3.8 — Polynomial Regression

ECON 480 • Econometrics • Fall 2021

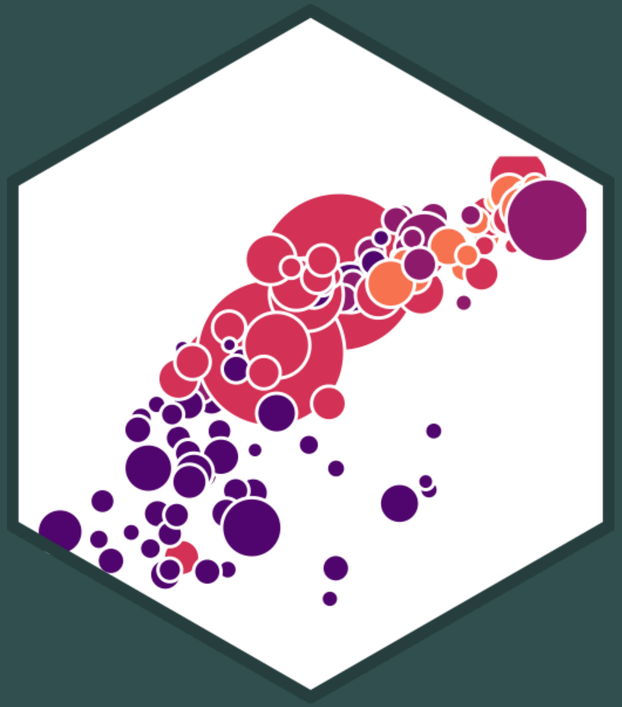
Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF21

