Linear Regression III

Nate Wells

Math 141, 3/1/21

Outline

In this lecture, we will...

Outline

In this lecture, we will...

- Review conditions for linear regression
- Create linear models comparing a quantitative response and categorical explanatory variables

Section 1

Introduction to Linear Regression

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

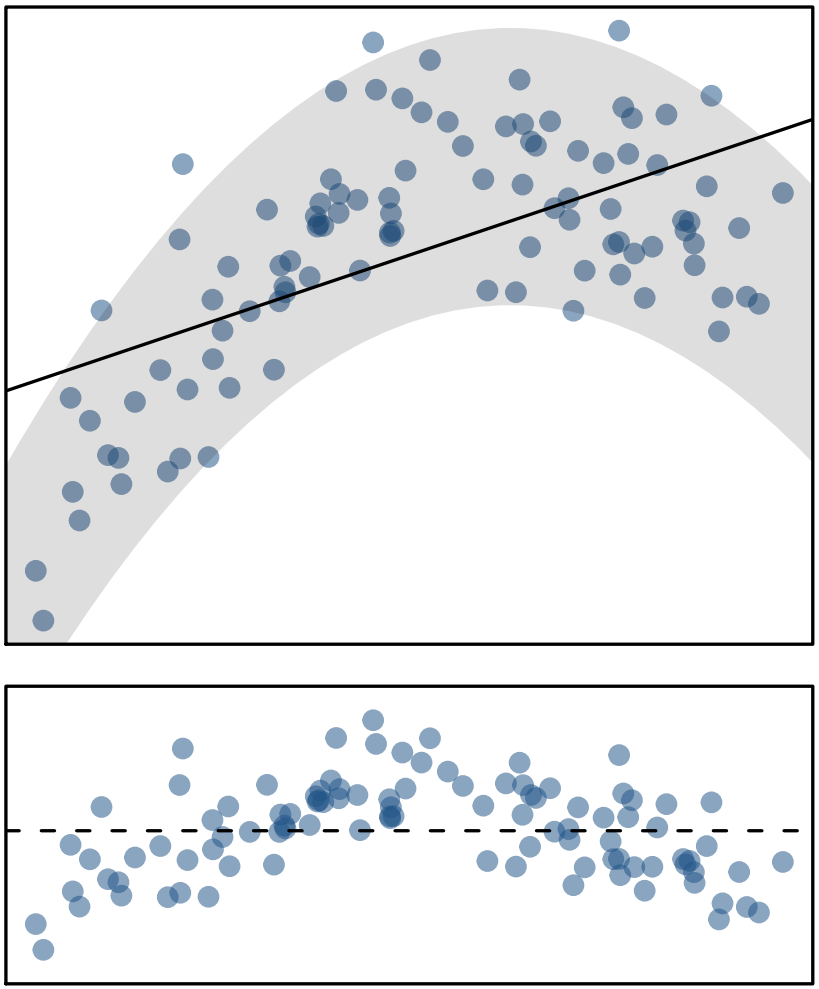
Conditions for Using Linear Regression

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.

