Linear Models

Nate Wells

Math 141, 2/18/22

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Discuss accuracy and appropriateness of linear models
- Work through an example of linear regression

Section 1

Assessing Accuracy of Linear Models

Review: 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 $\bar{x}, \bar{y}$, 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
$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

  - The least squares line minimizes the sum of squared residuals $e_1^2 + e_2^2 + \cdots + e_n^2$.

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

  - The least squares line minimizes the sum of squared residuals $e_1^2 + e_2^2 + \cdots + e_n^2$.
  - The line *always* goes through the point $(\bar{x}, \bar{y})$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

  - The least squares line minimizes the sum of squared residuals $e_1^2 + e_2^2 + \cdots + e_n^2$.
  - The line *always* goes through the point $(\bar{x}, \bar{y})$
  - The slope has the same sign as the correlation coefficient

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

  - The least squares line minimizes the sum of squared residuals $e_1^2 + e_2^2 + \cdots + e_n^2$.
  - The line *always* goes through the point $(\bar{x}, \bar{y})$
  - The slope has the same sign as the correlation coefficient
  - The slope is close to 0 when either $R \approx 0$ *or* when s_x is much bigger than s_y

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

  - The least squares line minimizes the sum of squared residuals $e_1^2 + e_2^2 + \cdots + e_n^2$.
  - The line *always* goes through the point $(\bar{x}, \bar{y})$
  - The slope has the same sign as the correlation coefficient
  - The slope is close to 0 when either $R \approx 0$ *or* when s_x is much bigger than s_y
  - A large slope does not indicate strong correlation and a small slope does not indicate lack of correlation

Goals for Regression

Least squared regression is used for 3 primary tasks:

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
    - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
    - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.
2. **Predicting** values of the response variable based on values of the explanatory variable

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
   - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.
2. **Predicting** values of the response variable based on values of the explanatory variable
   - EX: Using data between 1960 and 2015, we predict the atmosphere will contain 410 ppm CO_2 in 2025.

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
    - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.
2. **Predicting** values of the response variable based on values of the explanatory variable
    - EX: Using data between 1960 and 2015, we predict the atmosphere will contain 410 ppm CO_2 in 2025.
3. **Inferring** relationships about a population based on relationships observed in a sample.

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
   - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.
2. **Predicting** values of the response variable based on values of the explanatory variable
   - EX: Using data between 1960 and 2015, we predict the atmosphere will contain 410 ppm CO2 in 2025.
3. **Inferring** relationships about a population based on relationships observed in a sample.
   - EX: Based on the negative correlation between poverty and graduation rate observed in the sample of states in 2020, we infer that in general, a state's poverty and graduation rate are negatively correlated.

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
    - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.
2. **Predicting** values of the response variable based on values of the explanatory variable
    - EX: Using data between 1960 and 2015, we predict the atmosphere will contain 410 ppm CO2 in 2025.
3. **Inferring** relationships about a population based on relationships observed in a sample.
    - EX: Based on the negative correlation between poverty and graduation rate observed in the sample of states in 2020, we infer that in general, a state's poverty and graduation rate are negatively correlated.

- We can *always* find the line of best fit to explore data.

Goals for Regression

Least squared regression is used for 3 primary tasks:

1. **Exploring** and **summarizing** relationships between quantitative variables in a data set.
    - EX: In 2020, we observe that countries with higher GDP per capita consistently have higher average life expectancy.
2. **Predicting** values of the response variable based on values of the explanatory variable
    - EX: Using data between 1960 and 2015, we predict the atmosphere will contain 410 ppm CO_2 in 2025.
3. **Inferring** relationships about a population based on relationships observed in a sample.
    - EX: Based on the negative correlation between poverty and graduation rate observed in the sample of states in 2020, we infer that in general, a state's poverty and graduation rate are negatively correlated.

- We can *always* find the line of best fit to explore data.
- However, if we want to make accurate predictions or justified inference, we need to ensure certain conditions are satisfied.

Conditions for Using Linear Regression

In order to responsibly use linear regression for prediction or inference, we require:

Conditions for Using Linear Regression

In order to responsibly use linear regression for prediction or inference, we require:

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

Conditions for Using Linear Regression

In order to responsibly use linear regression for prediction or inference, we require:

1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
    - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
    - Check using scatterplot/residual plot, as well as sample design
3. 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 for prediction or inference, we require:

1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
   - Check using scatterplot/residual plot, as well as sample design
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
   - Check using histogram of residuals
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)
   - Check using residual plot.

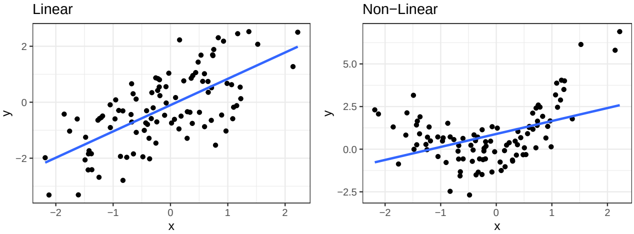
Conditions for Using Linear Regression

In order to responsibly use linear regression for prediction or inference, we require:

1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
    - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
    - Check using scatterplot/residual plot, as well as sample design
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
    - Check using histogram of residuals
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)
    - Check using residual plot.