metricsF21.classes.ryansafner.com

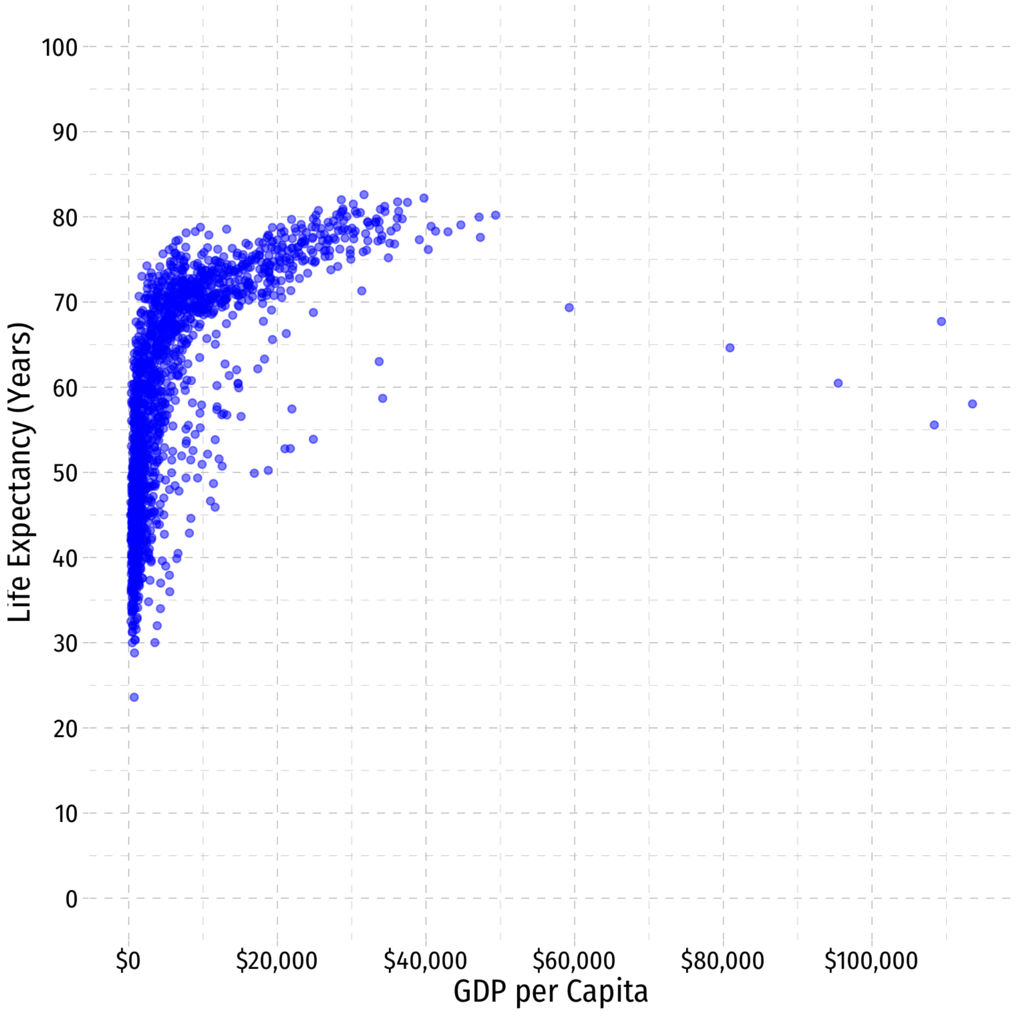


Outline

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

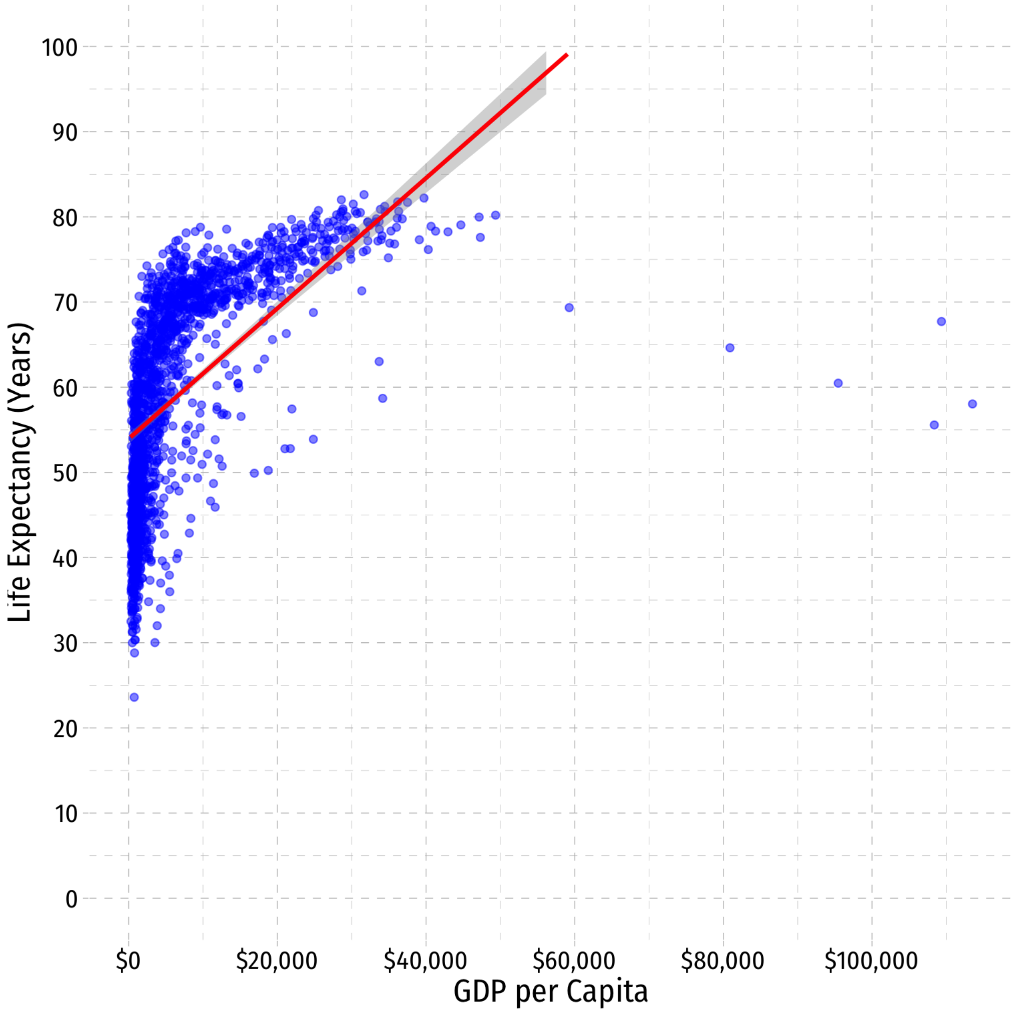


Linear Regression

- 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

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



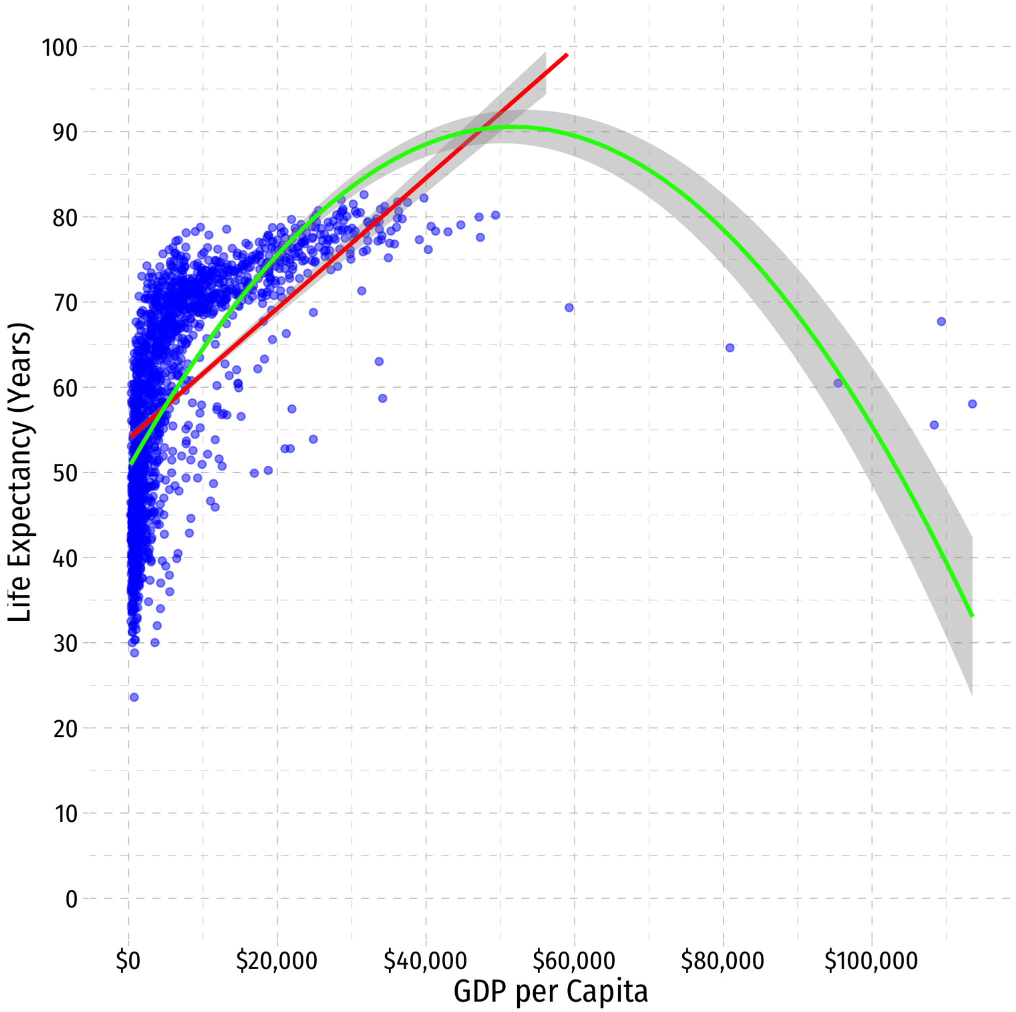
Linear Regression

- 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

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

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

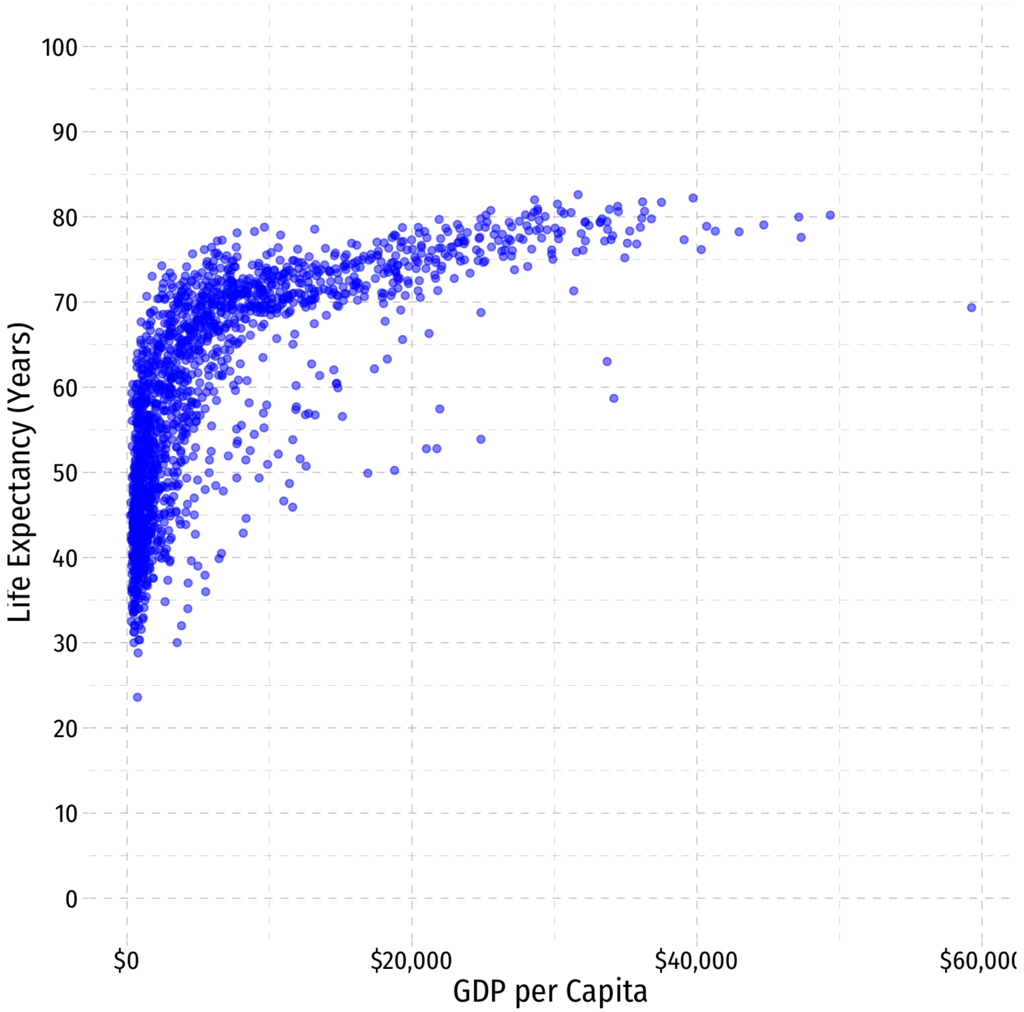


Linear Regression

- 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

- Get rid of the outliers (>$60,000)



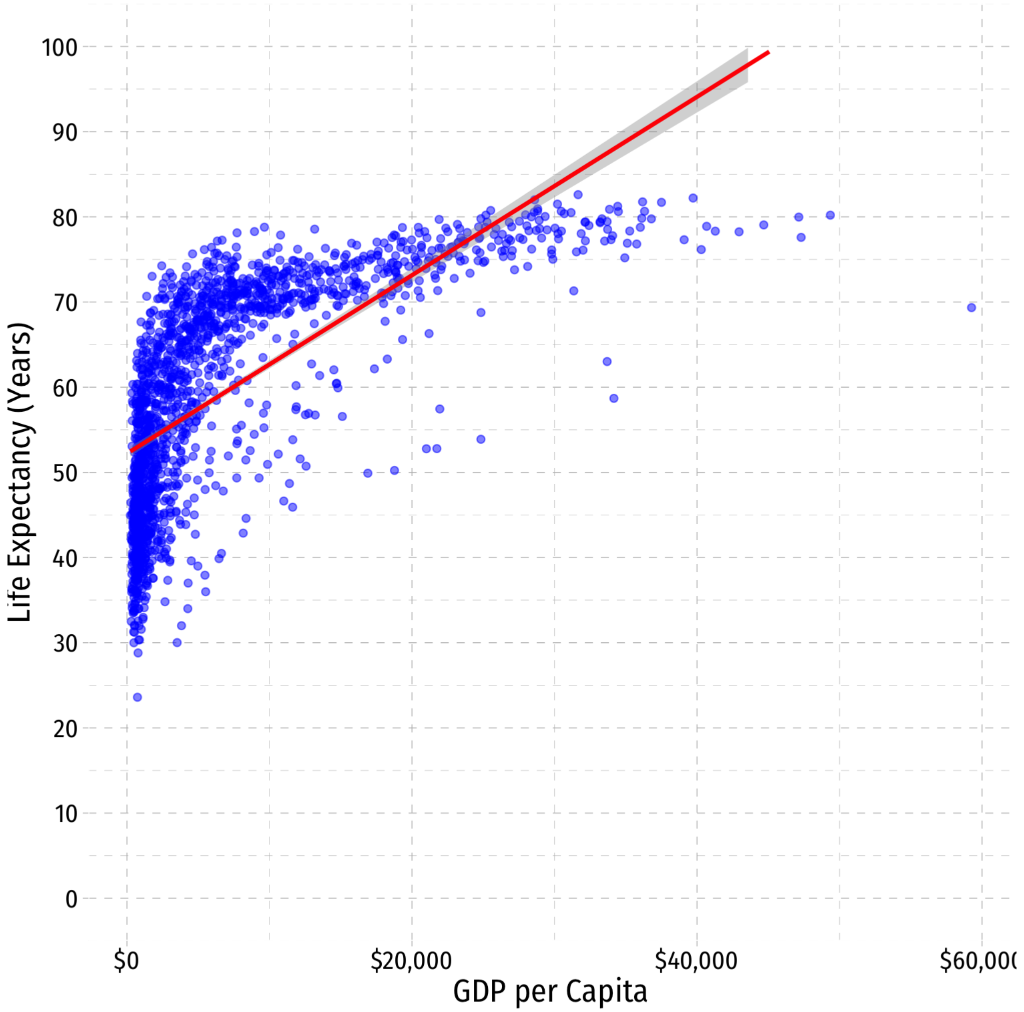
Linear Regression

- 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

- Get rid of the outliers (>$60,000)

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



Linear Regression

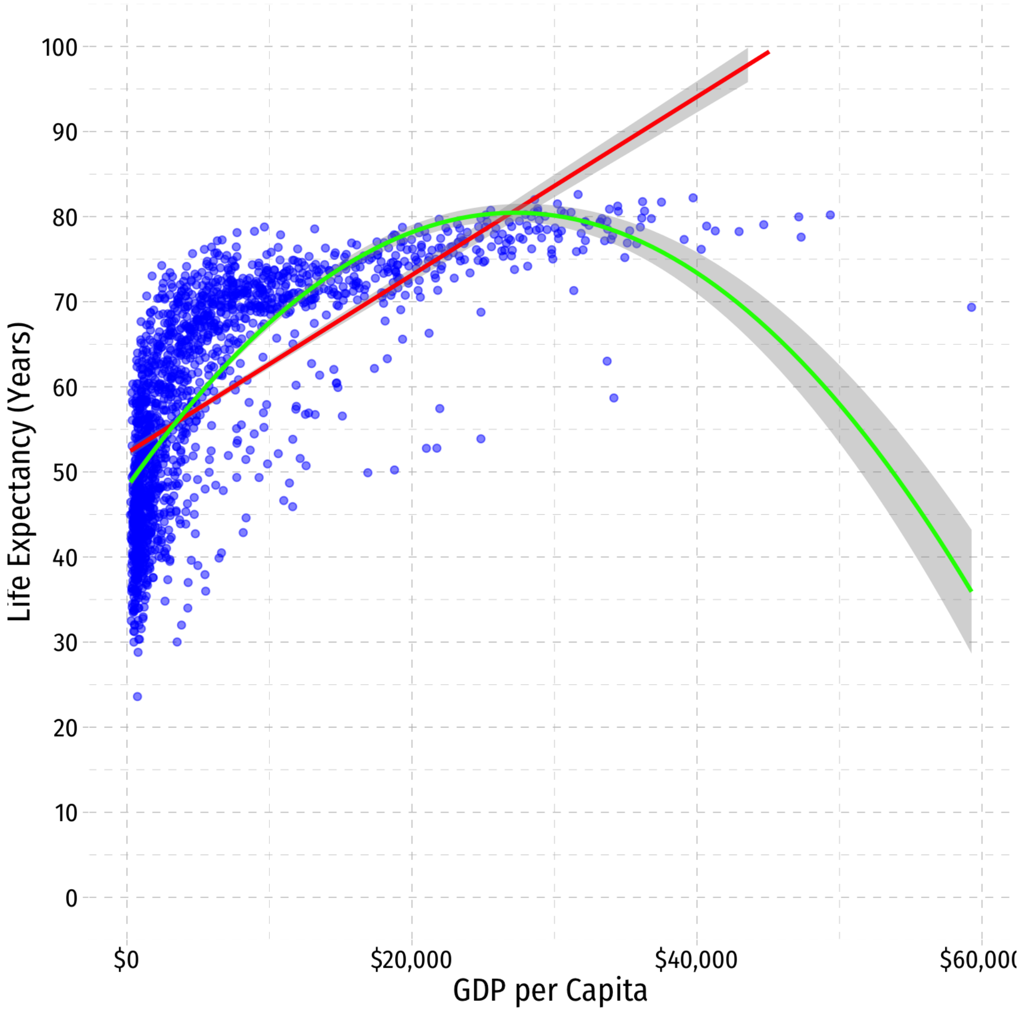
- OLS is commonly known as "**linear_ regression**" as it fits a **straight line** to data points

- Often, data and relationships between variables may *not* be linear

- Get rid of the outliers (>$60,000)

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

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



Linear Regression

- OLS is commonly known as "**linear_ regression**" as it fits a **straight line** to data points

