Linear Regression I

Nate Wells

Math 141, 2/25/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Introduce statistical modeling
- Investigate the linear model
- Discuss predictions and residuals

Section 1

Introduction to Linear Regression

Overview

"All models are wrong, but some are useful."

— George Box, statistician

Overview

"All models are wrong, but some are useful."

— George Box, statistician

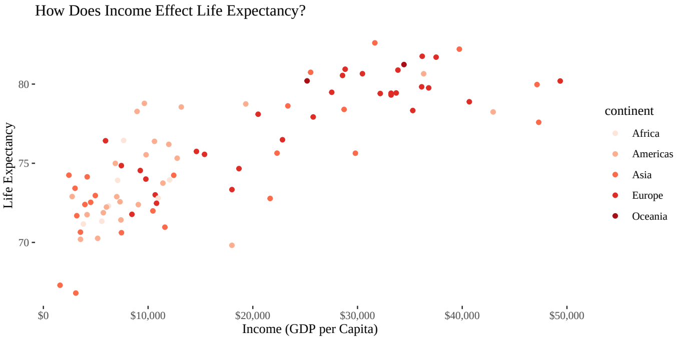
- Linear regression is both an accessible and potent tool in statistical analysis.

Overview

"All models are wrong, but some are useful."

— George Box, statistician

- Linear regression is both an accessible and potent tool in statistical analysis.

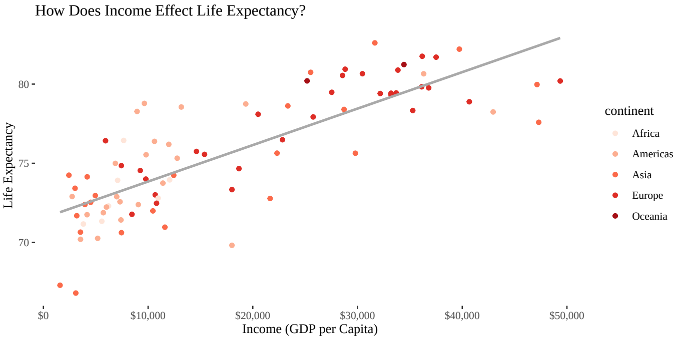


Overview

“All models are wrong, but some are useful.”

— George Box, statistician

- Linear regression is both an accessible and potent tool in statistical analysis.



Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

$$Y = \beta_0 + \beta_1 X \qquad \text{with } \beta_0, \beta_1 \text{ fixed constants}$$

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

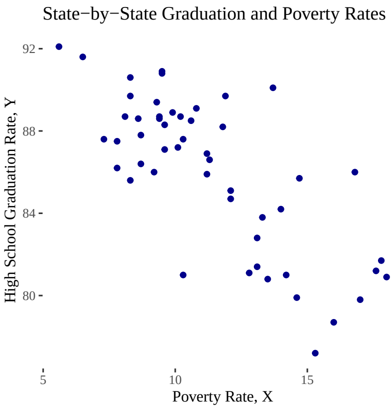
- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

$$Y = \beta_0 + \beta_1 X \qquad \text{with } \beta_0, \beta_1 \text{ fixed constants}$$

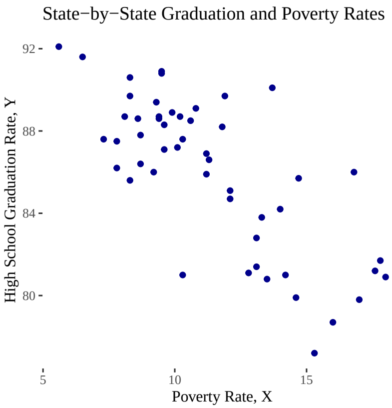
- Of course, in the wild, the observed values of Y will **not** be perfectly predicted by the values of X.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \text{where } \epsilon \text{ is the error}$$

