

Stat 401 B – Lecture 19

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.

1

Example

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

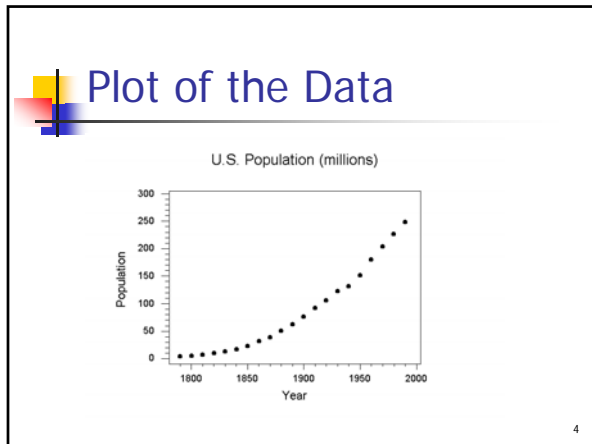
2

Data

Year	Population	Year	Population	Year	Population
1790	3.929	1860	31.443	1930	123.203
1800	5.308	1870	38.558	1940	132.165
1810	7.240	1880	50.189	1950	151.326
1820	9.638	1890	62.980	1960	179.323
1830	12.861	1900	76.212	1970	203.302
1840	17.063	1910	92.228	1980	226.542
1850	23.192	1920	106.022	1990	248.710

3

Stat 401 B – Lecture 19



- ## 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.
- 5

- ## Simple Linear Model
- How well will a simple linear model relating population to year do at explaining the relationship between these two variables?
- 6

Stat 401 B – Lecture 19

Simple Linear Model

- Predicted Population = $-2211.3 + 1.215 \cdot \text{Year}$
- 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.

7

Simple Linear Model

- Predicted Population = $-2211.3 + 1.215 \cdot \text{Year}$
- The estimated slope can be interpreted as follows: for every additional year, the population increases 1.215 (million), on average.


8

Model Utility

- $F=224.33$, $P\text{-value} < 0.0001$
- The small P-value indicates that there is a statistically significant linear relationship between population and year.

9


Stat 401 B – Lecture 19



Statistical Significance

- Year
 - $t=14.98$, $P\text{-value}<0.0001$
 - $F=224.33$, $P\text{-value}<0.0001$
- The P-value is small, therefore there is a statistically significant linear relationship between population and year.


10



Simple Linear Model

- $R^2=0.922$ or 92.2% of the variation in population can be explained by the linear relationship with year.
- $RMSE=22.52$

11



Summary - SLR

- The model is useful.
- The linear relationship with year is statistically significant.
- 92.2% of the variation in population is explained by the simple linear model.

12

Stat 401 B – Lecture 19

Problems with SLR

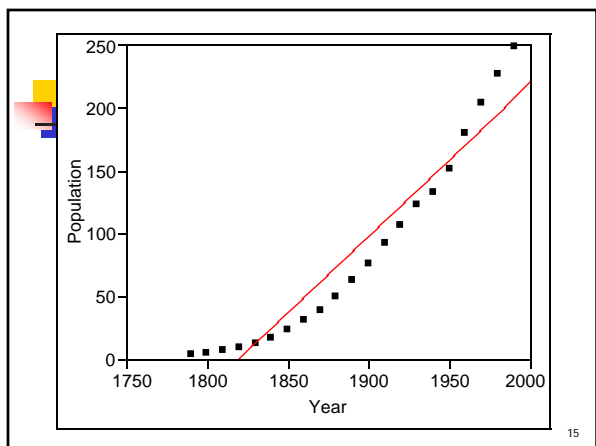
- Year 2000
 - Predicted Population = $-2211.3 + 1.215 \cdot (2000) = 218.7$ million
- This predicted value is smaller than the actual population in 1980 or 1990.

13

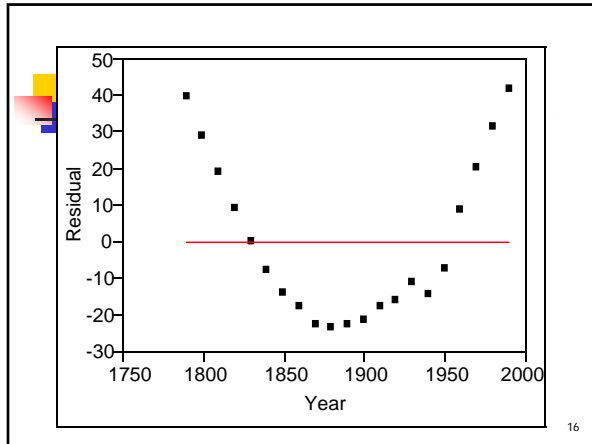
Problems with SLR

- Year 1800
 - Predicted Population = $-2211.3 + 1.215 \cdot (1800) = -24.3$ million
- This predicted value is negative!
- What does a negative predicted population mean?

14



Stat 401 B – Lecture 19



Plot of Residuals

- There is a curved pattern to the plot of residuals versus year.
- The SLR under-predicts up to 1830, over-predicts from 1840 through 1950, and under-predicts from 1960 to 1990.


17

Prediction for 2000

- The pattern in the residuals suggests that the prediction for 2000 (218.7 million) is under what the true population in that year was.

18


Stat 401 B – Lecture 19



Prediction for 1800

- The pattern in the residuals suggests that the prediction for 2000 (–24.3 million) is under what the true population in that year was.


19



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.

20



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.

21