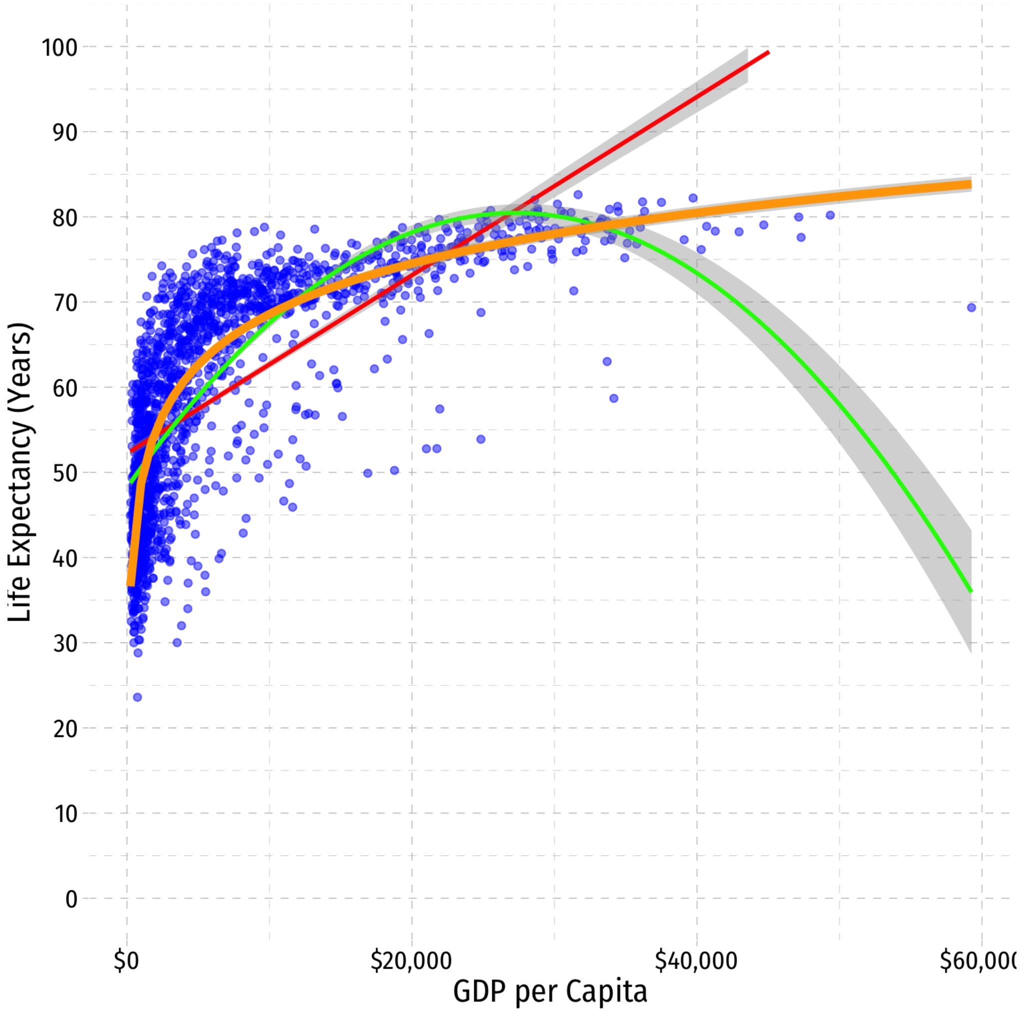
- Often, data and relationships between variables may *not* be linear

- Get rid of the outliers (>$60,000)

$$\widehat{\text{Life Expectancy}}_i = \hat{\beta_0} + \hat{\beta_1}\text{GDP}_i$$

$$\widehat{\text{Life Expectancy}}_i = \hat{\beta_0} + \hat{\beta_1}\text{GDP}_i + \hat{\beta_2}\text{GDP}_i^2$$

$$\widehat{\text{Life Expectancy}}_i = \hat{\beta_0} + \hat{\beta_1}\ln(\text{GDP}_i)$$



Nonlinear Effects in Linear Regression

- Despite being "linear regression", OLS can handle this with an easy fix

- OLS requires all *parameters* (i.e. the β's) to be linear, the *regressors* $(X$'s) can be nonlinear:

$$Y_i = \beta_0 + \beta_1 X_i^2 \quad ✅$$

$$Y_i = \beta_0 + \beta_1^2 X_i \quad ❌$$

$$Y_i = \beta_0 + \beta_1 \sqrt{X_i} \quad ✅$$

$$Y_i = \beta_0 + \sqrt{\beta_1} X_i \quad ❌$$

$$Y_i = \beta_0 + \beta_1 (X_{1i} \times X_{2i}) \quad ✅$$

$$Y_i = \beta_0 + \beta_1 ln(X_i) \quad ✅$$

- In the end, each X is always just a number in the data, OLS can always estimate parameters for it; but *plotting* the modelled points $(X_i, \hat{Y}_i)$ can result in a curve!

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

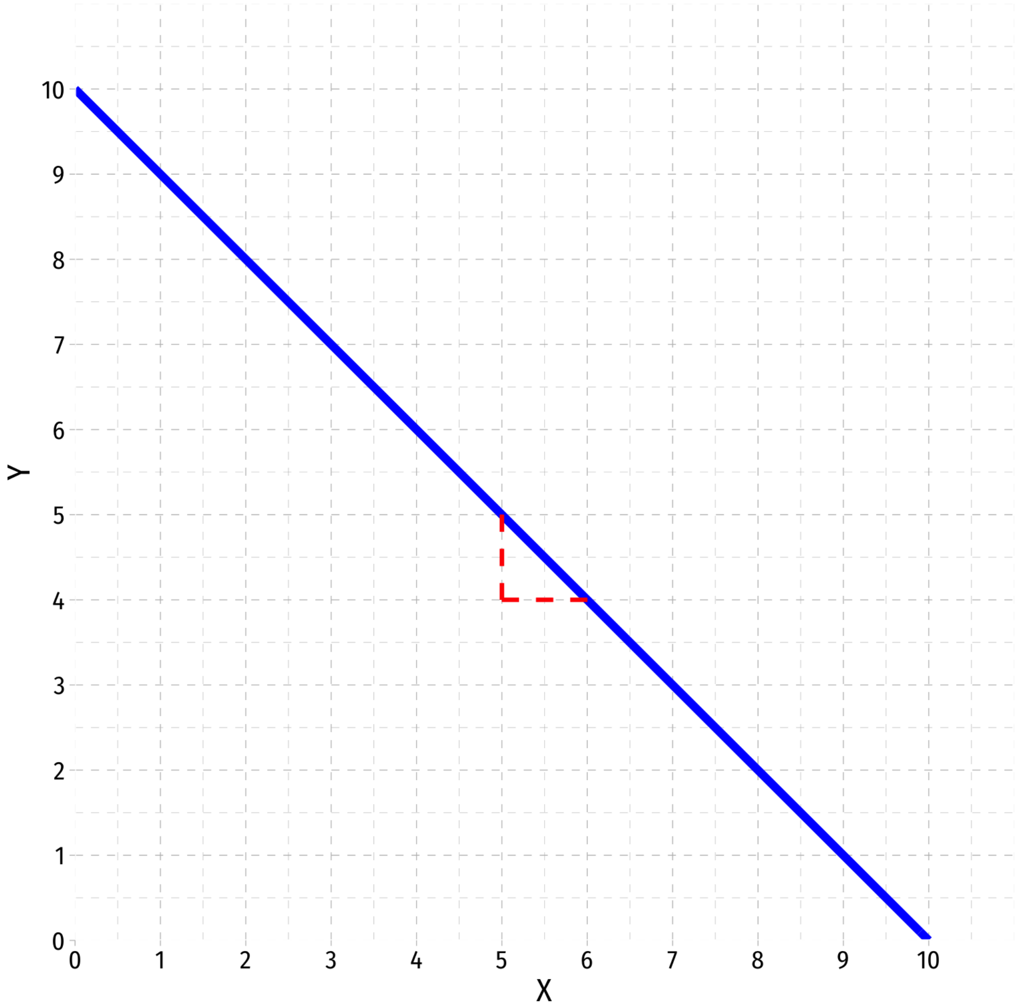
2. $X_1 \rightarrow Y$ is different for different levels of X_2
    - e.g. interaction effects (last lesson)

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

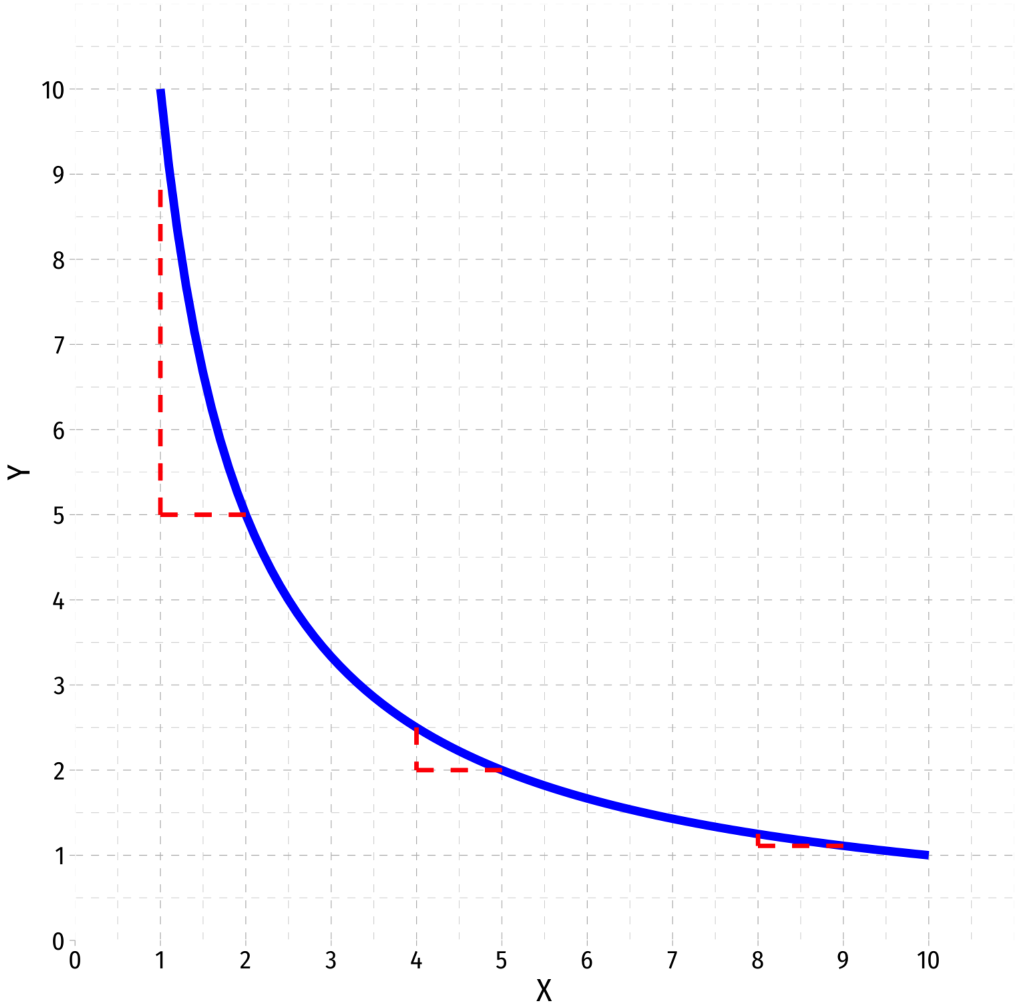


Nonlinearities Alter Marginal Effects

- **Polynomial**:

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

- Marginal effect, "slope" $\left(\neq \hat{\beta}_1 \right)$ *depends on the value of* X!