Scatterplots and Linear Relationships I

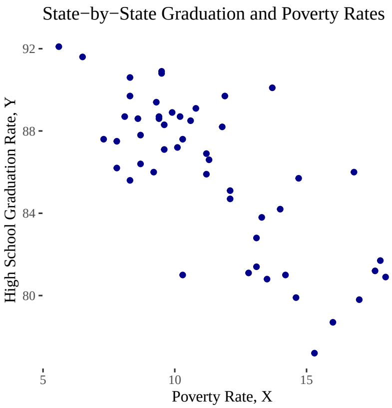


Scatterplots and Linear Relationships I



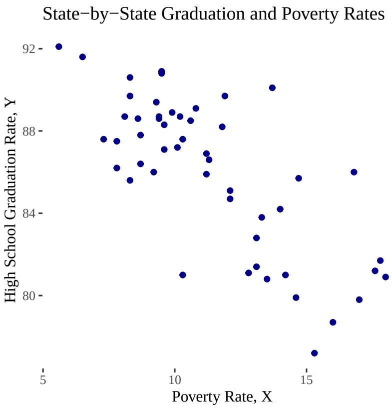
- Explanatory Variable:

Scatterplots and Linear Relationships I



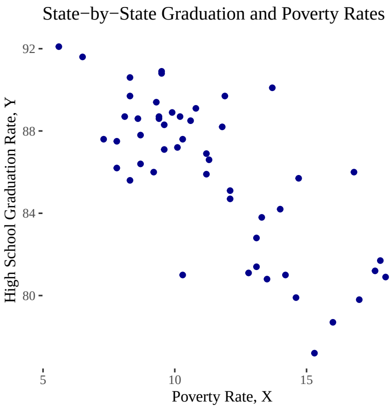
- Explanatory Variable:
  - Poverty Rate (X)

Scatterplots and Linear Relationships I



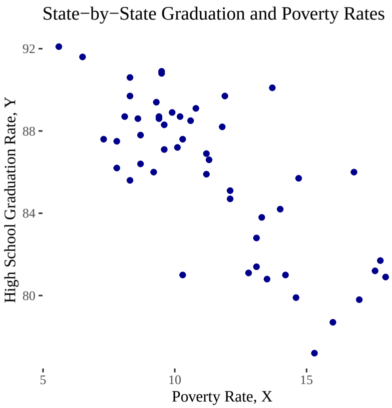
- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:

Scatterplots and Linear Relationships I



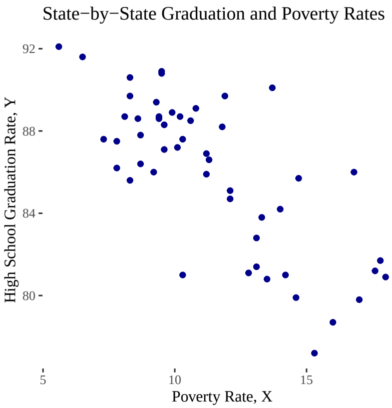
- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)

Scatterplots and Linear Relationships I



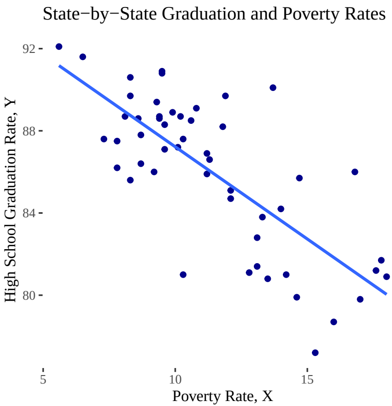
- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)
- Relationship:

Scatterplots and Linear Relationships I



- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)
- Relationship:
  - Linear, negative, moderately strong

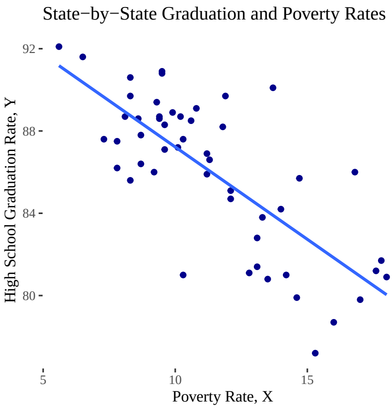
Scatterplots and Linear Relationships II



- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 96.2 - 0.9X$$

Scatterplots and Linear Relationships II

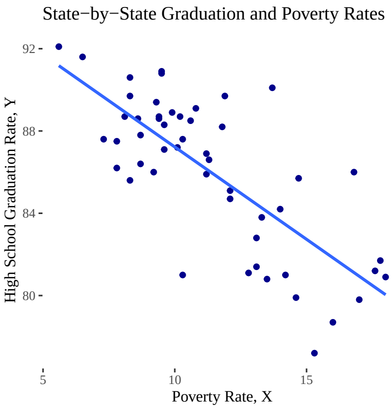


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 96.2 - 0.9X$$

- Hat ($\hat{Y}$) indicates this is an estimate of Y

Scatterplots and Linear Relationships II

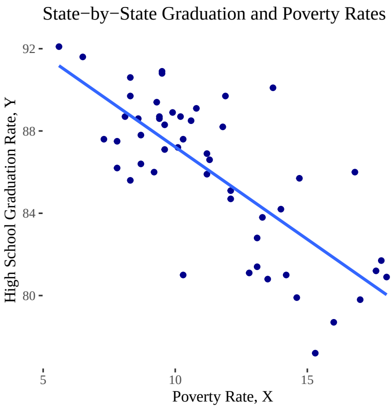


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 96.2 - 0.9X$$

- Hat ($\hat{Y}$) indicates this is an estimate of Y
- **Slope** of $\beta_1 = -0.9$ means every 1 unit increase in Poverty corresponds to a 0.9 unit decrease on average in Graduation.

Scatterplots and Linear Relationships II

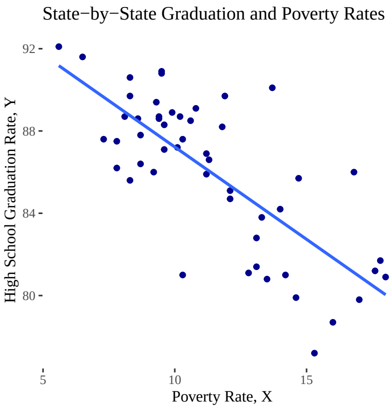


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 96.2 - 0.9X$$

- Hat ($\hat{Y}$) indicates this is an estimate of Y
- **Slope** of $\beta_1 = -0.9$ means every 1 unit increase in Poverty corresponds to a 0.9 unit decrease on average in Graduation.
- **Intercept** of $\beta_0 = 96.2$ means model predicts graduation rate of 96.2% when poverty rate is 0%.

Scatterplots and Linear Relationships III

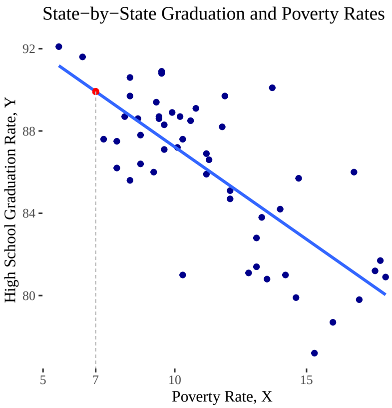


- Model:

$$\hat{Y} = 96.2 - 0.9 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 7%?

Scatterplots and Linear Relationships III

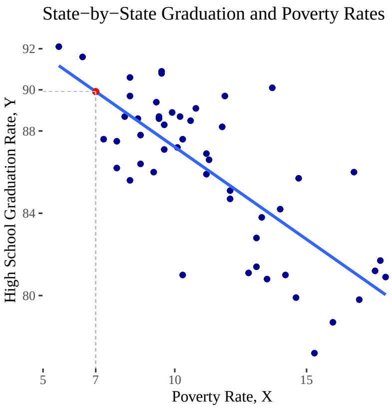


- Model:

$$\hat{Y} = 96.2 - 0.9 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 7%?