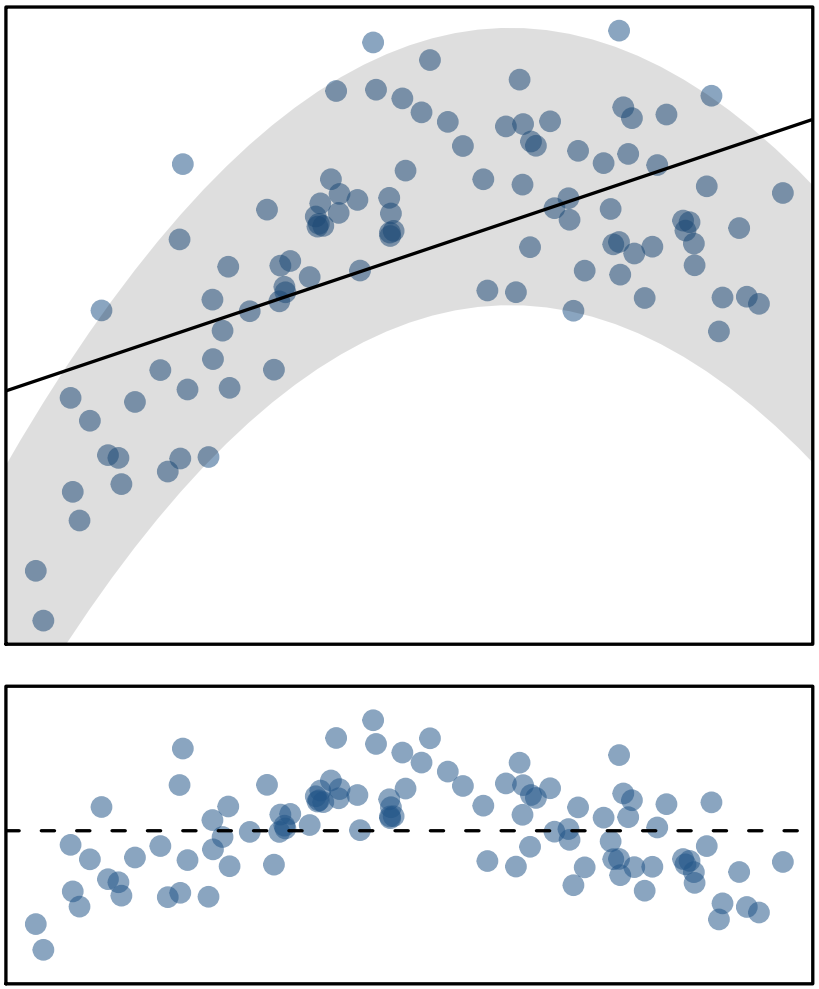
Checking Conditions I

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



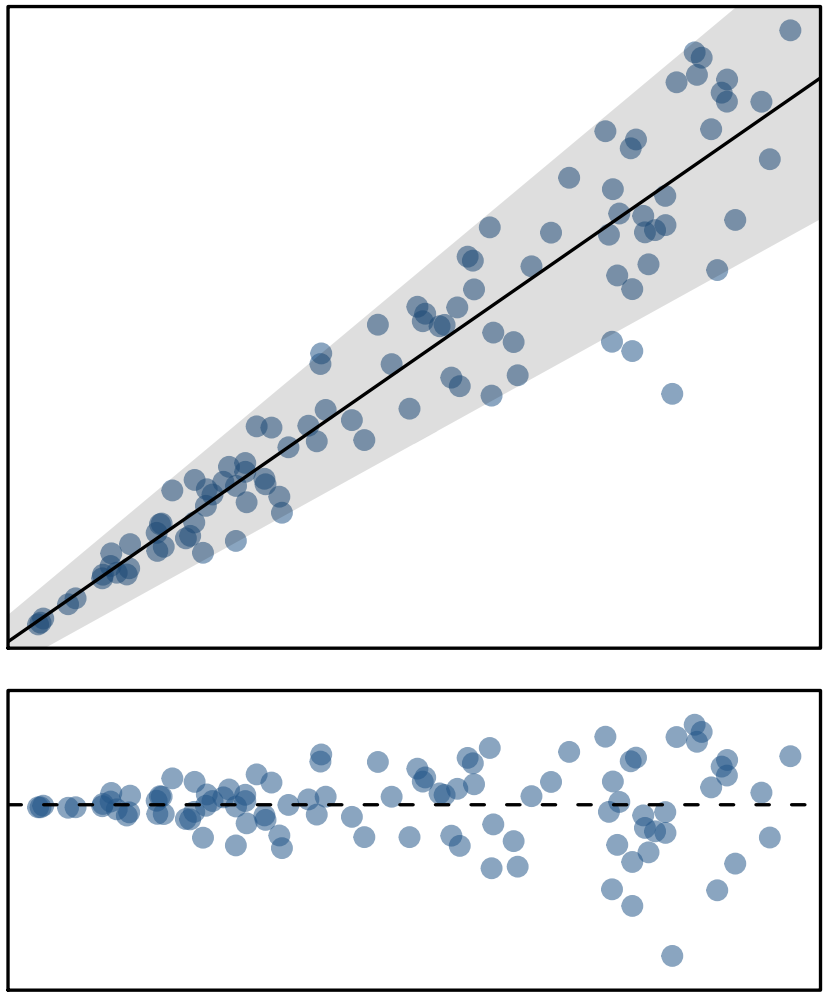
Checking Conditions I

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



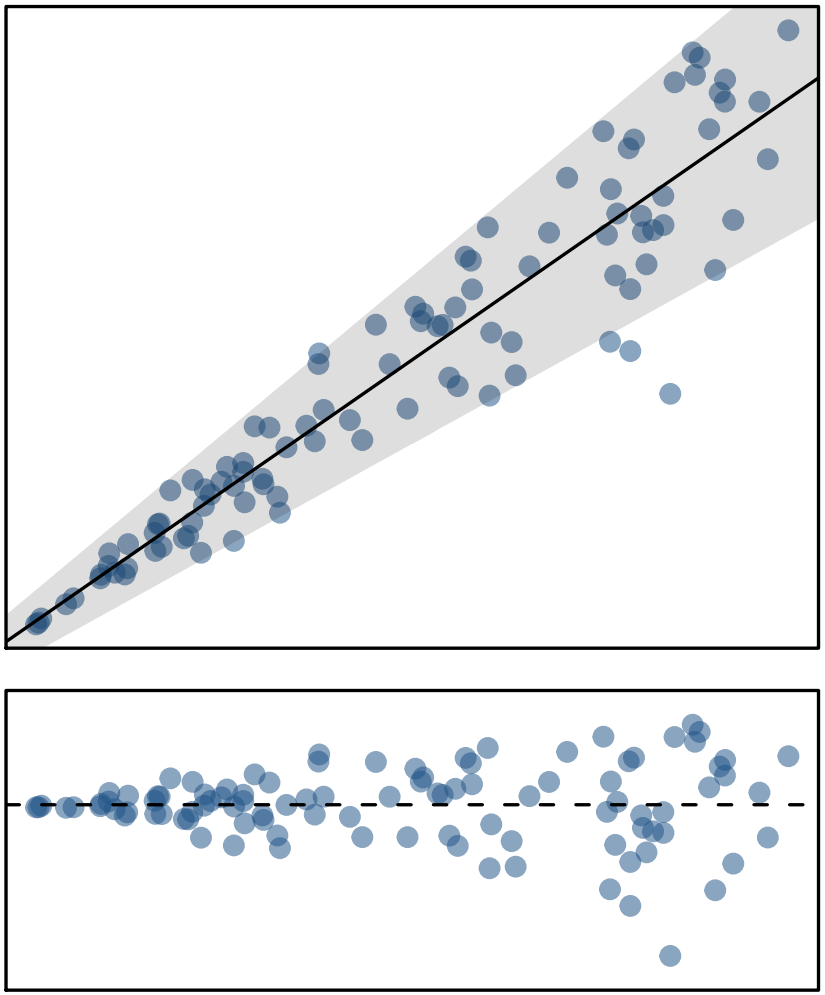
Checking Conditions II

- 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



Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`

Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`
2. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) +geom_point()`

Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`
2. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) +geom_point()`
3. Compute correlation for pair of variables
   - `the_data %>% get_correlation(function = var2 ~ var1)`

Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`
2. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) +geom_point()`
3. Compute correlation for pair of variables
   - `the_data %>% get_correlation(function = var2 ~ var1)`
4. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`

Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`
2. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) +geom_point()`
3. Compute correlation for pair of variables
   - `the_data %>% get_correlation(function = var2 ~ var1)`
4. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
5. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`

Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`
2. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) +geom_point()`
3. Compute correlation for pair of variables
   - `the_data %>% get_correlation(function = var2 ~ var1)`
4. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
5. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`
6. Plot regression line
   - `ggplot( ... ) + geom_point() + geom_smooth(method = "lm")`

Review of Regression Using ModernDive

The basic procedure for linear regression using ModernDive:

1. Load data and store as a variable
   - `the_data <- read_csv("example.csv")`
2. Perform exploratory data analysis (using dplyr and ggplot)
   - `ggplot(the_data, aes(x = var1, y = var2) ) +geom_point()`
3. Compute correlation for pair of variables
   - `the_data %>% get_correlation(function = var2 ~ var1)`
4. Fit a linear model to the data
   - `nice_model<- lm(var2 ~ var1, data = the_data)`
5. Get equation of regression line from regression table
   - `get_regression_table(nice_model)`
6. Plot regression line
   - `ggplot( ... ) + geom_point() + geom_smooth(method = "lm")`
7. Calculate residuals and create residual plot
   - `get_regression_points(nice_model)`
   - `ggplot( aes(x = var1, y = residual) ) + geom_point() + geom_smooth(method = "lm")`

Section 2

Regression for 1 Categorical Variable

Overview of Regression for a Categorical Variable

- Basic linear regression models a linear relationship between two **quantitative** variables.