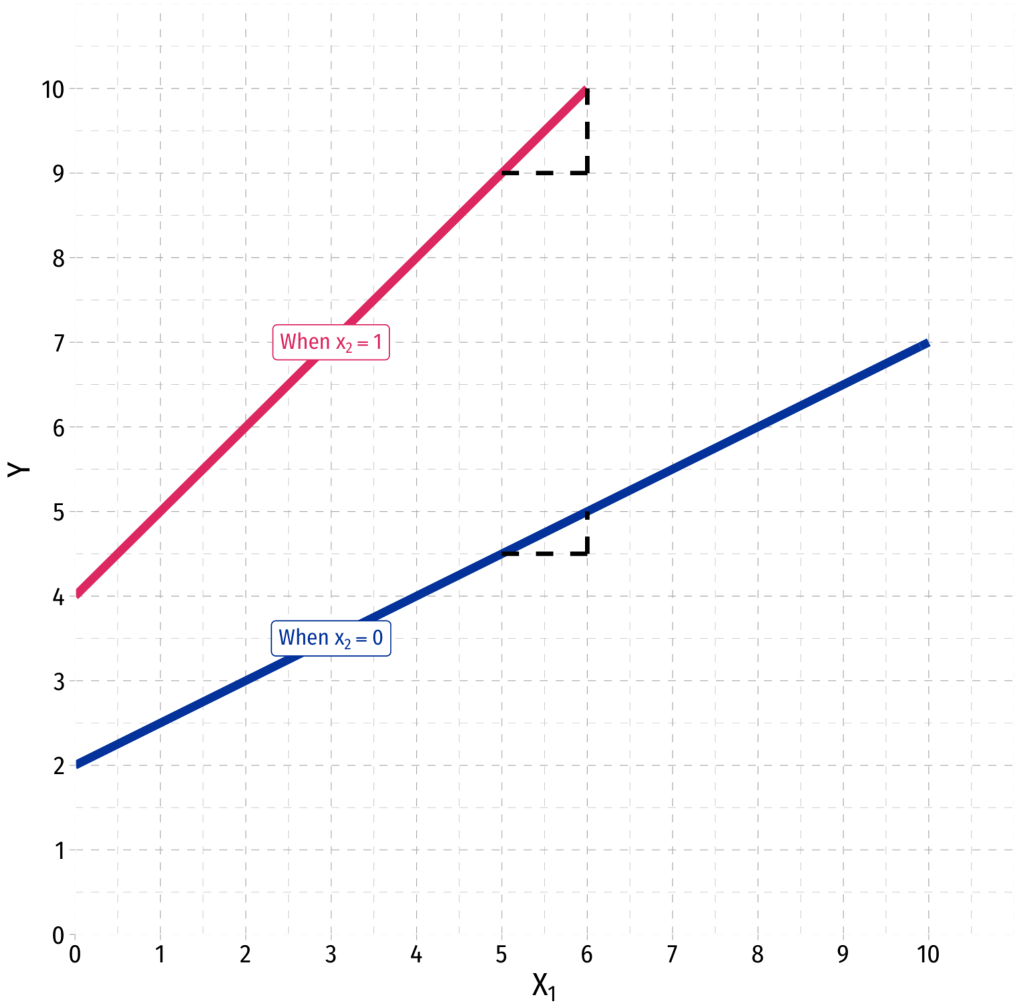
Sources of Nonlinearities III

- **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!

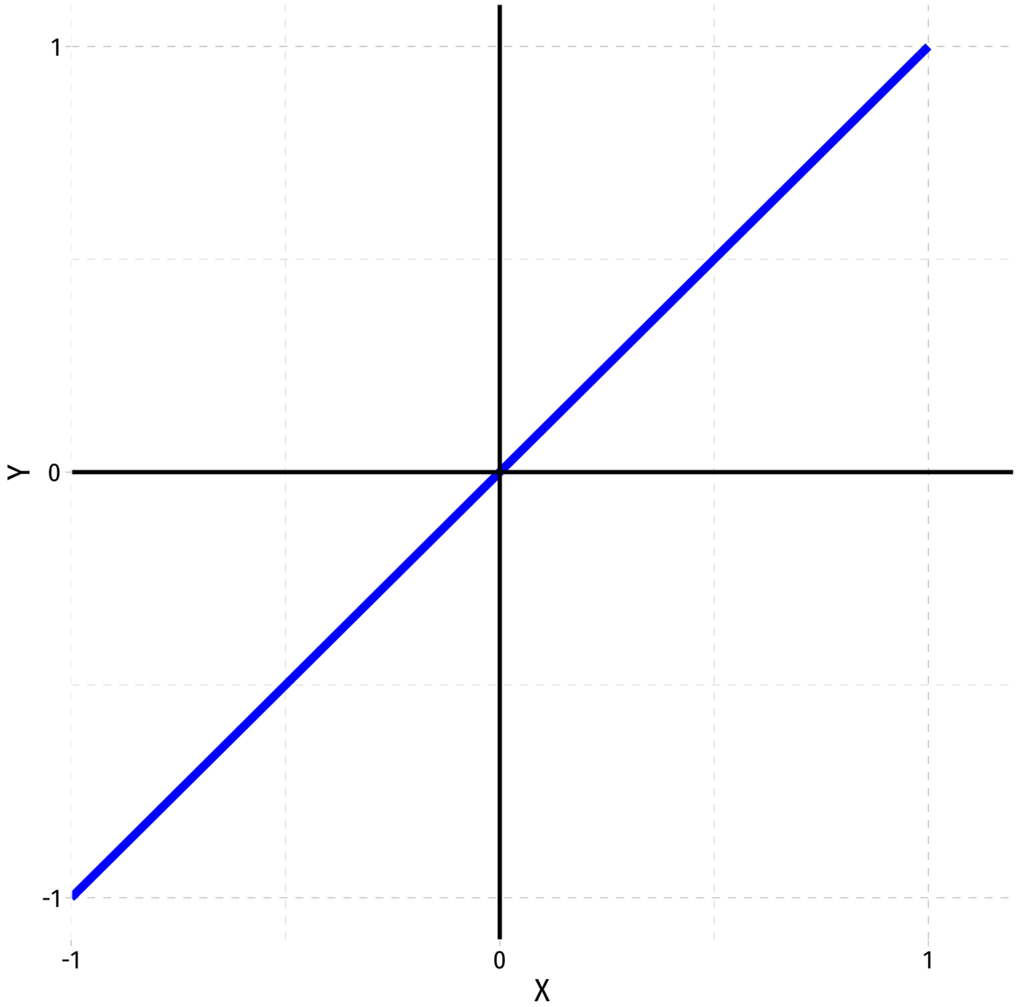
- Easy example: if X_2 is a dummy variable:
    - $X_2 = 0$ (control) vs. $X_2 = 1$ (treatment)