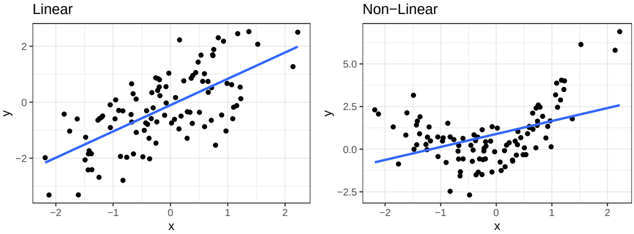
Linearity

1. The relationship between explanatory and response variables must be approximately linear.



Linearity

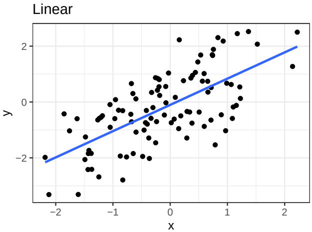
1. The relationship between explanatory and response variables must be approximately linear.

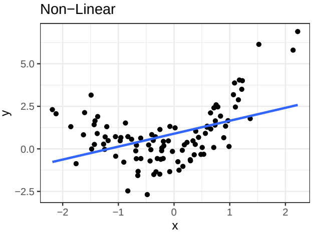


- If data is non-linear. . .

Linearity

1. The relationship between explanatory and response variables must be approximately linear.

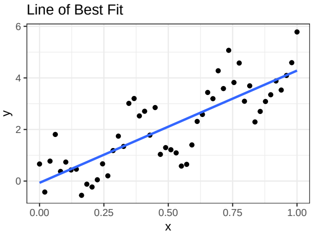


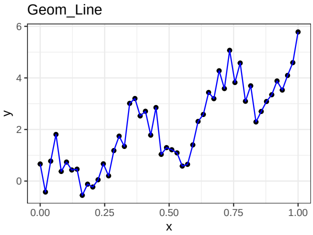


- If data is non-linear...
  - Slope does not adequately describe relationship
  - Predictions can be very inaccurate
  - More advanced modeling techniques should be used (Math 243)

Independent Observations

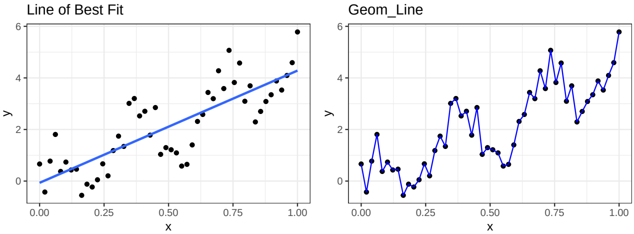
2. The observations should be independent of one another.





Independent Observations

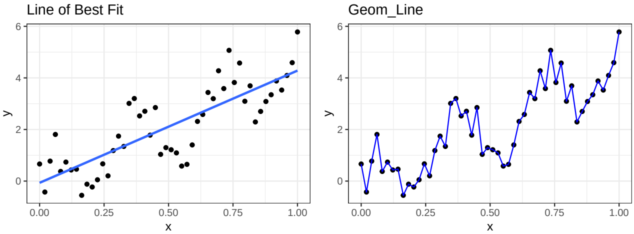
2. The observations should be independent of one another.



- If observations are not independent. . .

Independent Observations

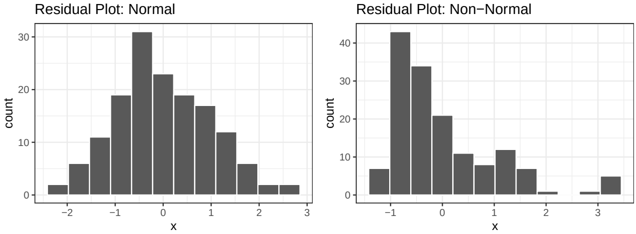
2. The observations should be independent of one another.



- If observations are not independent...
  - Coincidental trends more likely to appear
  - Slope and intercept estimates are more variable in sample
  - More advanced modeling techniques should be used (Math 243)

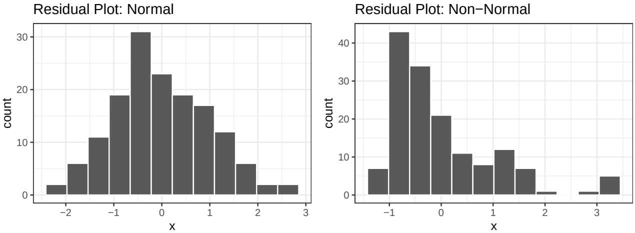
Normal Residuals

3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0.



Normal Residuals

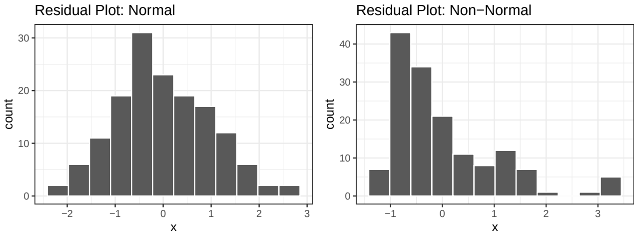
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0.