Overview of Regression for a Categorical Variable

- Basic linear regression models a linear relationship between two **quantitative** variables.
- But research questions often concern one quantitative and one categorical variable

Overview of Regression for a Categorical Variable

- Basic linear regression models a linear relationship between two **quantitative** variables.
- But research questions often concern one quantitative and one categorical variable
  - Sometimes a quantitative variable is grouped into discrete levels to form a categorical variable
  - In other cases, interested in measuring magnitude of effect of different levels of another variable

Overview of Regression for a Categorical Variable

- Basic linear regression models a linear relationship between two **quantitative** variables.
- But research questions often concern one quantitative and one categorical variable
  - Sometimes a quantitative variable is grouped into discrete levels to form a categorical variable
  - In other cases, interested in measuring magnitude of effect of different levels of another variable
- We can still use a basic linear regression to represent the quantitative variable as function of the categorical variable.
  - But the interpretation of "best" fitting line is slightly different.

An Extended Example

- As proof-of-concept for Categorical Regression, we will investigate the research question:

An Extended Example

- As proof-of-concept for Categorical Regression, we will investigate the research question:
  What factors are associated with increased rate of hate crimes?

An Extended Example

- As proof-of-concept for Categorical Regression, we will investigate the research question:
  What factors are associated with increased rate of hate crimes?
- This analysis is based on the Jan. 23, 2017 *Five-Thirty-Eight* article "Higher Rates Of Hate Crimes Are Tied To Income Inequality"

An Extended Example

- As proof-of-concept for Categorical Regression, we will investigate the research question:
  What factors are associated with increased rate of hate crimes?
- This analysis is based on the Jan. 23, 2017 *Five-Thirty-Eight* article "Higher Rates Of Hate Crimes Are Tied To Income Inequality"
- To begin, we load the data

The Data

```
## Rows: 47
## Columns: 9
## $ state             <chr> "New Mexico", "Maine", "New York", "Illinois", "De...
## $ median_house_inc  <chr> "low", "low", "low", "low", "high", "high", "high"...
## $ share_pop_metro   <dbl> 0.69, 0.54, 0.94, 0.90, 0.90, 1.00, 0.87, 0.86, 0....
## $ hs                <dbl> 83, 90, 85, 86, 87, 85, 89, 90, 81, 91, 89, 89, 87...
## $ hate_crimes       <dbl> 0.295, 0.616, 0.351, 0.195, 0.323, 0.095, 0.833, 0...
## $ trump_support     <chr> "low", "low", "low", "low", "low", "low", "low", "...
## $ unemployment      <chr> "high", "low", "low", "high", "low", "high", "high...
## $ urbanization      <chr> "low", "low", "high", "high", "high", "high", "hig...
## $ income            <dbl> 46686, 51710, 54310, 54916, 57522, 58633, 58875, 5...
```

The Data

```
## Rows: 47
## Columns: 9
## $ state             <chr> "New Mexico", "Maine", "New York", "Illinois", "De...
## $ median_house_inc  <chr> "low", "low", "low", "low", "high", "high", "high"...
## $ share_pop_metro   <dbl> 0.69, 0.54, 0.94, 0.90, 0.90, 1.00, 0.87, 0.86, 0....
## $ hs                <dbl> 83, 90, 85, 86, 87, 85, 89, 90, 81, 91, 89, 89, 87...
## $ hate_crimes       <dbl> 0.295, 0.616, 0.351, 0.195, 0.323, 0.095, 0.833, 0...
## $ trump_support     <chr> "low", "low", "low", "low", "low", "low", "low", "...
## $ unemployment      <chr> "high", "low", "low", "high", "low", "high", "high...
## $ urbanization      <chr> "low", "low", "high", "high", "high", "high", "hig...
## $ income            <dbl> 46686, 51710, 54310, 54916, 57522, 58633, 58875, 5...
```

- Each observation is a state in the U.S.

The Data

```
## Rows: 47
## Columns: 9
## $ state             <chr> "New Mexico", "Maine", "New York", "Illinois", "De...
## $ median_house_inc  <chr> "low", "low", "low", "low", "high", "high", "high"...
## $ share_pop_metro   <dbl> 0.69, 0.54, 0.94, 0.90, 0.90, 1.00, 0.87, 0.86, 0....
## $ hs                <dbl> 83, 90, 85, 86, 87, 85, 89, 90, 81, 91, 89, 89, 87...
## $ hate_crimes       <dbl> 0.295, 0.616, 0.351, 0.195, 0.323, 0.095, 0.833, 0...
## $ trump_support     <chr> "low", "low", "low", "low", "low", "low", "low", "...
## $ unemployment      <chr> "high", "low", "low", "high", "low", "high", "high...
## $ urbanization      <chr> "low", "low", "high", "high", "high", "high", "hig...
## $ income            <dbl> 46686, 51710, 54310, 54916, 57522, 58633, 58875, 5...
```

