Linear Regression II

Nate Wells

Math 141, 2/25/21

Outline

In this lecture, we will. . .

Outline

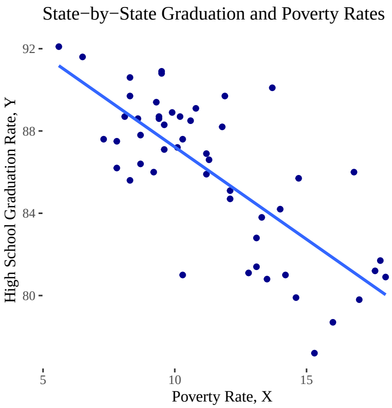
In this lecture, we will. . .

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

Section 1

Introduction to Linear Regression

Scatterplots and Linear Relationships



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

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

Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.

Residuals

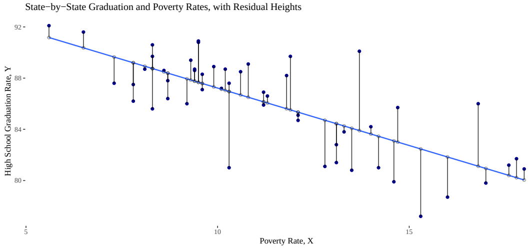
- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

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

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:

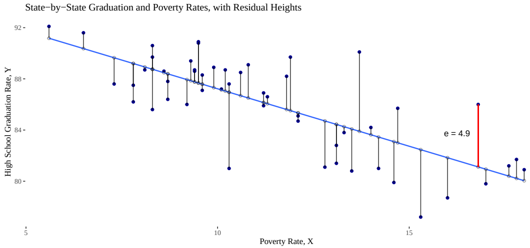
$$e_i = Y_i - \hat{Y}_i$$



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

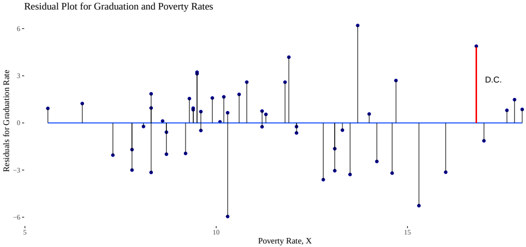
$$e = Y - \hat{Y} = 86 - 81.1 = 4.9$$

Residual Plot

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

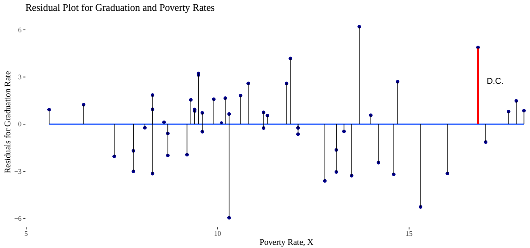
Residual Plot

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



Residual Plot

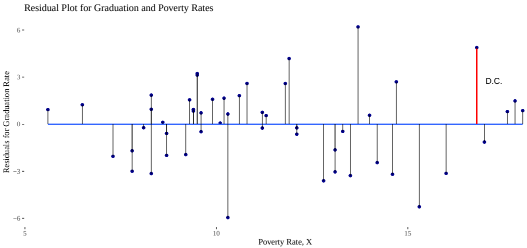
- To visualize the degree of accuracy of a linear model, we use residual plots:



- 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.
- Tighter clustering around the horizontal axis indicates stronger fit.

Quantifying Goodness-of-Fit

- **Correlation** R describes the strength of a linear relationship between two variables, and is always a number between -1 and 1.

Quantifying Goodness-of-Fit

- **Correlation** R describes the strength of a linear relationship between two variables, and is always a number between -1 and 1.

If R is close to ...	Then linear relationship is...
1	strong, positive
-1	strong, negative
0	weak

Quantifying Goodness-of-Fit

- **Correlation** R describes the strength of a linear relationship between two variables, and is always a number between -1 and 1.

If R is close to ...	Then linear relationship is...
1	strong, positive
-1	strong, negative
0	weak

- Correlation can be computed via formula using the mean and standard deviation of each variable.

Quantifying Goodness-of-Fit

- **Correlation** R describes the strength of a linear relationship between two variables, and is always a number between -1 and 1.

If R is close to ...	Then linear relationship is...
1	strong, positive
-1	strong, negative
0	weak

- Correlation can be computed via formula using the mean and standard deviation of each variable.

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Quantifying Goodness-of-Fit

- **Correlation** R describes the strength of a linear relationship between two variables, and is always a number between -1 and 1.