- If residuals are non-Normal...

Normal Residuals

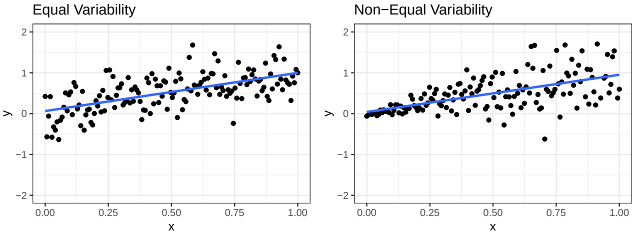
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0.



- If residuals are non-Normal. . .
  - Cannot estimate trends in population
  - Some predictions can be very inaccurate
  - More advanced modeling techniques should be used (Math 243)

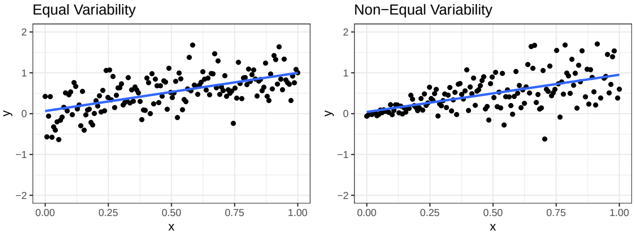
Equal Variability

4. The variability of residuals should be roughly constant across entire data set.



Equal Variability

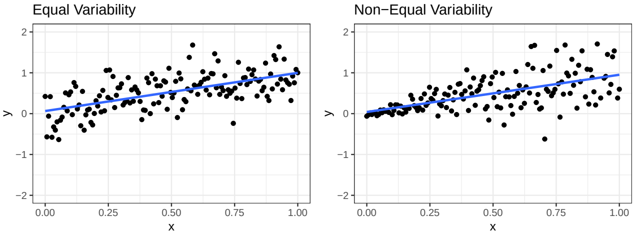
4. The variability of residuals should be roughly constant across entire data set.



- If residuals don't have equal variability. . .

Equal Variability

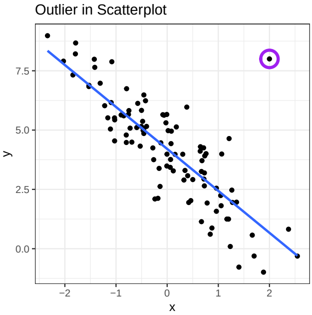
4. The variability of residuals should be roughly constant across entire data set.

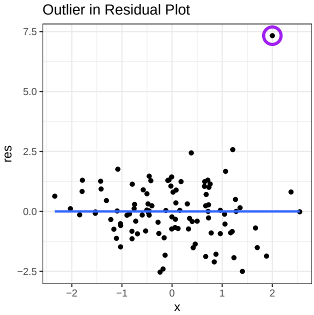


- If residuals don't have equal variability. . .
  - Inference about the population may be misleading
  - Outliers in high-variability range are more influential
  - More advanced modeling techniques should be used (Math 243)

Outliers

- An **outlier** in regression is an observation that lies far from the cloud of data points.

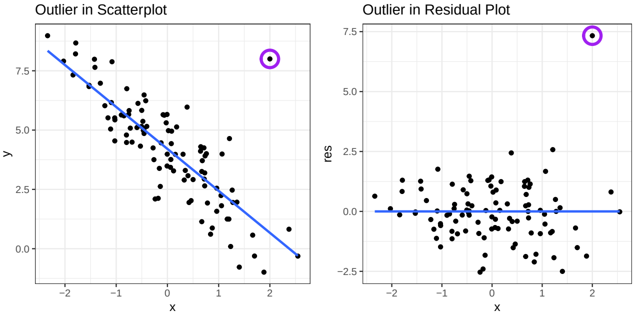




- Outliers can arise for a variety of reasons. . .

Outliers

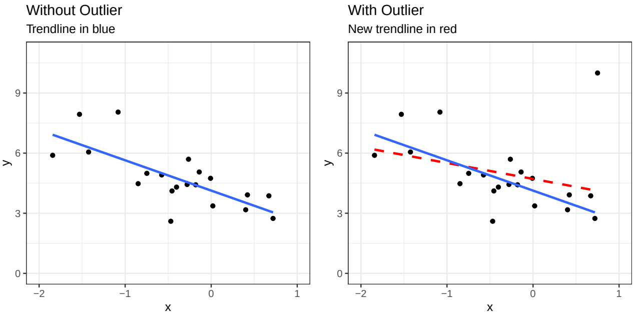
- An **outlier** in regression is an observation that lies far from the cloud of data points.