- Each observation is a state in the U.S.
- We will focus on `hate_crimes` as a response variable

The Data

```
## Rows: 47
## Columns: 9
## $ state             <chr> "New Mexico", "Maine", "New York", "Illinois", "De...
## $ median_house_inc  <chr> "low", "low", "low", "low", "high", "high", "high"...
## $ share_pop_metro   <dbl> 0.69, 0.54, 0.94, 0.90, 0.90, 1.00, 0.87, 0.86, 0....
## $ hs                <dbl> 83, 90, 85, 86, 87, 85, 89, 90, 81, 91, 89, 89, 87...
## $ hate_crimes       <dbl> 0.295, 0.616, 0.351, 0.195, 0.323, 0.095, 0.833, 0...
## $ trump_support     <chr> "low", "low", "low", "low", "low", "low", "low", "...
## $ unemployment      <chr> "high", "low", "low", "high", "low", "high", "high...
## $ urbanization      <chr> "low", "low", "high", "high", "high", "high", "hig...
## $ income            <dbl> 46686, 51710, 54310, 54916, 57522, 58633, 58875, 5...
```

- Each observation is a state in the U.S.
- We will focus on `hate_crimes` as a response variable
- We investigate 3 different categorical explanatory variables:

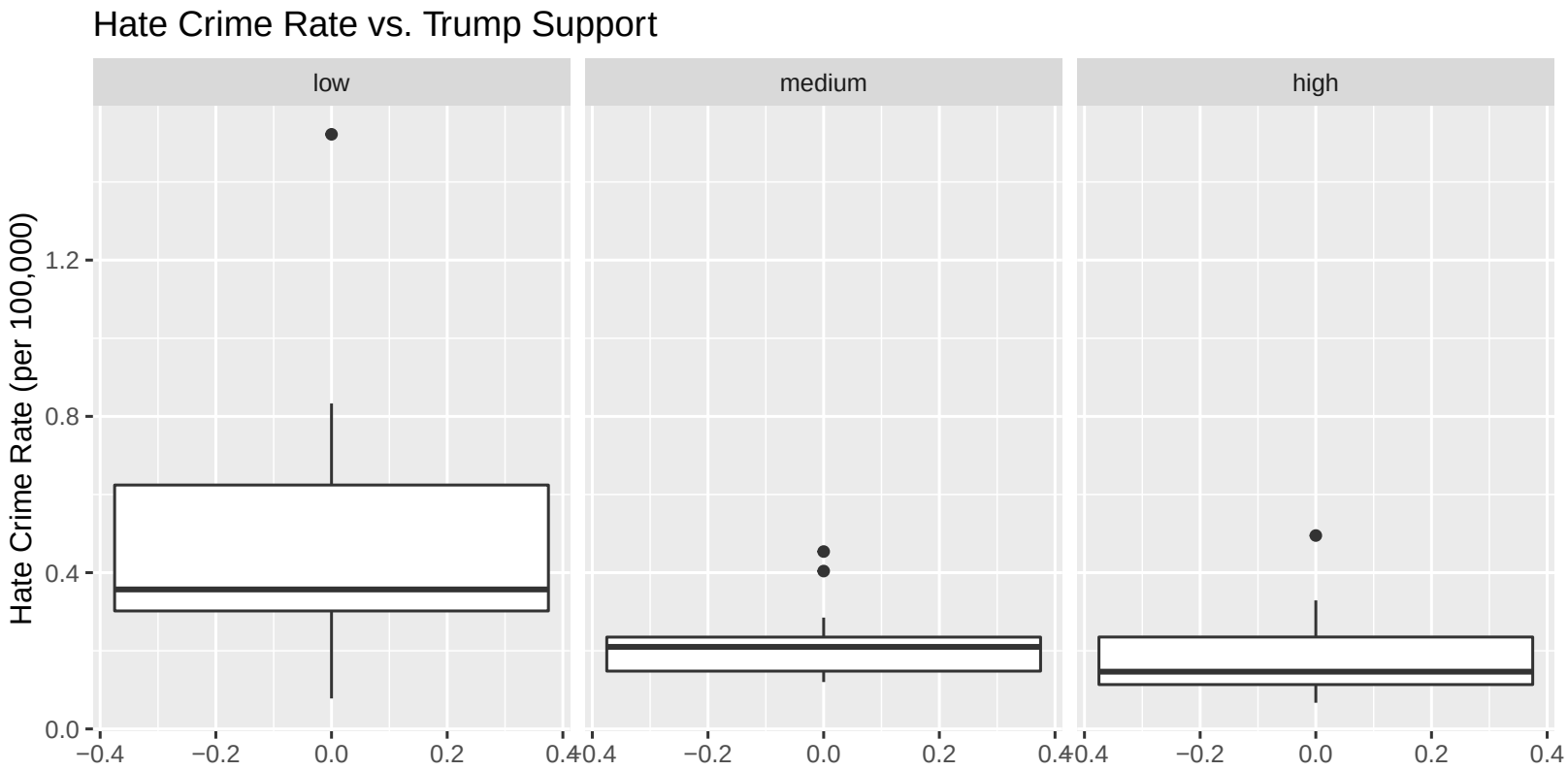
The Data

```
## Rows: 47
## Columns: 9
## $ state             <chr> "New Mexico", "Maine", "New York", "Illinois", "De...
## $ median_house_inc  <chr> "low", "low", "low", "low", "high", "high", "high"...
## $ share_pop_metro   <dbl> 0.69, 0.54, 0.94, 0.90, 0.90, 1.00, 0.87, 0.86, 0....
## $ hs                <dbl> 83, 90, 85, 86, 87, 85, 89, 90, 81, 91, 89, 89, 87...
## $ hate_crimes       <dbl> 0.295, 0.616, 0.351, 0.195, 0.323, 0.095, 0.833, 0...
## $ trump_support     <chr> "low", "low", "low", "low", "low", "low", "low", "...
## $ unemployment      <chr> "high", "low", "low", "high", "low", "high", "high...
## $ urbanization      <chr> "low", "low", "high", "high", "high", "high", "hig...
## $ income            <dbl> 46686, 51710, 54310, 54916, 57522, 58633, 58875, 5...
```