If R is close to ...	Then linear relationship is...
1	strong, positive
-1	strong, negative
0	weak

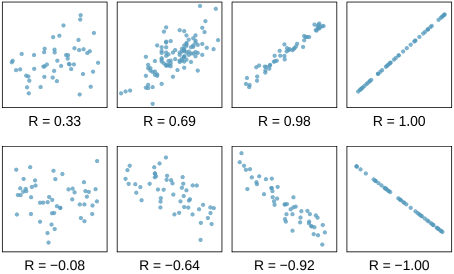
- Correlation can be computed via formula using the mean and standard deviation of each variable.

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  - But in practice, we will always use technology to compute R.

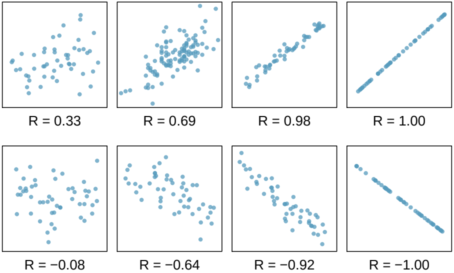
Correlation

- Correlation gives a **relative** sense of the strength of a linear relationship



Correlation

- Correlation gives a **relative** sense of the strength of a linear relationship



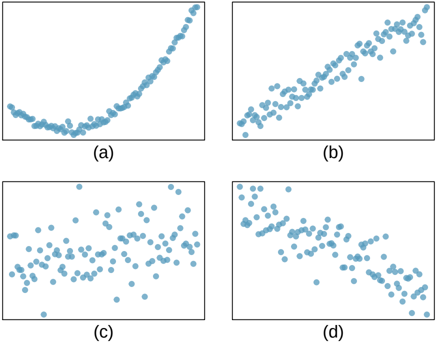
- In practice, correlation is...
  - strong, if $|R| > 0.7$
  - moderate, if $0.3 < |R| < 0.7$
  - weak, if $|R| < 0.3$

Correlation is not Association

- Correlation measures strength of **LINEAR** relationship:

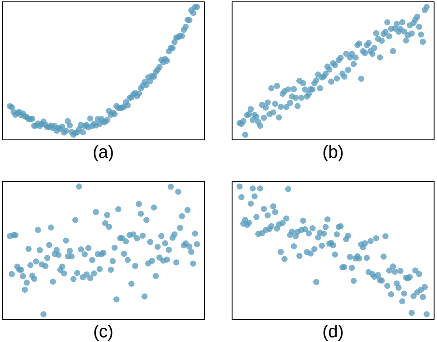
Correlation is not Association

- Correlation measures strength of **LINEAR** relationship:
- Which of the following has the strongest correlation (largest value of $|R|$)?



Correlation is not Association

- Correlation measures strength of **LINEAR** relationship:
- Which of the following has the strongest correlation (largest value of $|R|$)?



- Answer: (b), not (a)

Correlation isn't the Whole Story

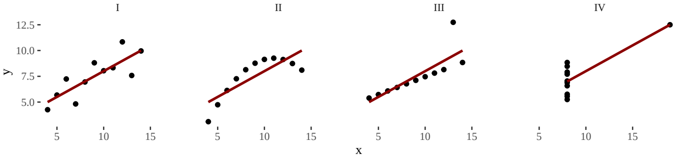
- Computing a correlation coefficient is no substitute for data visualization.

Correlation isn't the Whole Story

- Computing a correlation coefficient is no substitute for data visualization.
- All of the following have identical, strong positive correlation ($R = 0.8$):

Correlation isn't the Whole Story

- Computing a correlation coefficient is no substitute for data visualization.
- All of the following have identical, strong positive correlation ($R = 0.8$):



Section 2

Fitting a Line by Least-Squares Regression

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values
    $$|e_1| + |e_2| + \cdots + |e_n|$$
  - Option 2: Minimize the sum of squares
    $$e_1^2 + e_2^2 + \cdots + e_n^2$$
- Option 2 is usually preferred.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.
  3. Has theoretical advantages (by analogy with distance and pythagorean thm.)

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.
  3. Has theoretical advantages (by analogy with distance and pythagorean thm.)
  4. Appropriately weights one large residuals as “worse” than many small ones.

An Aside on Some Important Formulas

- You **do not** need to memorize these formulas

An Aside on Some Important Formulas

- You **do not** need to memorize these formulas
  - But you should understand where they come from and what they mean

An Aside on Some Important Formulas

- You **do not** need to memorize these formulas
  - But you should understand where they come from and what they mean
- Suppose $x_1, x_2, \dots, x_n$ are a list of numerica observations. . .
- The **mean** of this data set is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

An Aside on Some Important Formulas

- You **do not** need to memorize these formulas
  - But you should understand where they come from and what they mean
- Suppose $x_1, x_2, \ldots, x_n$ are a list of numerica observations...
- The **mean** of this data set is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

  - The mean is a measure of center.
- The standard deviation of this data is

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

An Aside on Some Important Formulas

- You **do not** need to memorize these formulas
  - But you should understand where they come from and what they mean
- Suppose $x_1, x_2, \ldots, x_n$ are a list of numerica observations...
- The **mean** of this data set is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

  - The mean is a measure of center.
- The standard deviation of this data is

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

  - The standard deviation is a measure of spread.