- Outliers can arise for a variety of reasons. . .
  - Measurement, recording, or reporting error
  - Evidence of possible confounding variable
  - Random chance in sampling

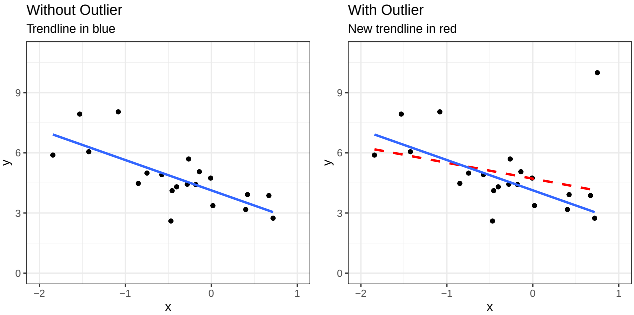
Effect of Outliers on Least Squares

- The least squares line is *not* robust against outliers



Effect of Outliers on Least Squares

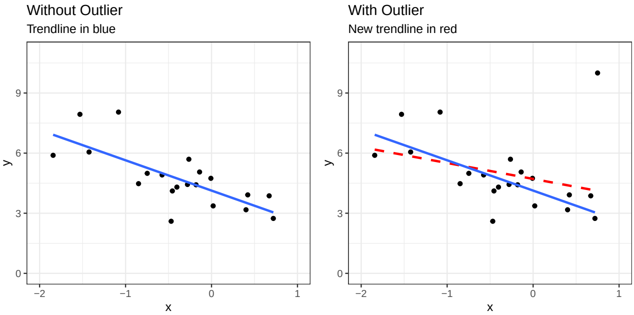
- The least squares line is *not* robust against outliers



- Outliers that have both extreme y values and extreme x values have the potential to significantly change slope and intercept of regression line

Effect of Outliers on Least Squares

- The least squares line is *not* robust against outliers



- Outliers that have both extreme y values and extreme x values have the potential to significantly change slope and intercept of regression line
- But unless you have very good reason to, do not remove outliers (they tell an important story about the data)

Coefficient of Variation

- The correlation coefficient R measures the strength and direction of a linear relationship.

Coefficient of Variation

- The correlation coefficient R measures the strength and direction of a linear relationship.
- But another common measure of the *strength* of a linear relationship is the **coefficient of variation** R^2 (sometimes just called "R-squared")
    - Since R is a number between -1 and 1, then R^2 will always be between 0 and 1.

Coefficient of Variation

- The correlation coefficient R measures the strength and direction of a linear relationship.
- But another common measure of the *strength* of a linear relationship is the **coefficient of variation** R^2 (sometimes just called "R-squared")
  - Since R is a number between -1 and 1, then R^2 will always be between 0 and 1.
- The value of R^2 measures the proportion of variation in the response variable Y that is explained by its linear relationship with the explanatory variable X.

Coefficient of Variation

- The correlation coefficient R measures the strength and direction of a linear relationship.
- But another common measure of the *strength* of a linear relationship is the **coefficient of variation** R^2 (sometimes just called "R-squared")
  - Since R is a number between -1 and 1, then R^2 will always be between 0 and 1.
- The value of R^2 measures the proportion of variation in the response variable Y that is explained by its linear relationship with the explanatory variable X.
- R^2 can also be computed as

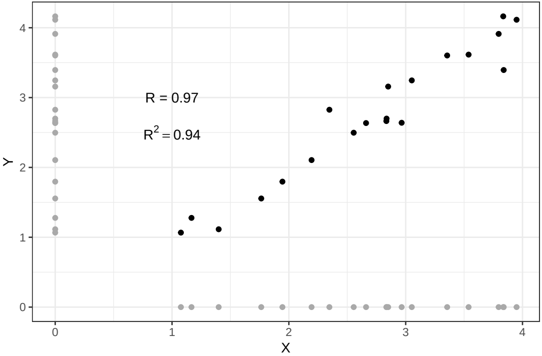
$$R^2 = \frac{\text{Variability in Y explained by X}}{\text{Variability in Y}} = \frac{s_y^2 - s_{\text{res}}^2}{s_y^2}$$

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the explanatory variable.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the explanatory variable.

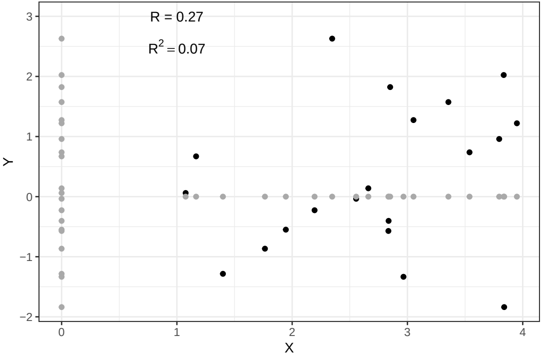


Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the explanatory variable.

Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the explanatory variable.



Section 2

Linear Regression in Practice

Review of Regression in R

1. State research question and identify variables

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`
5. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`
5. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
6. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`
5. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
6. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`
7. Plot regression line
   - `ggplot( ... ) + geom_point() + geom_smooth(method = "lm", se = F)`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`
5. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
6. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`
7. Plot regression line
   - `ggplot( ... ) + geom_point() + geom_smooth(method = "lm", se = F)`
8. Calculate residuals and create residual plot
   - `model_residuals <- get_regression_points(nice_model)`
   - `ggplot(model_residuals, aes(x = var1, y = residual) ) + geom_point() + geom_smooth(method = "lm", se = F)`

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`
5. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
6. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`
7. Plot regression line
   - `ggplot( ... ) + geom_point() + geom_smooth(method = "lm", se = F)`
8. Calculate residuals and create residual plot
   - `model_residuals <- get_regression_points(nice_model)`
   - `ggplot(model_residuals, aes(x = var1, y = residual) ) + geom_point() + geom_smooth(method = "lm", se = F)`
9. Assess model conditions and investigate outliers.