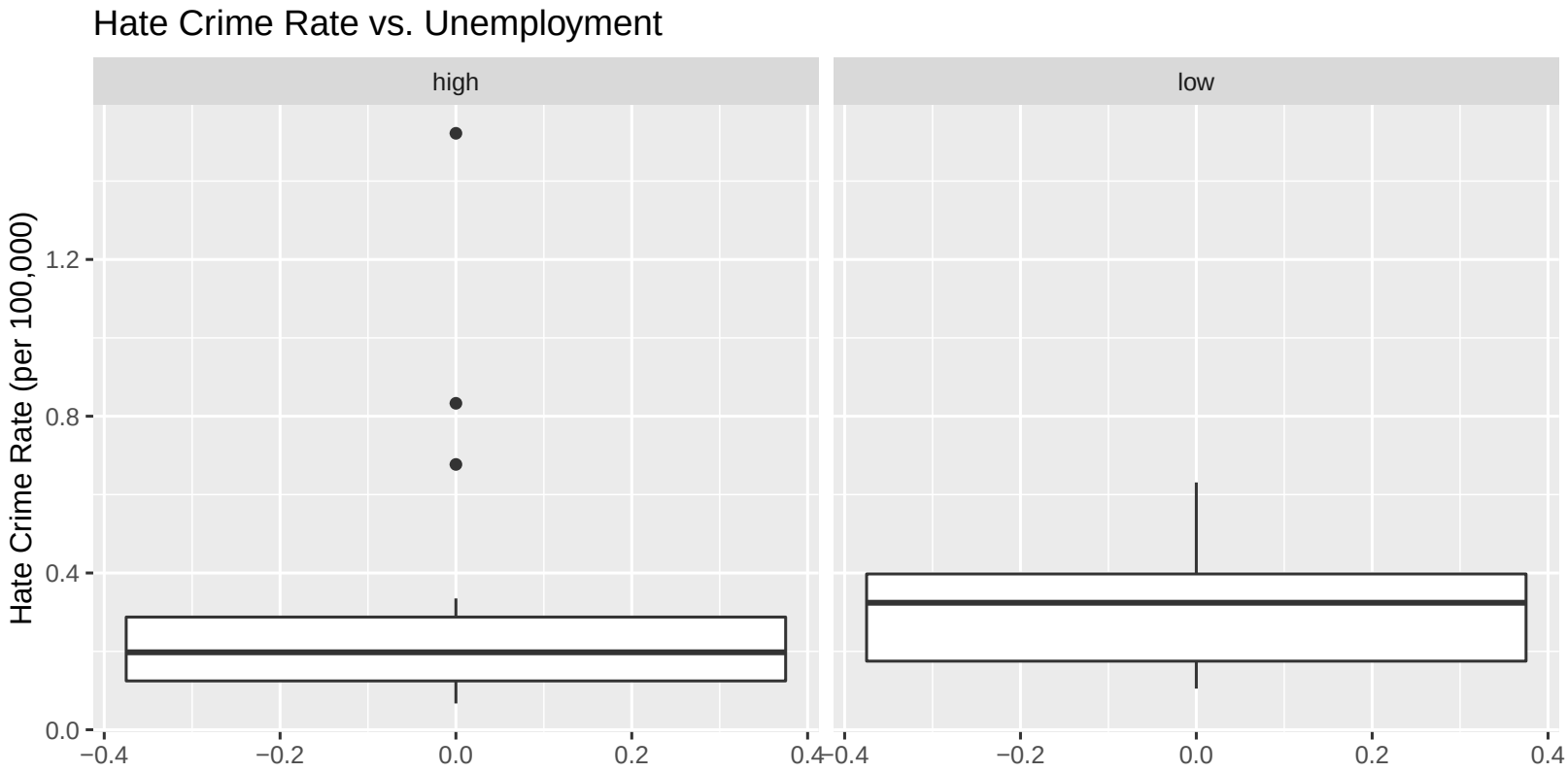
- Each observation is a state in the U.S.
- We will focus on `hate_crimes` as a response variable
- We investigate 3 different categorical explanatory variables:
  1. `trump_support`: level of Trump support in 2016 election (low, medium or high - split into roughly equal number of cases)
  2. `unemployment`: level of unemployment in a state (low or high - split below or above mean)
  3. `median_house_inc`: median household income in the state (low or high - split below or above mean)