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

with means $\bar{x}, \bar{y}$ and standard deviations s_x, s_y and correlation R.

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

  with means $\bar{x}, \bar{y}$ and standard deviations s_x, s_y and correlation R.

- The **Least Squares Regression Line** modeling Y as a function of X is

$$\hat{Y} = \beta_0 + \beta_1 X$$

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

with means $\bar{x}, \bar{y}$ and standard deviations s_x, s_y and correlation R.

- The **Least Squares Regression Line** modeling Y as a function of X is

$$\hat{Y} = \beta_0 + \beta_1 X$$

where the slope β_1 is given by

$$\beta_1 = \frac{s_y}{s_x} R$$

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

with means $\bar{x}, \bar{y}$ and standard deviations s_x, s_y and correlation R.

- The **Least Squares Regression Line** modeling Y as a function of X is

$$\hat{Y} = \beta_0 + \beta_1 X$$

where the slope β_1 is given by

$$\beta_1 = \frac{s_y}{s_x} R$$

and where the intercept is given by

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Conditions for Using Linear Regression

In order to responsibly use linear regression. . .

Conditions for Using Linear Regression

In order to responsibly use linear regression...

1. Relationship between explanatory and response variables must be approximately linear. (**Linear**)
    - Check using scatterplot and/or residual plot

Conditions for Using Linear Regression

In order to responsibly use linear regression. . .

1. Relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot and/or residual plot
2. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
   - Check using histogram of residuals

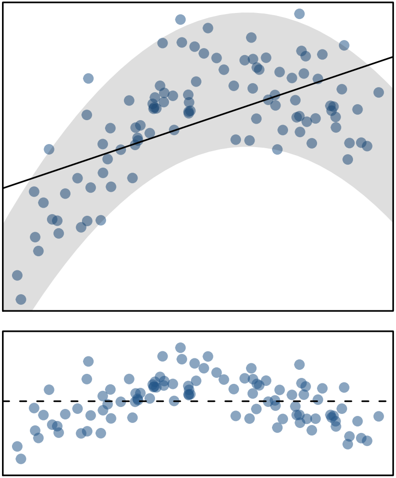
Conditions for Using Linear Regression

In order to responsibly use linear regression...

1. Relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot and/or residual plot
2. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
   - Check using histogram of residuals
3. The variability of residuals should be roughly constant across entire data set. (**Homoscedastic**)
   - Check using resdidual plot.

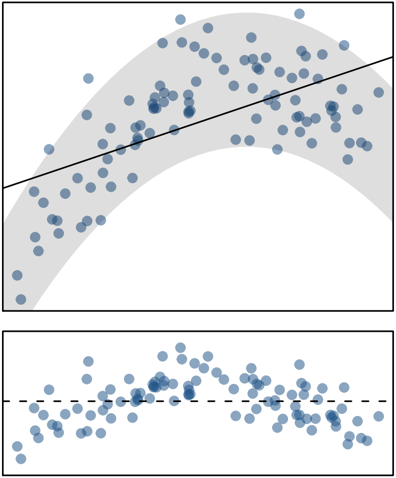
Checking Conditions I

- What condition is this linear model most obviously violating?
  a. Linearity
  b. Normalacy
  c. Homoscedasticity
  d. Extreme Outliers



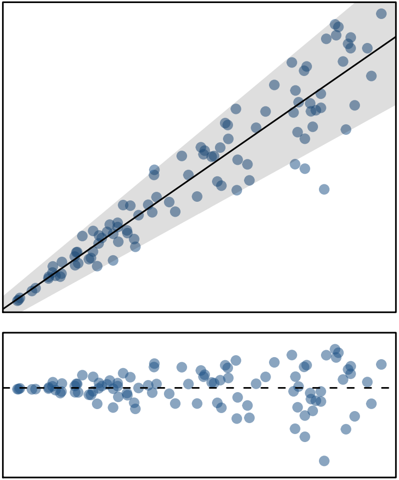
Checking Conditions I

- What condition is this linear model most obviously violating?
  a. Linearity
  b. Normalacy
  c. Homoscedasticity
  d. Extreme Outliers



Checking Conditions II

- What condition is this linear model most obviously violating?
  a. Linearity
  b. Normalacy
  c. Homoscedasticity
  d. Extreme Outliers



Checking Conditions II

- What condition is this linear model most obviously violating?
  a. Linearity
  b. Normalacy
  c. Homoscedasticity
  d. Extreme Outliers