Review of Regression in R

1. State research question and identify variables
2. Load data
   - `the_data <- read_csv("example.csv")`
3. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) + geom_point()`
4. Compute correlation and R^2 for pair of variables
   - `the_data %>% summarize(cor = cor(var1, var2)) %>% mutate(R_sq = cor^2)`
5. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
6. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`
7. Plot regression line
   - `ggplot( ... ) + geom_point() + geom_smooth(method = "lm", se = F)`
8. Calculate residuals and create residual plot
   - `model_residuals <- get_regression_points(nice_model)`
   - `ggplot(model_residuals, aes(x = var1, y = residual) ) + geom_point() + geom_smooth(method = "lm", se = F)`
9. Assess model conditions and investigate outliers.
10. Make conclusions.

Poverty and Graduation Rate

- **Research Question**: In the contemporary United States, what is the relationship between poverty rate and graduation rate at the state level?

Poverty and Graduation Rate

- **Research Question**: In the contemporary United States, what is the relationship between poverty rate and graduation rate at the state level?
    - **Explanatory Variable**: Poverty Rate
    - **Response Variable**: Graduation Rate
    - **Population**: The contemporary United States
    - **Sample**: US States (2018 - 2020)

Poverty and Graduation Rate

- **Research Question**: In the contemporary United States, what is the relationship between poverty rate and graduation rate at the state level?
  - **Explanatory Variable**: Poverty Rate
  - **Response Variable**: Graduation Rate
  - **Population**: The contemporary United States
  - **Sample**: US States (2018 - 2020)
- **Research Method**: Build a linear model for graduation rate as a function of poverty rate, using individual states as observations.

Poverty and Graduation Rate

- **Research Question**: In the contemporary United States, what is the relationship between poverty rate and graduation rate at the state level?
  - **Explanatory Variable**: Poverty Rate
  - **Response Variable**: Graduation Rate
  - **Population**: The contemporary United States
  - **Sample**: US States (2018 - 2020)
- **Research Method**: Build a linear model for graduation rate as a function of poverty rate, using individual states as observations.
- **Data**: We've obtained data called `states` on poverty rate from the 2020 US Census, and data on graduation rate from a 2018-2019 report by NCES

Poverty and Graduation Rate

- **Research Question**: In the contemporary United States, what is the relationship between poverty rate and graduation rate at the state level?
    - **Explanatory Variable**: Poverty Rate
    - **Response Variable**: Graduation Rate
    - **Population**: The contemporary United States
    - **Sample**: US States (2018 - 2020)
- **Research Method**: Build a linear model for graduation rate as a function of poverty rate, using individual states as observations.
- **Data**: We've obtained data called `states` on poverty rate from the 2020 US Census, and data on graduation rate from a 2018-2019 report by NCES
    - `grad_rate` denotes the adjusted cohort graduation rate (percent of high school freshmen who finish with regular diploma within 4 years of starting 9th grade)

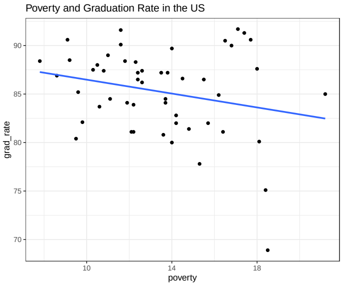
Poverty and Graduation Rate

- **Research Question**: In the contemporary United States, what is the relationship between poverty rate and graduation rate at the state level?
    - **Explanatory Variable**: Poverty Rate
    - **Response Variable**: Graduation Rate
    - **Population**: The contemporary United States
    - **Sample**: US States (2018 - 2020)
- **Research Method**: Build a linear model for graduation rate as a function of poverty rate, using individual states as observations.
- **Data**: We've obtained data called `states` on poverty rate from the 2020 US Census, and data on graduation rate from a 2018-2019 report by NCES
    - `grad_rate` denotes the adjusted cohort graduation rate (percent of high school freshmen who finish with regular diploma within 4 years of starting 9th grade)
    - `poverty` denotes the proportion of state population living below poverty threshold (26,246 per person, for family of 4 with two children)

Exploratory Analysis

- Visualize Relationship using ggplot2

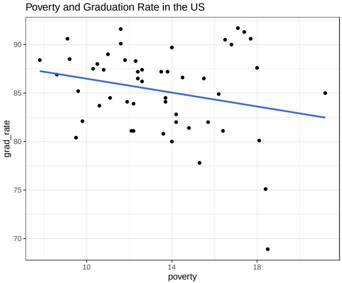
```
ggplot(states, aes(y = grad_rate, x = poverty)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Poverty and Graduation Rate in the US")+
  theme_bw()
```



Exploratory Analysis

- Visualize Relationship using ggplot2

```
ggplot(states, aes(y = grad_rate, x = poverty)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Poverty and Graduation Rate in t
  theme_bw()
```



