Polynomial Models
- An interaction model includes a new explanatory variable that is the product of two original explanatory variables.
- A polynomial model includes new explanatory variables that are powers of original explanatory variables.

Example
- Response: Population of the U.S. (millions)
- Explanatory: Year the census was taken.

Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Year</th>
<th>Population</th>
<th>Year</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1790</td>
<td>3.929</td>
<td>1870</td>
<td>38.558</td>
<td>1950</td>
<td>151.326</td>
</tr>
<tr>
<td>1800</td>
<td>5.308</td>
<td>1880</td>
<td>50.189</td>
<td>1960</td>
<td>179.323</td>
</tr>
<tr>
<td>1810</td>
<td>7.240</td>
<td>1890</td>
<td>62.980</td>
<td>1970</td>
<td>203.302</td>
</tr>
<tr>
<td>1820</td>
<td>9.638</td>
<td>1900</td>
<td>76.212</td>
<td>1980</td>
<td>226.542</td>
</tr>
<tr>
<td>1830</td>
<td>12.861</td>
<td>1910</td>
<td>92.228</td>
<td>1990</td>
<td>248.710</td>
</tr>
<tr>
<td>1840</td>
<td>17.063</td>
<td>1920</td>
<td>106.022</td>
<td>2000</td>
<td>281.422</td>
</tr>
<tr>
<td>1850</td>
<td>23.192</td>
<td>1930</td>
<td>123.203</td>
<td>2010</td>
<td>308.746</td>
</tr>
<tr>
<td>1860</td>
<td>31.443</td>
<td>1940</td>
<td>132.165</td>
<td>2020</td>
<td>???</td>
</tr>
</tbody>
</table>
General Trend

- As the years pass, population tends to grow, but not at the same rate (non-linear).
- In the 1800’s the population grew slowly.
- In the 1900’s the population grew more quickly.

Simple Linear Model

- How well will a simple linear model relating population to year do at explaining the relationship between these two variables?
Simple Linear Model

- Predicted Population = -2480.85 + 1.360\times Year
- The estimated intercept is not interpretable because although Year = 0 makes sense, Year = 0 is way outside the values for Year in the data set.

Simple Linear Model

- Predicted Population = -2480.85 + 1.360\times Year
- The estimated slope can be interpreted as follows: for every additional year, the population increases 1.360 (million), on average.

Model Utility

- F=239.13, P-value<0.0001
- The small P-value indicates that there is a statistically significant linear relationship between population and year.
Statistical Significance

- Year
  - $t=15.46$, $P$-value $< 0.0001$
  - $F=239.13$, $P$-value $< 0.0001$

- The $P$-value is small, therefore there is a statistically significant linear relationship between population and year.

Simple Linear Model

- $R^2=0.919$ or 91.9% of the variation in population can be explained by the linear relationship with year.
- RMSE = 27.99

Summary - SLR

- The model is useful.
- The linear relationship with year is statistically significant.
- 91.9% of the variation in population is explained by the simple linear model.
Problems with SLR

- Year 2010
  - Predicted Population = \(-2480.85 + 1.360 \times 2010\) = 252.75 million
  - This predicted value is much smaller than the 308.746 million in 2010.

- Year 1800
  - Predicted Population = \(-2480.85 + 1.360 \times 1800\) = \(-32.85\) million
  - This predicted value is negative!
  - What does a negative predicted population mean?
Plot of Residuals

- There is a curved pattern to the plot of residuals versus year.
- The SLR under-predicts up to 1840, over-predicts from 1840 through 1960, and under-predicts from 1970 to 2010.

Prediction for 2020

- The pattern in the residuals suggests that the prediction for 2020 (266.35 million) is under what the true population in that year will be.
Prediction for 1800
- The pattern in the residuals suggests that the prediction for 1800 (-32.85 million) is under what the true population in that year was.

Plot of Residuals
- Although the simple linear regression model is useful and explains a lot of the variation in population, we can do better with a model that accounts for the curvature.

How can we do better?
- We need to add a variable to the simple linear regression model that can account for the curved nature of the relationship between population and year.