Data Exploration (Visualization) Trump

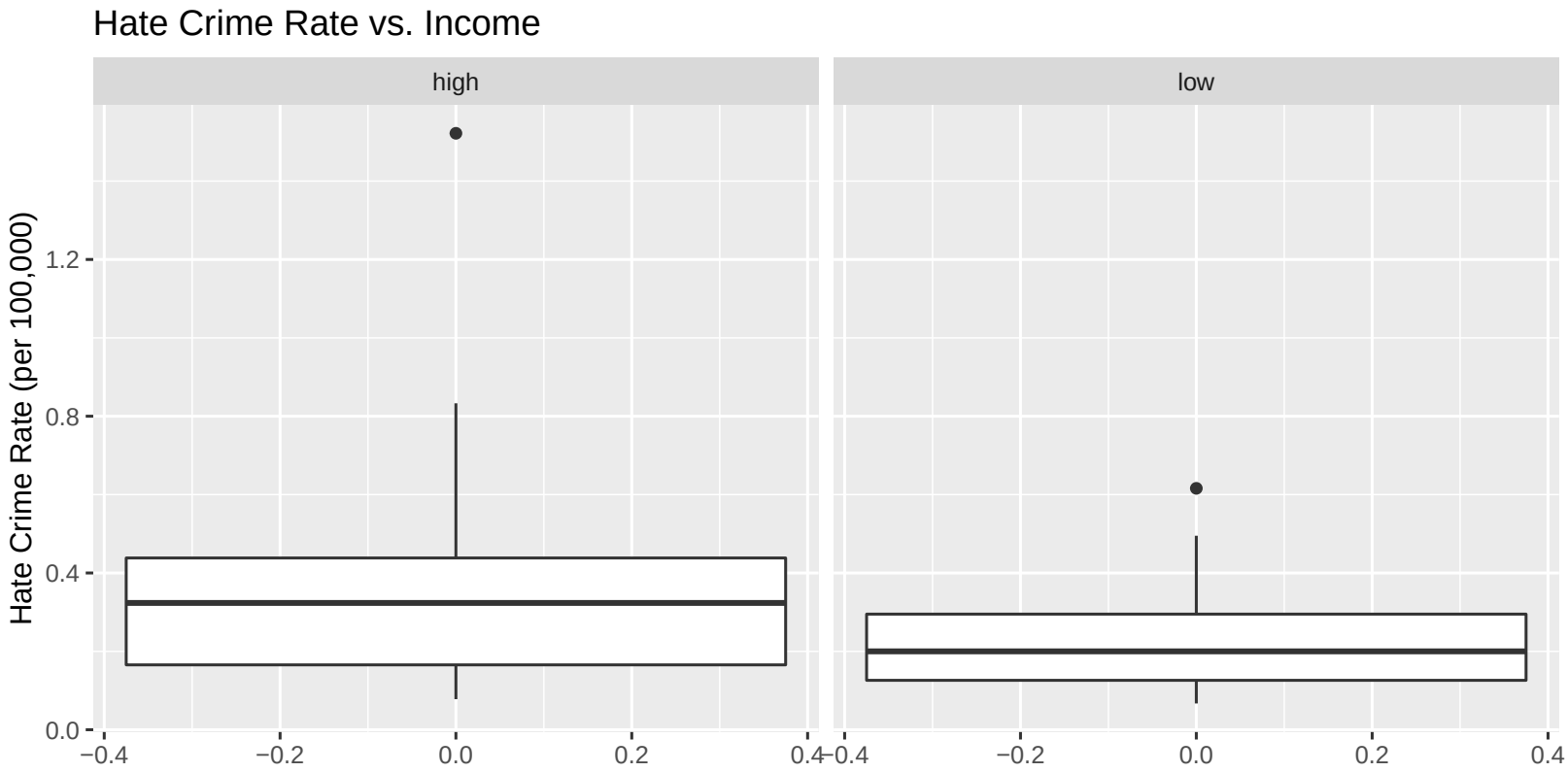
- Create a visual models comparing the response variable to each explanatory variable



