Quadratic Model

In order to account for curvature in the relationship between an explanatory and a response variable, one often adds the square of the explanatory variable to the simple linear model.

Quadratic Model

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \]

Conditions on \( \epsilon \)
- Independent
- Identically distributed
- Normally distributed with common standard deviation, \( \sigma \)

Example

- Response, \( Y \): Population of the U.S. (millions)
- Explanatory, \( X \): Year the census was taken.
Quadratic Model

- Predicted Population = 21952.25 – 24.390*Year + 0.0067764*Year^2
- We cannot interpret the estimated slope coefficients because we cannot change Year by 1 while holding Year^2 constant.

Model Utility

- F=11091.30, P-value<0.0001
- The small P-value indicates that the quadratic model relating population to Year and Year^2 is statistically significant (useful).

Statistical Significance

- Year (added to Year^2)
  - t=-39.87, P-value<0.0001
  - The P-value is small, therefore the addition of Year is statistically significant.
Statistical Significance

- Year^2 (added to Year)
  - t=42.10, P-value<0.0001
- The P-value is small, therefore the addition of Year^2 is statistically significant.

Quadratic Model

- R^2=0.999 or 99.9% of the variation in population can be explained by the quadratic model.
- RMSE=3.03 million

Summary - Quadratic

- The model is useful.
- Each term is a statistically significant addition.
- 99.9% of the variation in population is explained by the quadratic model.
Prediction

- **Year 2010**
  - Not bad as the actual figure in 2010 was 308.746 million.

Prediction for 2020

- **Year 2020**
  - Predicted Population = 21952.25 - 24.390(2020) + 0.0067764*(2020)^2 = 334.873 million

Prediction

- **Year 1800**
  - Predicted Population = 21952.25 - 24.390(1800) + 0.0067764*(1800)^2 = 5.786 million
  - Very close to the actual value of 5.308 million.
The residuals wiggle around the zero line. Hard to say whether this is a pattern or not.

The residuals for 1940 and 1950 stick out. The quadratic model over predicts for these years.
Possible explanation.

- 1940 census followed the Great Depression when immigration was very low.
- 1950 census followed WW-II when birth rates and immigration were low.

Can we do better?

- Could try higher order polynomial terms like Year³ or Year⁴.
- Year³ is statistically significant (P-value = 0.0335) in cubic model.
- Year⁴ is not statistically significant in a quartic model.

Quadratic Model

- There is still the issue of trying to interpret the coefficients in the quadratic model.
- Again, creating a new explanatory variable, Year², has introduced multicollinearity into the quadratic model.
Correlation

- Year and Year²
  - Correlation: $r = 0.9999$
  - For the values that Year takes on, there is an extremely strong positive linear correlation with Year².

Centering

- Center Year by subtracting off the mean before constructing the squared term in the quadratic model.
- Mean year is 1900.
Quadratic Model

- Predicted Population = $-2510.67 + 1.360 \times \text{Year} + 0.000067764 \times (\text{Year} - 1900)^2$
- Note that the estimated slope for year is exactly the same as in the simple linear model.

Correlation

- Year and $(\text{Year} - 1900)^2$
- Correlation: $r = -0.0000$
- For the values that Year takes on, there is no linear correlation with $(\text{Year} - 1900)^2$. 
Centering

- Centering has completely removed the multicollinearity resulting from the inclusion of the quadratic term in the quadratic model.

Quadratic Model

- Predicted Population = 74.166 + 1.360*(Year – 1900) + 0.0067764*(Year – 1900)^2
- The predicted population in 1900 is 74.166 million.

Quadratic Model

- Predicted Population = 74.166 + 1.360*(Year – 1900) + 0.0067764*(Year – 1900)^2
- For each additional year, the population goes up, on average, 1.360 million.
Quadratic Model

- Predicted Population = 74.166 + 1.360*(Year – 1900) + 0.0067764*(Year – 1900)^2
- In addition to the average change per year, there is a bigger adjustment to this rate of change the further away you are from 1900.

Ten Year Change

- 1880: Pred = 49.668 million
- 1890: Pred = 61.240 million
  - Difference of 11.572 million
  - Average increase of 1.1572 million per year.

Ten Year Change

- 1980: Pred = 226.371 million
- 1990: Pred = 251.495 million
  - Difference of 25.124 million
  - Average increase of 2.5124 million per year.