Scatterplots and Linear Relationships III



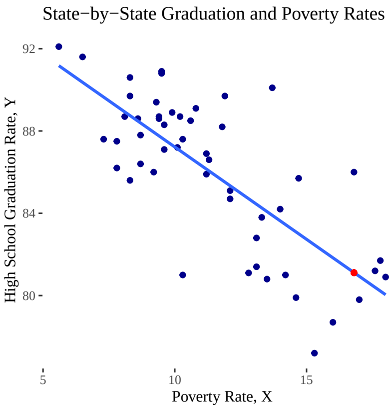
- Model:

$$\hat{Y} = 96.2 - 0.9 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 7%?

$$\hat{Y} = 96.2 - 0.9 \cdot 7 = 89.9$$

Scatterplots and Linear Relationships IV



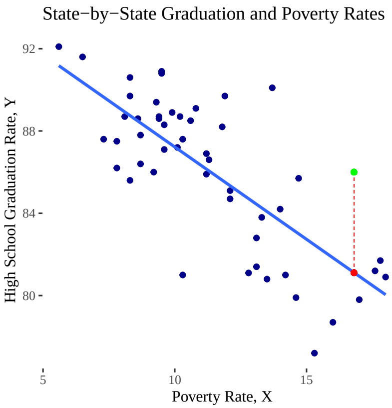
- Model:

$$\hat{Y} = 96.2 - 0.9 \cdot X$$

- Washington D.C. has a poverty rate of 16.8. What does the model predict is D.C.'s graduation rate?

$$\hat{Y} = 96.2 - 0.9 \cdot 16.8 = 81.1$$

Scatterplots and Linear Relationships IV



- Model:

$$\hat{Y} = 96.2 - 0.9 \cdot X$$

- Washington D.C. has a poverty rate of 16.8. What does the model predict is D.C.'s graduation rate?

$$\hat{Y} = 96.2 - 0.9 \cdot 16.8 = 81.1$$

But D.C.'s actual graduation rate is 86.0

Residuals

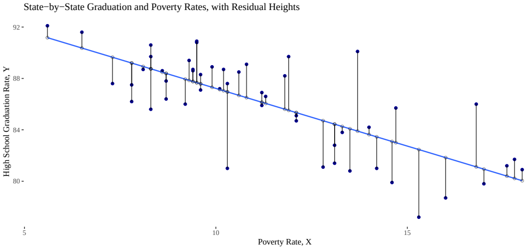
- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

$$e_i = Y_i - \hat{Y}_i$$

Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

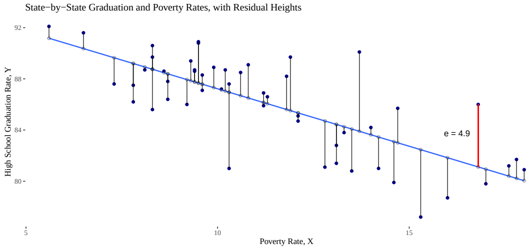
$$e_i = Y_i - \hat{Y}_i$$



Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

$$e_i = Y_i - \hat{Y}_i$$

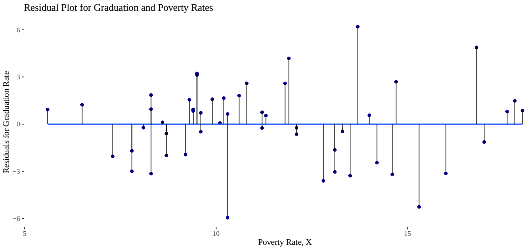


- D.C.'s residual is

$$e = Y - \hat{Y} = 86 - 81.1 = 4.9$$

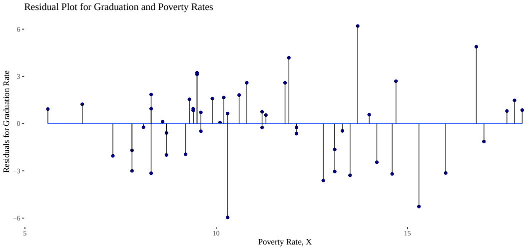
Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



Residual Plot

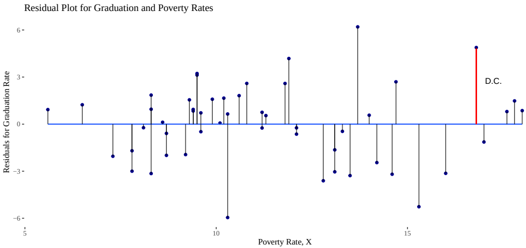
- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.

Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.