- Compute relevant summary statistics using dplyr

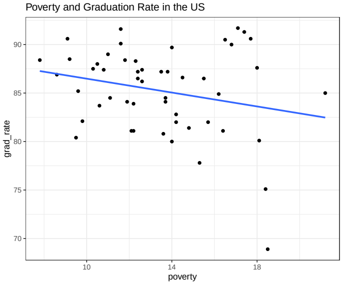
```
states %>% summarize(
  mean_poverty = mean(poverty),
  sd_poverty = sd(poverty),
  mean_grad = mean(grad_rate),
  sd_grad = sd(grad_rate))
```

```
## # A tibble: 1 x 4
##   mean_poverty sd_poverty mean_grad sd_grad
##          <dbl>      <dbl>     <dbl>   <dbl>
## 1         13.5       3.02      85.2    4.48
```

Exploratory Analysis

- Visualize Relationship using ggplot2

```
ggplot(states, aes(y = grad_rate, x = poverty)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Poverty and Graduation Rate in t
  theme_bw()
```



- Compute relevant summary statistics using dplyr

```
states %>% summarize(
  mean_poverty = mean(poverty),
  sd_poverty = sd(poverty),
  mean_grad = mean(grad_rate),
  sd_grad = sd(grad_rate))
```

```
## # A tibble: 1 x 4
##   mean_poverty sd_poverty mean_grad sd_grad
##          <dbl>      <dbl>     <dbl>   <dbl>
## 1         13.5       3.02      85.2    4.48
```

```
states %>% summarize(
  R = cor(grad_rate, poverty)) %>%
  mutate(R_sq = R^2)
```

```
## # A tibble: 1 x 2
##        R   R_sq
##    <dbl>  <dbl>
## 1 -0.241 0.0582
```

Fit the Linear Model

- Fit the linear modeling using `lm`

```
states_mod <- lm(grad_rate ~ poverty, data = states)
```

Fit the Linear Model

- Fit the linear modeling using `lm`

```
states_mod <- lm(grad_rate ~ poverty, data = states)
```

- Get the regression equation

```
get_regression_table(states_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   90.1       2.84      31.8    0         84.4     95.8
## 2 poverty     -0.358     0.206     -1.74   0.088    -0.771     0.055
```

Fit the Linear Model

- Fit the linear modeling using `lm`

```
states_mod <- lm(grad_rate ~ poverty, data = states)
```

- Get the regression equation

```
get_regression_table(states_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   90.1       2.84      31.8    0        84.4     95.8
## 2 poverty     -0.358     0.206     -1.74   0.088    -0.771    0.055
```

- Express the coefficients in terms of a linear function

$$\text{Grad Rate } = 90.1 - 0.358 \cdot \text{Poverty}$$

Fit the Linear Model

- Fit the linear modeling using `lm`

```
states_mod <- lm(grad_rate ~ poverty, data = states)
```

- Get the regression equation

```
get_regression_table(states_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   90.1       2.84      31.8    0        84.4     95.8
## 2 poverty     -0.358     0.206     -1.74   0.088    -0.771    0.055
```

- Express the coefficients in terms of a linear function

$$\text{Grad Rate} \; = 90.1 - 0.358 \cdot \text{Poverty}$$

- Interpret the coefficients

Fit the Linear Model

- Fit the linear modeling using `lm`

```
states_mod <- lm(grad_rate ~ poverty, data = states)
```

- Get the regression equation

```
get_regression_table(states_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   90.1       2.84      31.8    0        84.4     95.8
## 2 poverty     -0.358     0.206     -1.74   0.088    -0.771    0.055
```

- Express the coefficients in terms of a linear function

$$\text{Grad Rate } = 90.1 - 0.358 \cdot \text{Poverty}$$

- Interpret the coefficients
  - Every 1 unit increase in poverty corresponds to a .358 unit decrease in graduation rate.

Fit the Linear Model

- Fit the linear modeling using `lm`

```
states_mod <- lm(grad_rate ~ poverty, data = states)
```

- Get the regression equation

```
get_regression_table(states_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   90.1       2.84      31.8    0        84.4     95.8
## 2 poverty     -0.358     0.206     -1.74   0.088    -0.771    0.055
```

- Express the coefficients in terms of a linear function

$$\widehat{\text{Grad Rate}} = 90.1 - 0.358 \cdot \text{Poverty}$$

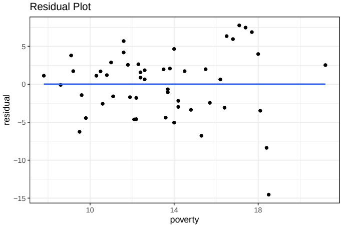
- Interpret the coefficients
  - Every 1 unit increase in poverty corresponds to a .358 unit decrease in graduation rate.
  - The predicted graduation rate for a state with 0 poverty is 90.062

Calculate Residuals

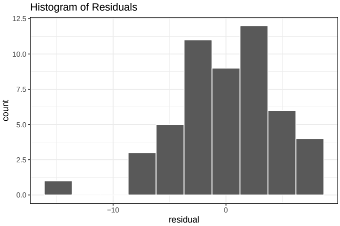
- Get residuals

```
state_residuals <- get_regression_points(states_mod)
```

```
ggplot(state_residuals,
       aes(x = poverty, y = residual)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual Plot")+
  theme_bw()
```




```
ggplot(state_residuals,
       aes(x = residual)) +
  geom_histogram(bins = 10, color = "white")+
  labs(title = "Histogram of Residuals")+
  theme_bw()
```