Polynomial Functions of X I

- Linear

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



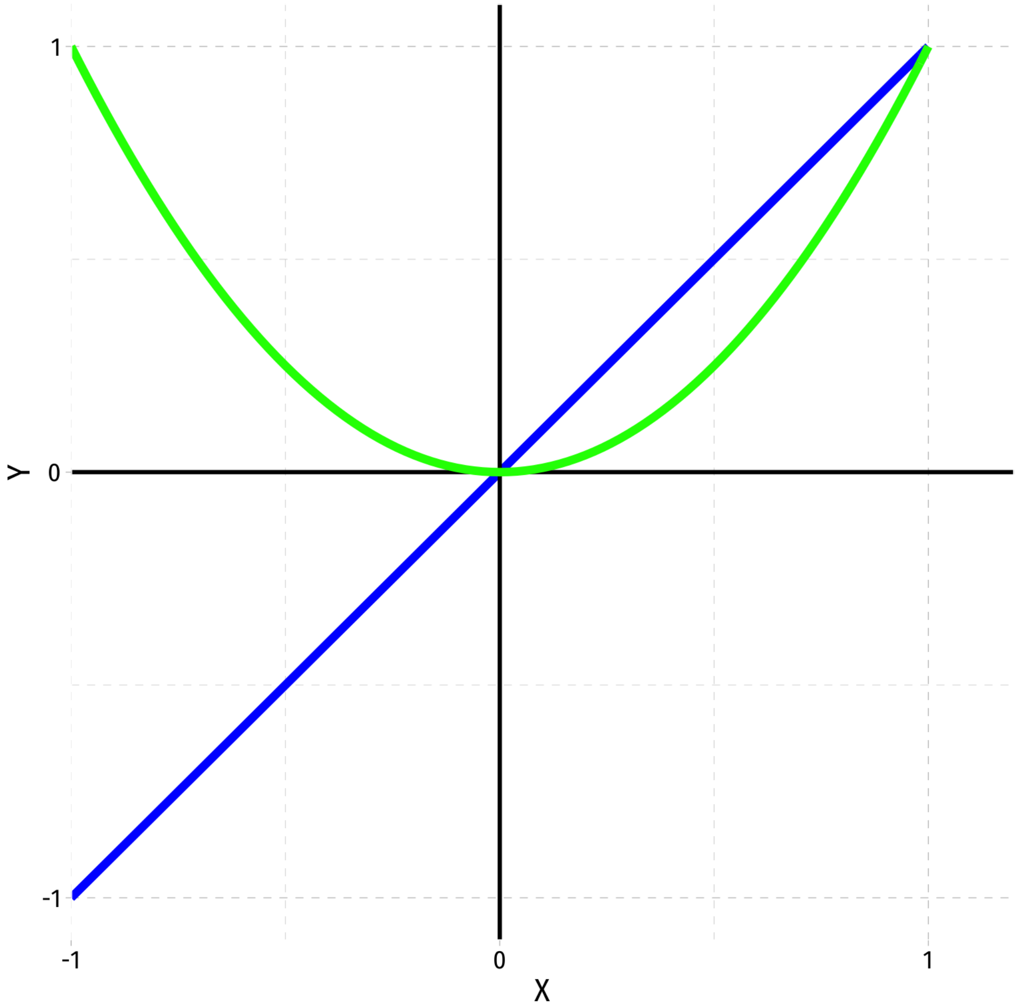
Polynomial Functions of X I

- 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

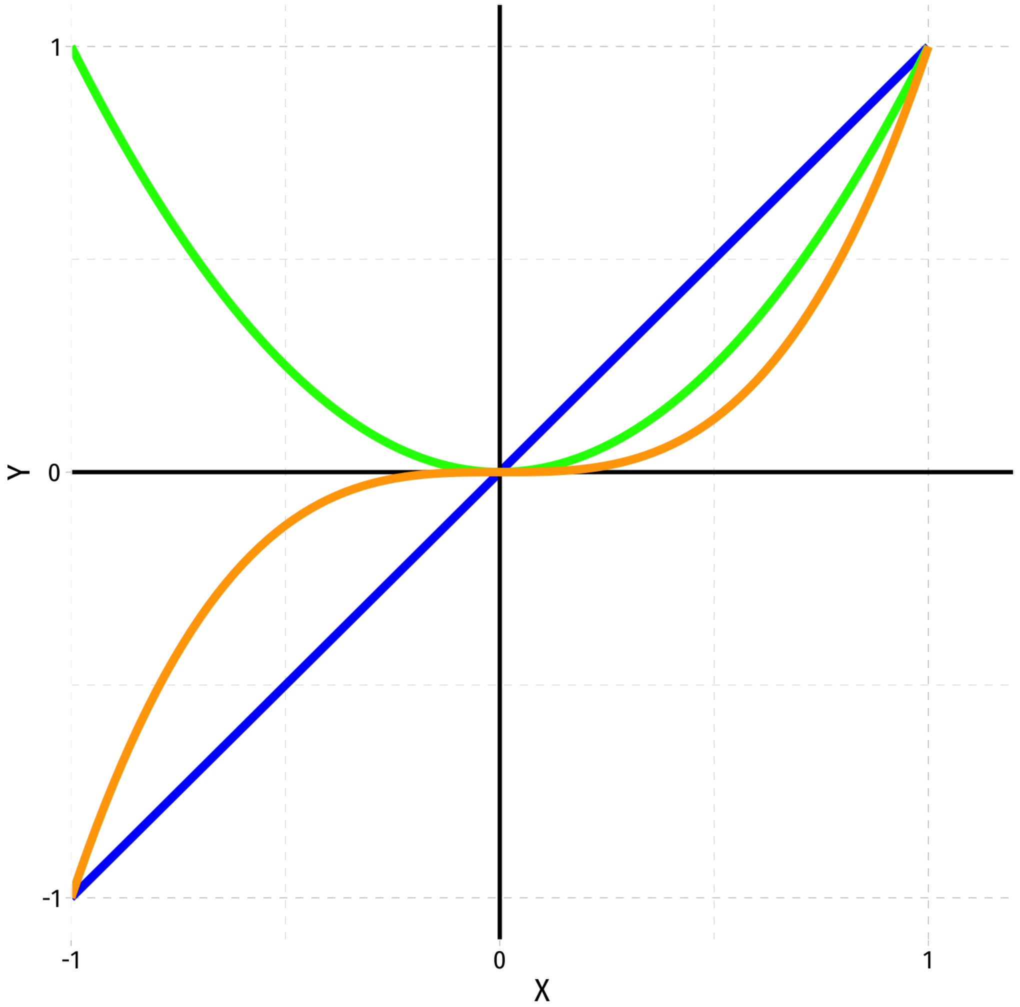
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- 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

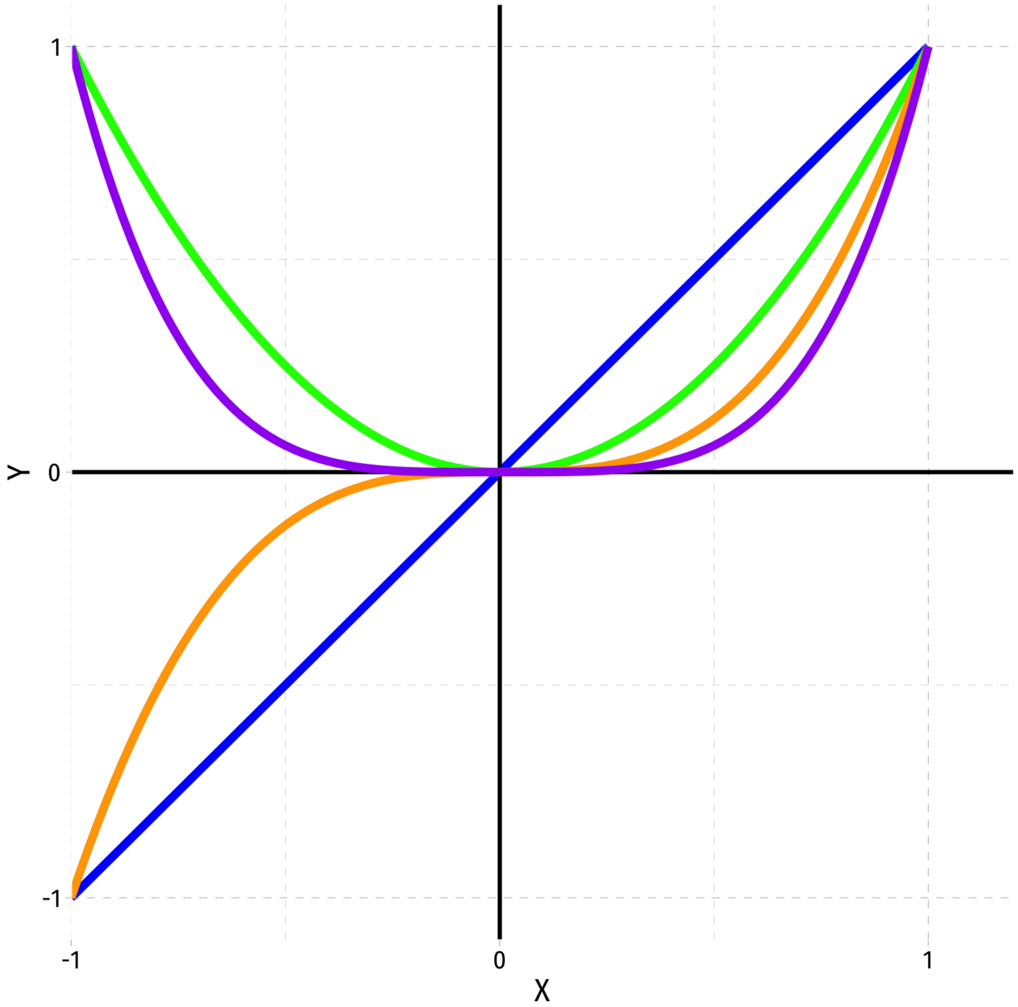
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

- 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 I

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

- 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!

The 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??[†]
- **Estimate marginal effects** by calculating predicted $\hat{Y}_i$ for different levels of X_i

[†] Note: this is *not* a perfect multicollinearity problem! Correlation only measures *linear* relationships!

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 $\left(\hat{\beta_1} + 2\hat{\beta_2} X_i\right)$ unit change in Y

Quadratic Model: Example I

Example:

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

- Use `gapminder` package and data

```r
library(gapminder)
```

Quadratic Model: Example II

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

```
gapminder <- gapminder %>%
  mutate(GDP_t = gdpPercap/1000)

gapminder %>% head() # look at it
```

country <fct>	continent <fct>	year <int>	lifeExp <dbl>	pop <int>	gdpPercap <dbl>	GDP_t <dbl>
Afghanistan	Asia	1952	28.801	8425333	779.4453	0.7794453
Afghanistan	Asia	1957	30.332	9240934	820.8530	0.8208530
Afghanistan	Asia	1962	31.997	10267083	853.1007	0.8531007
Afghanistan	Asia	1967	34.020	11537966	836.1971	0.8361971
Afghanistan	Asia	1972	36.088	13079460	739.9811	0.7399811
Afghanistan	Asia	1977	38.438	14880372	786.1134	0.7861134

6 rows

Quadratic Model: Example III

- Let's also create a squared term, `gdp_sq`

```
gapminder <- gapminder %>%
  mutate(GDP_sq = GDP_t^2)

gapminder %>% head() # look at it
```

country	continent	year	lifeExp	pop	gdpPercap	GDP_t	GDP_sq
<fct>	<fct>	<int>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
Afghanistan	Asia	1952	28.801	8425333	779.4453	0.7794453	0.6075350
Afghanistan	Asia	1957	30.332	9240934	820.8530	0.8208530	0.6737997
Afghanistan	Asia	1962	31.997	10267083	853.1007	0.8531007	0.7277808
Afghanistan	Asia	1967	34.020	11537966	836.1971	0.8361971	0.6992257
Afghanistan	Asia	1972	36.088	13079460	739.9811	0.7399811	0.5475720
Afghanistan	Asia	1977	38.438	14880372	786.1134	0.7861134	0.6179742

6 rows

Quadratic Model: Example IV

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

```
library(broom)
reg1 <- lm(lifeExp ~ GDP_t + GDP_sq, data = gapminder)

reg1 %>% tidy()
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	50.52400578	0.2978134673	169.64984	0.000000e+00
GDP_t	1.55099112	0.0373734945	41.49976	1.292863e-260
GDP_sq	-0.01501927	0.0005794139	-25.92149	3.935809e-125

3 rows

Quadratic Model: Example V

- OR use `gdp_t` and add the “transform” command in regression, `I(gdp_t^2)`

```
reg1_alt <- lm(lifeExp ~ GDP_t + I(GDP_t^2), data = gapminder)

reg1_alt %>% tidy()
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	50.52400578	0.2978134673	169.64984	0.000000e+00
GDP_t	1.55099112	0.0373734945	41.49976	1.292863e-260
I(GDP_t^2)	-0.01501927	0.0005794139	-25.92149	3.935809e-125

3 rows

Quadratic Model: Example VI

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	50.52400578	0.2978134673	169.64984	0.000000e+00
GDP_t	1.55099112	0.0373734945	41.49976	1.292863e-260
GDP_sq	-0.01501927	0.0005794139	-25.92149	3.935809e-125

3 rows

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

- 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 VII

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	50.52400578	0.2978134673	169.64984	0.000000e+00
GDP_t	1.55099112	0.0373734945	41.49976	1.292863e-260
GDP_sq	-0.01501927	0.0005794139	-25.92149	3.935809e-125

3 rows

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

$$\begin{aligned} \frac{\partial Y}{\partial X} &= \hat{\beta}_1 + 2\hat{\beta}_2 X_i \\ \frac{\partial \text{Life Expectancy}}{\partial \text{GDP}} &\approx 1.55 + 2(-0.02)\,\text{GDP} \\ &\approx 1.55 - 0.04\,\text{GDP} \end{aligned}$$

