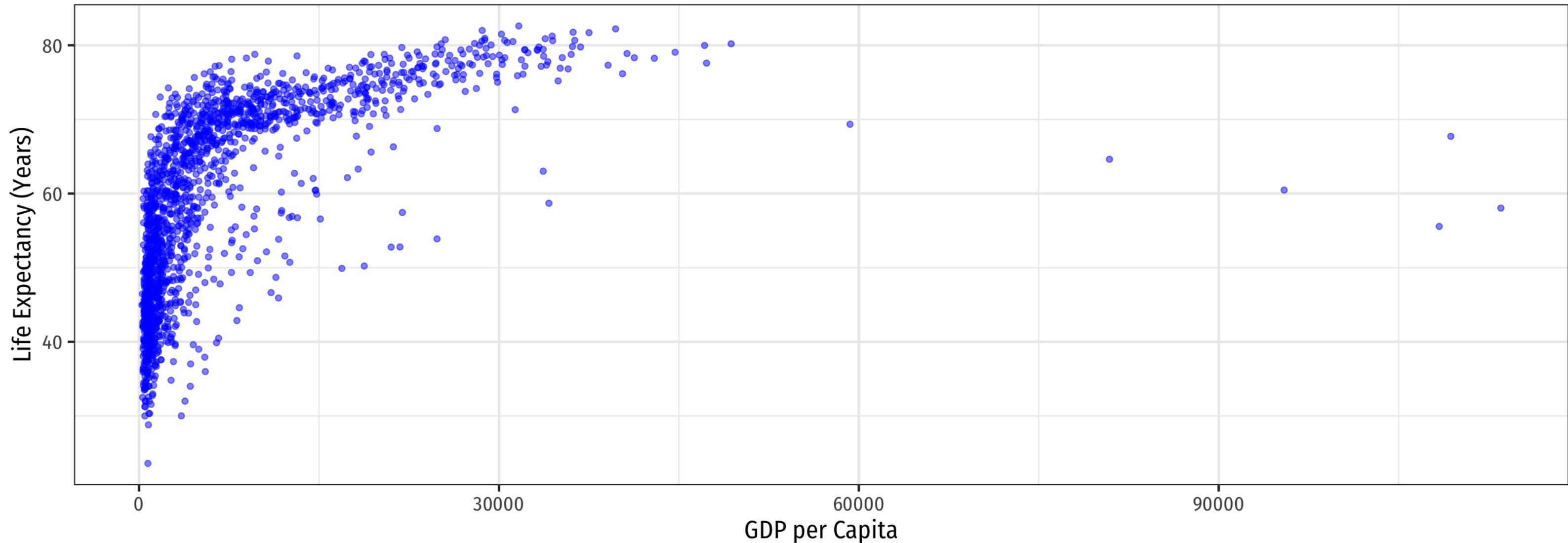
► Code

term	estimate
	<dbl>
(Intercept)	1.095650e-16
gdp_Z	5.837062e-01
2 rows 1-2 of 5 columns	



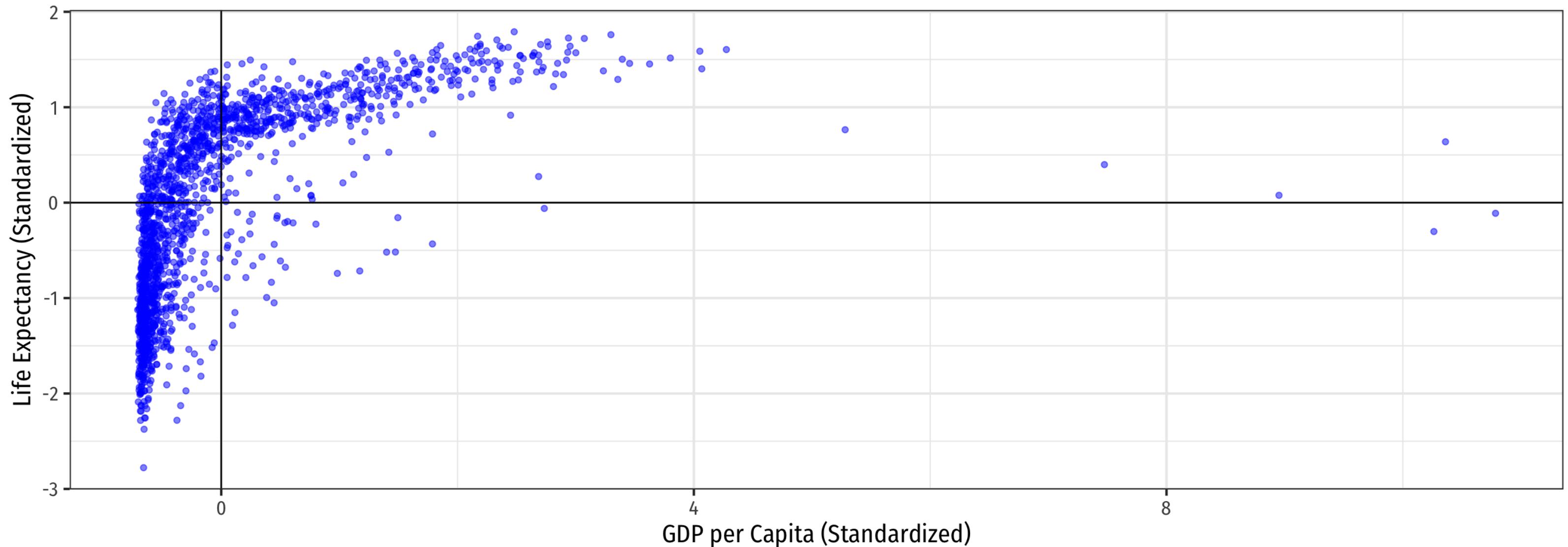
Rescaling: Visually

► Code



Rescaling: Visually

► Code



Rescaling: Visually

- Both X and Y now have means of 0 and sd of 1
- Code

mean_gdp	sd_gdp	mean_life
<dbl>	<dbl>	<dbl>
0	1	0
1 row 1-3 of 4 columns		



Joint Hypothesis Testing