Assess Conditions for Linear Regression

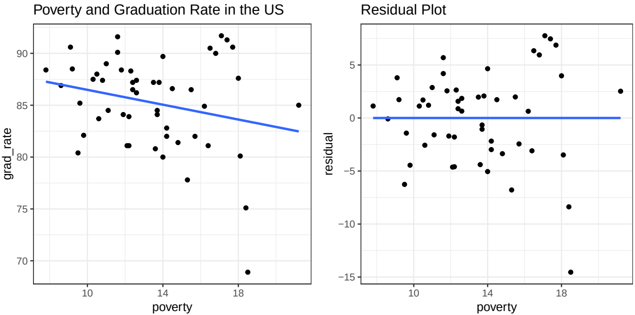
1. Linearity?

Assess Conditions for Linear Regression

1. Linearity?

Assess Conditions for Linear Regression

1. Linearity?



- Scatterplot suggests a weak linear relationship between `poverty` and `grad_rate`, but residual plot doesn't show any strong non-linear trends

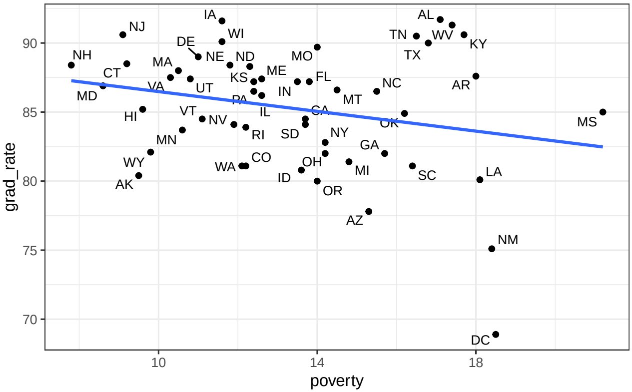
Assess Conditions for Linear Regression

2. Independence?

Assess Conditions for Linear Regression

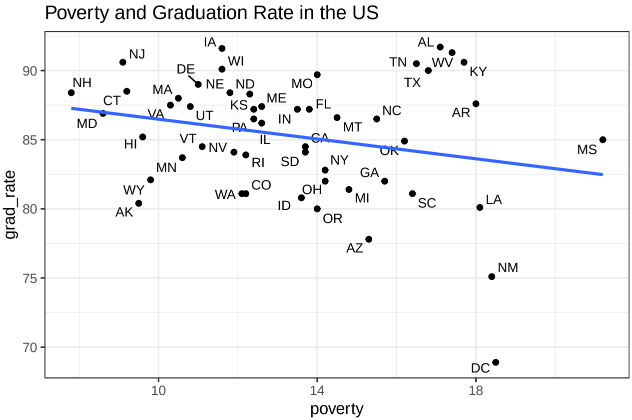
2. Independence?

Poverty and Graduation Rate in the US



Assess Conditions for Linear Regression

2. Independence?



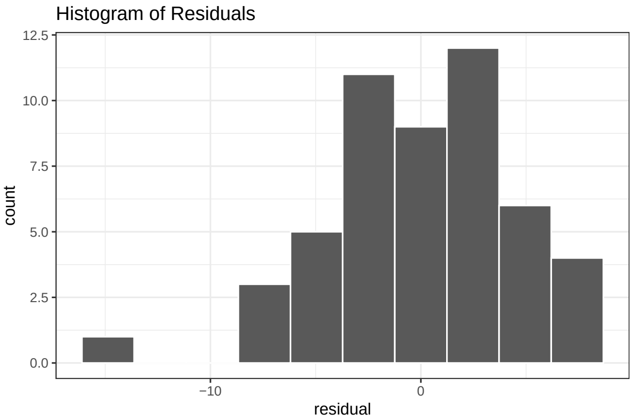
- States in close geographic proximity tend to have similar poverty and grad rates.