Quadratic Model: Example VIII

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

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 IX

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

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

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

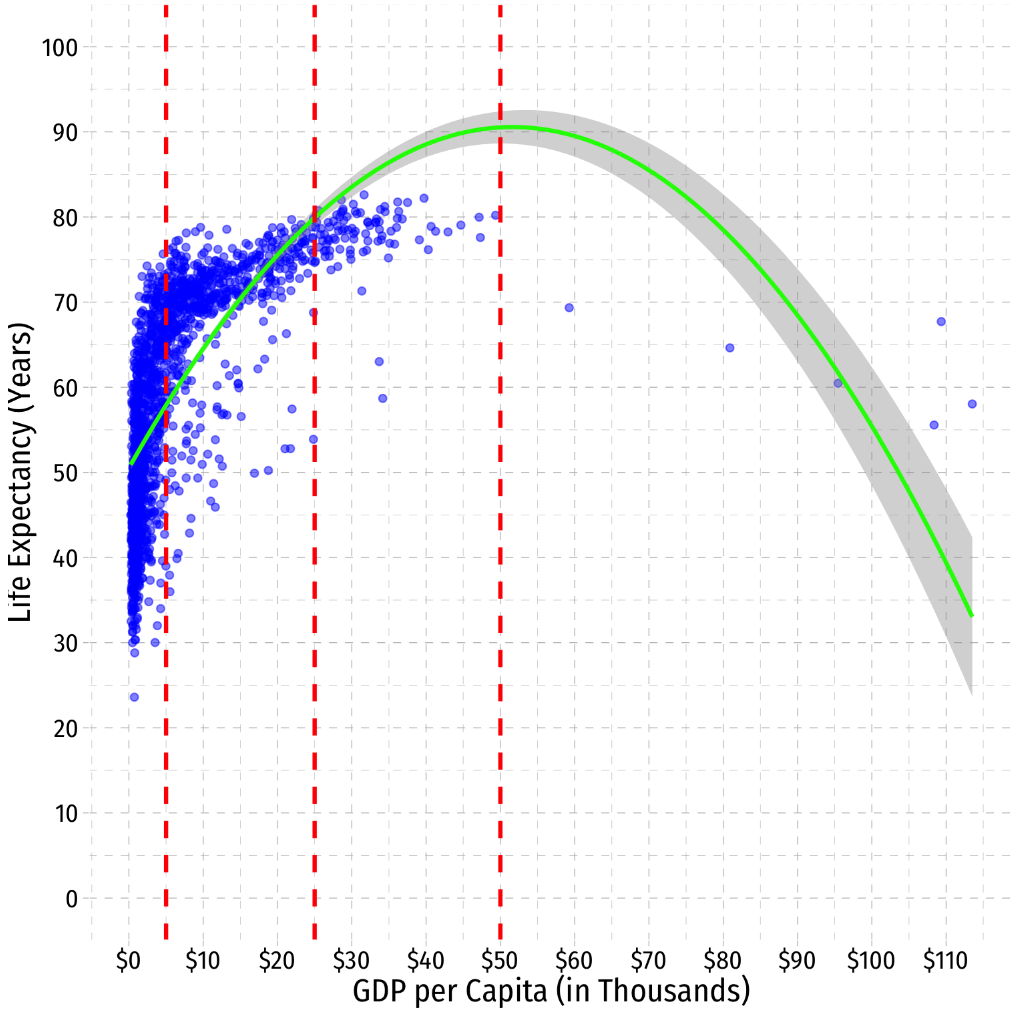
$$\frac{\partial\,\text{Life Expectancy}}{d\,\text{GDP}} = 1.55 - 0.04\text{GDP}$$

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

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

Quadratic Model: Example XII

```r
ggplot(data = gapminder)+
  aes(x = GDP_t,
      y = lifeExp)+
  geom_point(color="blue", alpha=0.5)+
  stat_smooth(method = "lm",
              formula = y ~ x + I(x^2),
              color="green")+
  geom_vline(xintercept=c(5,25,50),
             linetype="dashed",
             color="red", size = 1)+
  scale_x_continuous(labels=scales::dollar,
                     breaks=seq(0,120,10))+
  scale_y_continuous(breaks=seq(0,100,10),
                     limits=c(0,100))+
  labs(x = "GDP per Capita (in Thousands)",
       y = "Life Expectancy (Years)")+
  ggthemes::theme_pander(base_family = "Fira Sans Condensed",
           base_size=16)
```



The Quadratic Model: Maxima and Minima

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:

$$\begin{aligned}
\frac{\partial Y_i}{\partial X_i} &= 0 \\
\beta_1 + 2\beta_2 X_i &= 0 \\
2\beta_2 X_i &= -\beta_1 \\
X_i^* &= -\frac{\beta_1}{2\beta_2}
\end{aligned}$$

Quadratic Model: Maxima and Minima II

term	estimate	std.error	statistic
<chr>	<dbl>	<dbl>	<dbl>
(Intercept)	50.52400578	0.2978134673	169.64984
GDP_t	1.55099112	0.0373734945	41.49976
GDP_sq	-0.01501927	0.0005794139	-25.92149

3 rows | 1-4 of 5 columns

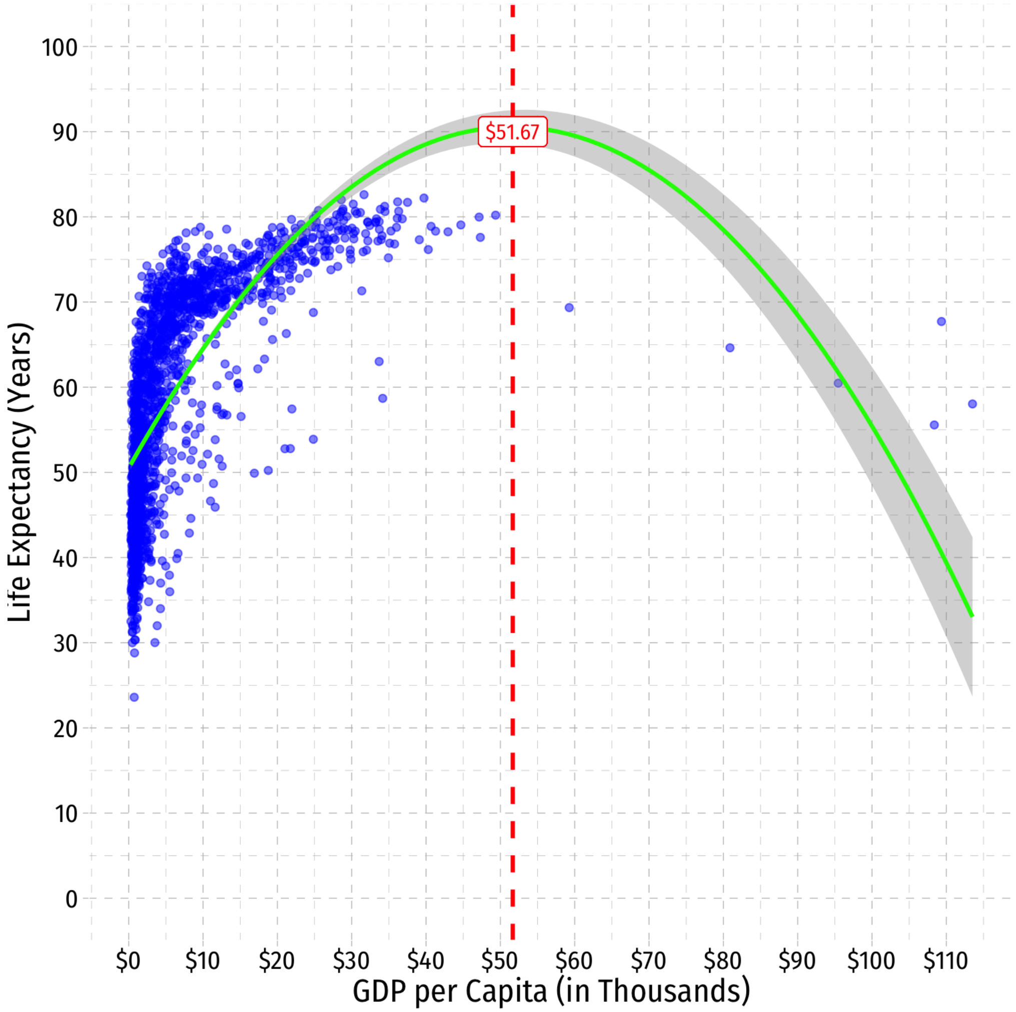
$$GDP_i^* = -\frac{\beta_1}{2\beta_2}$$

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

$$GDP_i^* \approx 51.67$$

Quadratic Model: Maxima and Minima III

```r
ggplot(data = gapminder)+
  aes(x = GDP_t,
      y = lifeExp)+
  geom_point(color="blue", alpha=0.5)+
  stat_smooth(method = "lm",
              formula = y ~ x + I(x^2),
              color="green")+
  geom_vline(xintercept=51.67, linetype="dashed", color="red"
  geom_label(x=51.67, y=90, label="$51.67", color="red")+
  scale_x_continuous(labels=scales::dollar,
                     breaks=seq(0,120,10))+
  scale_y_continuous(breaks=seq(0,100,10),
                     limits=c(0,100))+
  labs(x = "GDP per Capita (in Thousands)",
       y = "Life Expectancy (Years)")+
  ggthemes::theme_pander(base_family = "Fira Sans Condensed",
           base_size=16)
```



Are Polynomials Necessary?

Determining Polynomials are Necessary I

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	50.52400578	0.2978134673	169.64984	0.000000e+00
GDP_t	1.55099112	0.0373734945	41.49976	1.292863e-260
GDP_sq	-0.01501927	0.0005794139	-25.92149	3.935809e-125

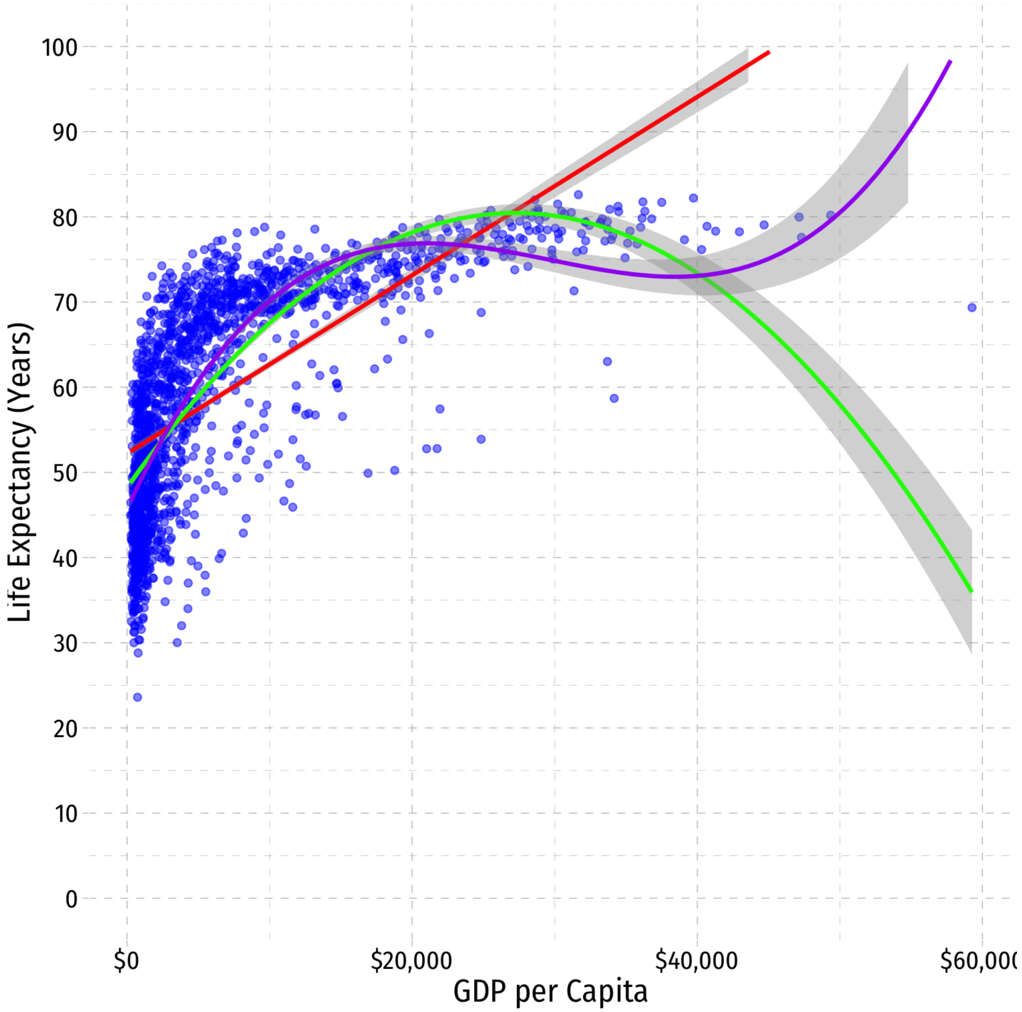
3 rows

- Is the quadratic term necessary?
- Determine if $\hat{\beta}_2$ (on X_i^2) is statistically significant:
  - $H_0 : \hat{\beta}_2 = 0$
  - $H_a : \hat{\beta}_2 \neq 0$
- Statistically significant $\implies$ we should keep the quadratic model
  - If we only ran a linear model, it would be incorrect!