Assess Conditions for Linear Regression

3. Normal residuals?

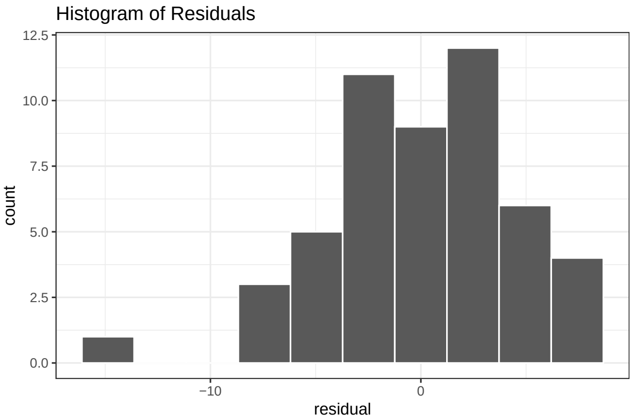
Assess Conditions for Linear Regression

3. Normal residuals?



Assess Conditions for Linear Regression

3. Normal residuals?



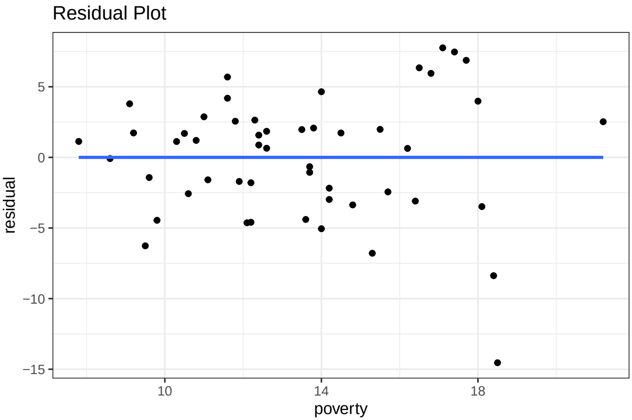
- Residual distribution appears to be (mostly) symmetric, unimodal, and centered at 0. Roughly bell-shaped. But with 1 notable outlier.

Assess Conditions for Linear Regression

4. Equal Variability?

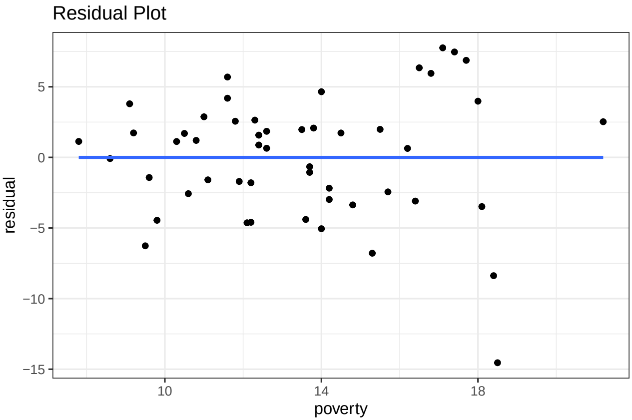
Assess Conditions for Linear Regression

4. Equal Variability?



Assess Conditions for Linear Regression

4. Equal Variability?



- Variability in outliers is relatively consistent across poverty range (with exception of outlier)

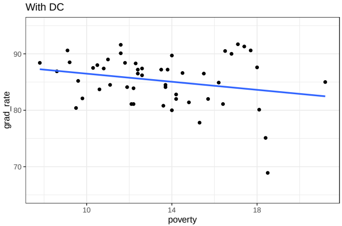
Investigate Outliers

- What's up with DC?

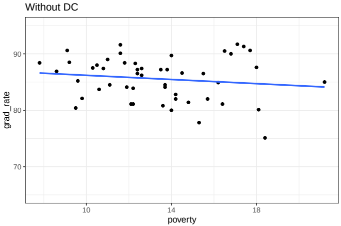
Investigate Outliers

- What's up with DC?

```
states %>%
ggplot(aes(y = grad_rate, x = poverty)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "With DC")+theme_bw()+
  scale_y_continuous(limits = c(65,95))
```




```
states %>% filter(abbr!="DC") %>%
ggplot(aes(y = grad_rate, x = poverty)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Without DC")+theme_bw()+
  scale_y_continuous(limits = c(65,95))
```




```
states %>%
  summarize(R = cor(poverty, grad_rate))
```

```
## # A tibble: 1 x 1
##         R
##     <dbl>
## 1 -0.241
```

```
states %>% filter(abbr != "DC") %>%
  summarize(R = cor(poverty, grad_rate))
```

```
## # A tibble: 1 x 1
##         R
##     <dbl>
## 1 -0.142
```

Conclusions

- What have we learned?

Conclusions

- What have we learned?
    - Based on data from 2018 - 2020, there seems to be some evidence of a negative linear relationship between poverty rate and graduation rate ($R = -.24$)

Conclusions

- What have we learned?
  - Based on data from 2018 - 2020, there seems to be some evidence of a negative linear relationship between poverty rate and graduation rate ($R = -.24$)
  - However, an outlier (Washington D.C.) was influential in the model, and with this outlier removed, the linear relationship was considerably weaker ($R = -.14$)

Conclusions

- What have we learned?
  - Based on data from 2018 - 2020, there seems to be some evidence of a negative linear relationship between poverty rate and graduation rate ($R = -.24$)
  - However, an outlier (Washington D.C.) was influential in the model, and with this outlier removed, the linear relationship was considerably weaker ($R = -.14$)
  - Geo-politically similar states appear to have similar graduation and poverty rates, raising concerns about independence of observations; variability of residuals in this sample may not represent variability overall

Conclusions

- What have we learned?
  - Based on data from 2018 - 2020, there seems to be some evidence of a negative linear relationship between poverty rate and graduation rate ($R = -.24$)
  - However, an outlier (Washington D.C.) was influential in the model, and with this outlier removed, the linear relationship was considerably weaker ($R = -.14$)
  - Geo-politically similar states appear to have similar graduation and poverty rates, raising concerns about independence of observations; variability of residuals in this sample may not represent variability overall
  - Further studies should be conducted to assess whether these trends (a) change over time, and (b) are replicated at smaller scale.