Determining Polynomials are Necessary II

- Should we keep going up in polynomials?

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

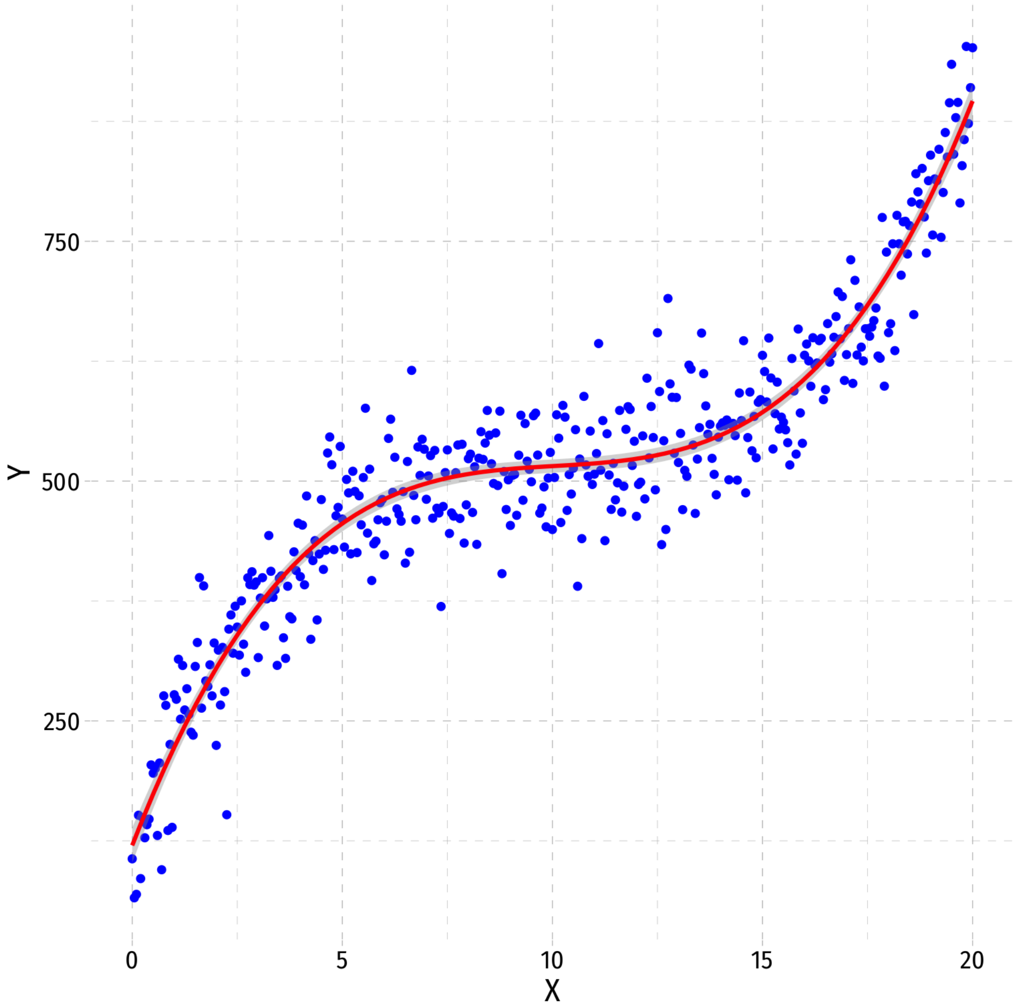


Determining Polynomials are Necessary III

- In general, you should have a **compelling theoretical reason** why data or relationships should **"change direction"** multiple times

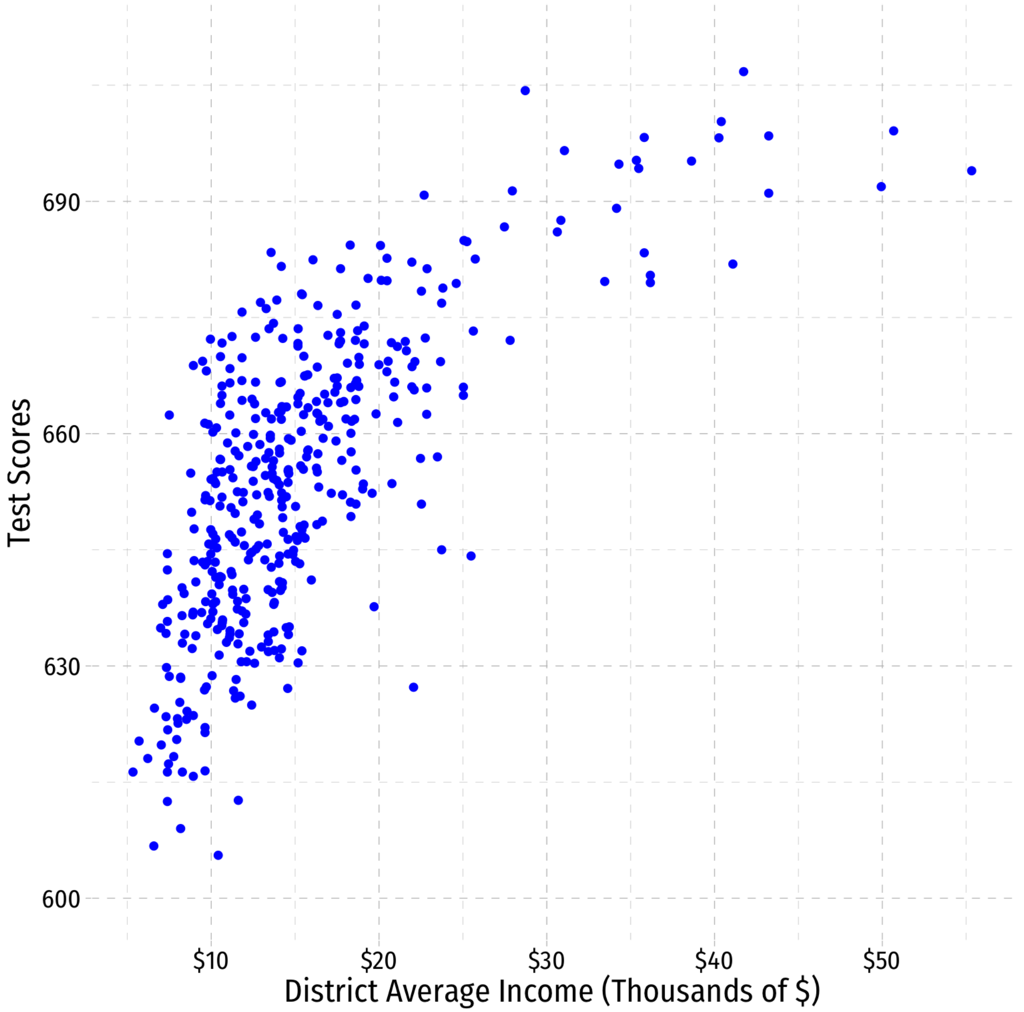
- Or clear data patterns that have multiple "bends"

- Recall, we care more about accurately measuring the causal effect between X and Y, rather than getting the most accurate prediction possible for $\hat{Y}$



A Second Polynomial Example I

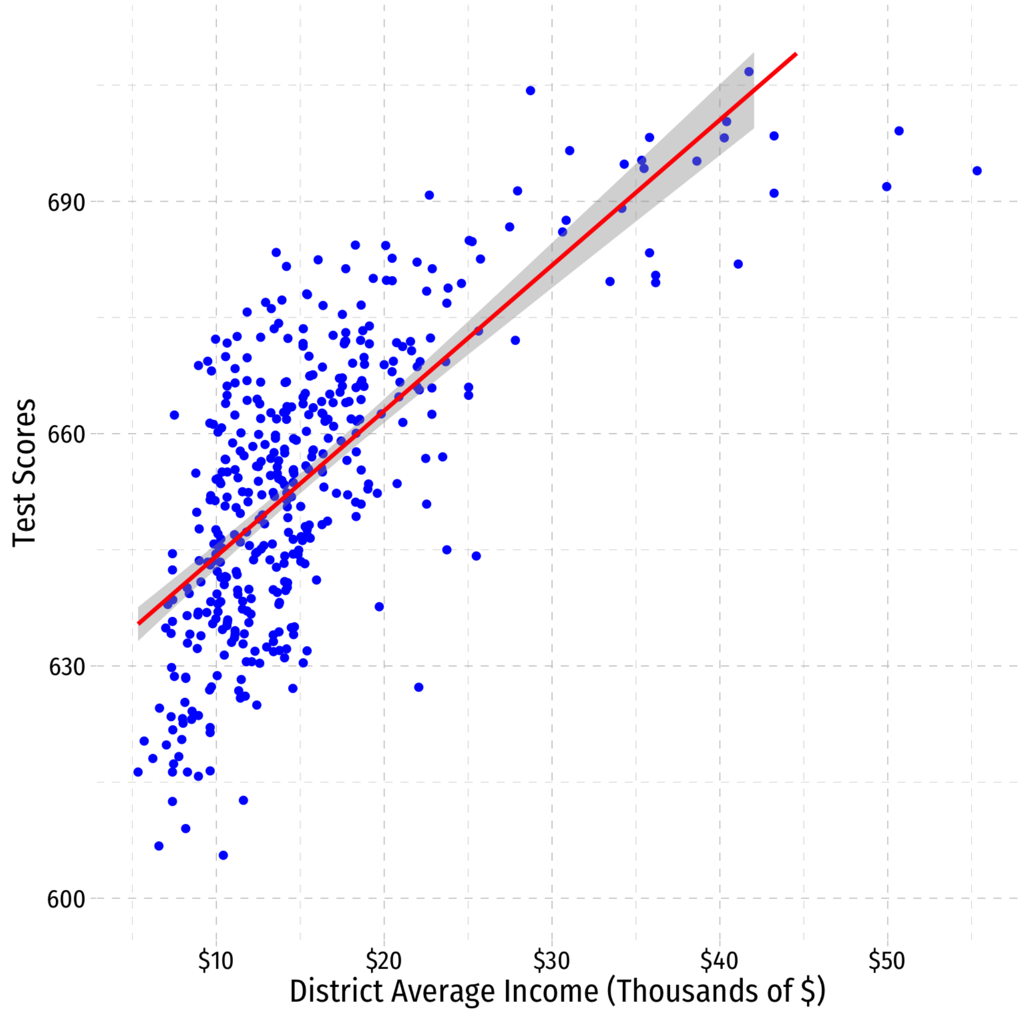
Example: How does a school district's average income affect Test scores?



A Second Polynomial Example I

Example: How does a school district's average income affect Test scores?

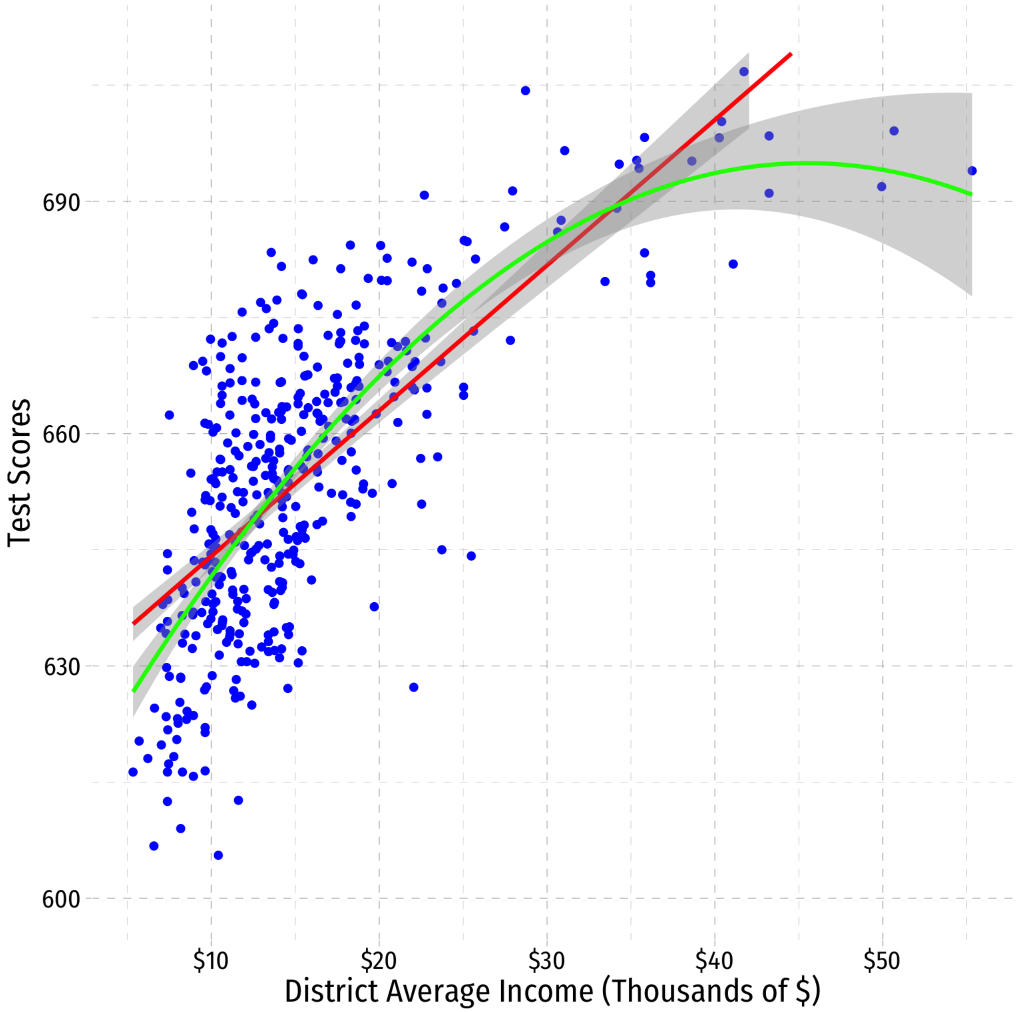
$$\widehat{\text{Test Score}_i} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}_i$$



A Second Polynomial Example I

Example: How does a school district's average income affect Test scores?

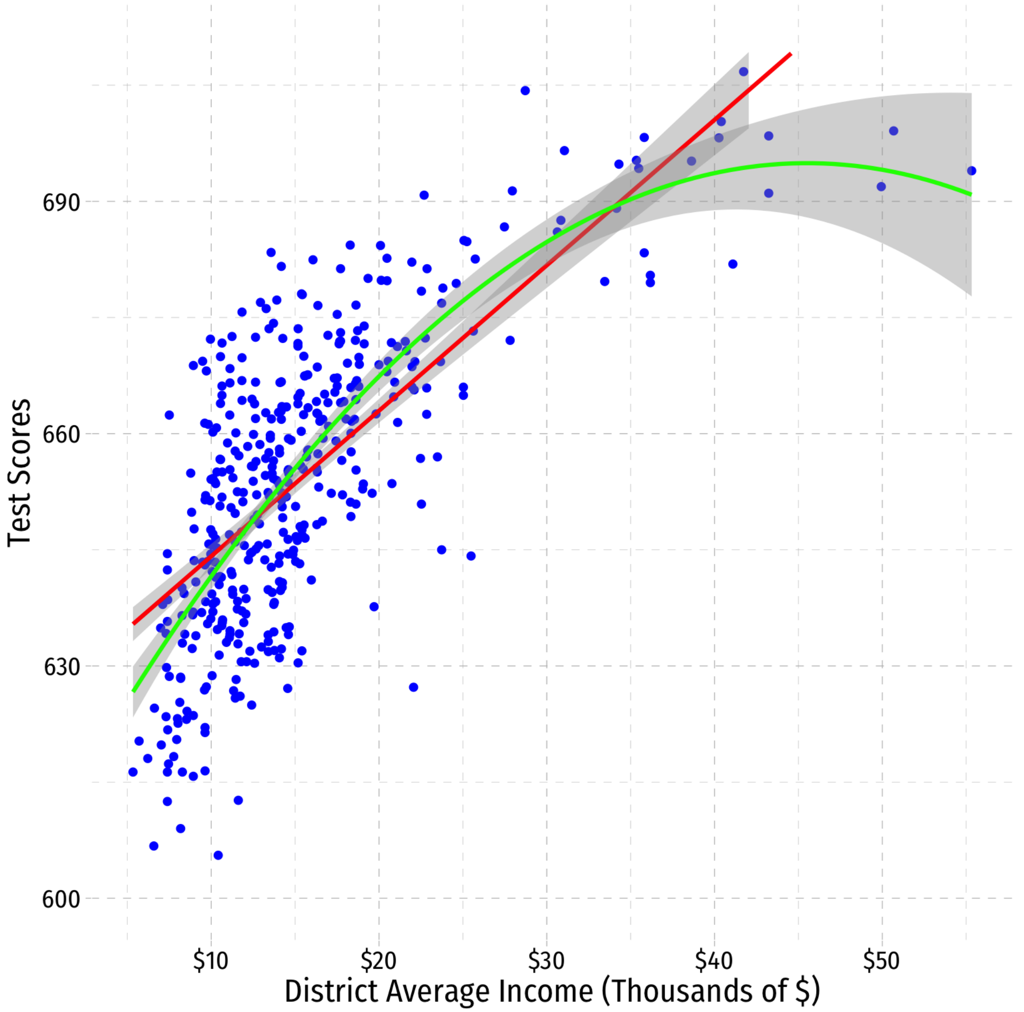
$$\widehat{\text{Test Score}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}_i + \hat{\beta}_1 \text{Income}_i^2$$



A Second Polynomial Example II

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	607.30173501	3.046219282	199.362449	0.000000e+00
avginc	3.85099474	0.304261693	12.656850	2.690099e-31
I(avginc^2)	-0.04230846	0.006260061	-6.758474	4.713383e-11

3 rows

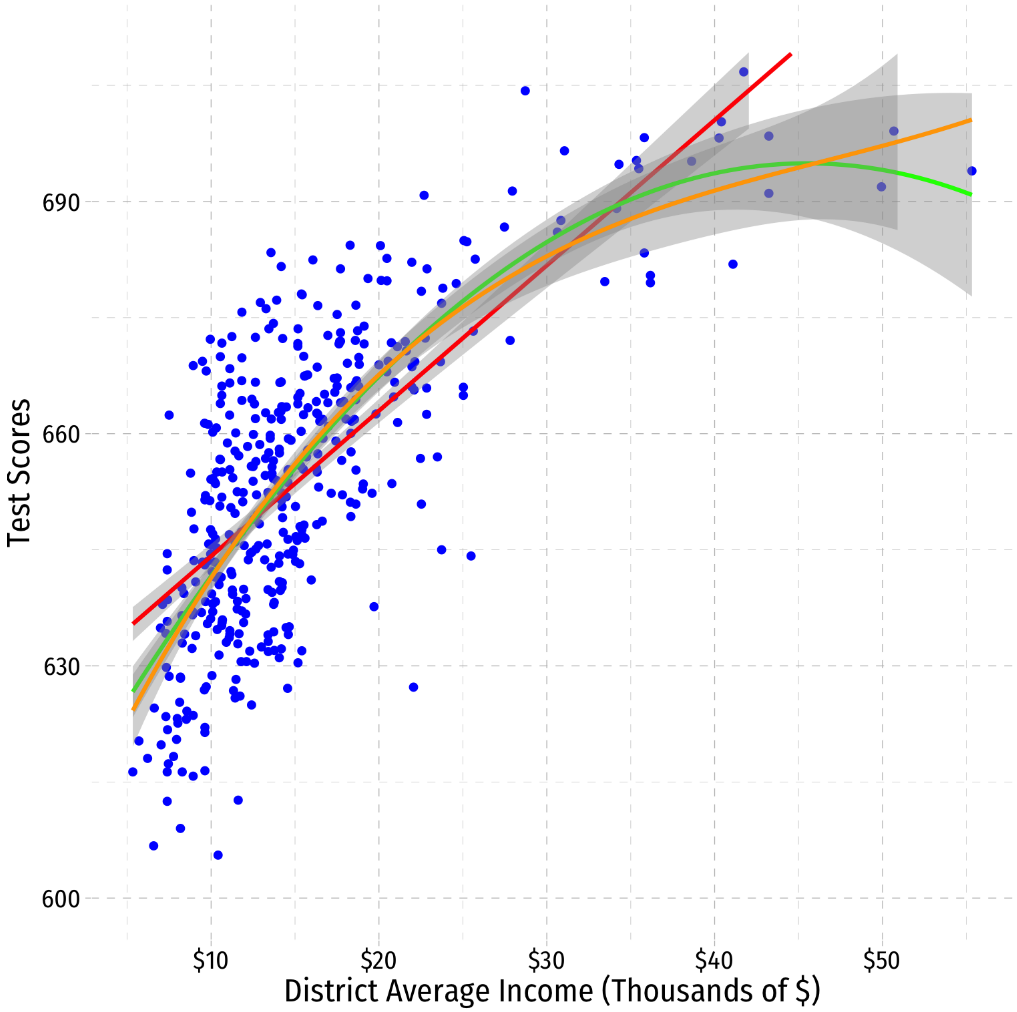


A Second Polynomial Example III

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	6.000790e+02	5.8295880342	102.936774	4.611745e-298
avginc	5.018677e+00	0.8594537744	5.839379	1.056874e-08
I(avginc^2)	-9.580515e-02	0.0373591998	-2.564433	1.068452e-02
I(avginc^3)	6.854842e-04	0.0004719549	1.452436	1.471343e-01

4 rows

- Should we keep going?



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