Data Exploration (Visualization) Unemployment



Data Exploration (Visualization) Median Income



Data Exploration (Summary Statistics)

- Let's take a closer look at the statistics. . .

Data Exploration (Summary Statistics)

- Let's take a closer look at the statistics...

trump_support	median_hate_crimes	mean_hate_crime
low	0.3570	0.4601111
medium	0.2100	0.2224000
high	0.1465	0.1910000

unemployment	median_hate_crimes	mean_hate_crime
high	0.1975	0.288750
low	0.3240	0.320087

median_house_inc	median_hate_crimes	mean_hate_crime
high	0.3235	0.3894091
low	0.2000	0.2290000

Differences in Mean

- Let's highlight the discrepancies in mean hate crime rate within each variable.

Differences in Mean

- Let's highlight the discrepancies in mean hate crime rate within each variable.
- Using the first entry as a baseline...

Differences in Mean

- Let's highlight the discrepancies in mean hate crime rate within each variable.
- Using the first entry as a baseline. . .

trump_support	mean_hate_crime	difference
low	0.460	0.000
medium	0.222	-0.238
high	0.191	-0.269

unemployment	mean_hate_crime	difference
high	0.289	0.000
low	0.320	0.031

median_house_inc	mean_hate_crime	difference
high	0.389	0.00
low	0.229	-0.16

Linear Regression Trump

- Let's boldly try to fit a linear model to the data, as if we were studying pairs of quantitative variables

Linear Regression Trump

- Let's boldly try to fit a linear model to the data, as if we were studying pairs of quantitative variables

```
hc_trump_model <- lm(hate_crimes ~ trump_support, data = hate_crimes1)
get_regression_table(hc_trump_model)
```

Linear Regression Trump

- Let's boldly try to fit a linear model to the data, as if we were studying pairs of quantitative variables

```
hc_trump_model <- lm(hate_crimes ~ trump_support, data = hate_crimes1)
get_regression_table(hc_trump_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.460	0.053	8.692	0.000	0.353	0.567
trump_supportmedium	-0.238	0.079	-3.028	0.004	-0.396	-0.079
trump_supporthigh	-0.269	0.080	-3.363	0.002	-0.430	-0.108

Linear Regression Trump

- Let's boldly try to fit a linear model to the data, as if we were studying pairs of quantitative variables

```
hc_trump_model <- lm(hate_crimes ~ trump_support, data = hate_crimes1)
get_regression_table(hc_trump_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.460	0.053	8.692	0.000	0.353	0.567
trump_supportmedium	-0.238	0.079	-3.028	0.004	-0.396	-0.079
trump_supporthigh	-0.269	0.080	-3.363	0.002	-0.430	-0.108

- But we've seen some of those numbers before. . .

Linear Regression Trump

- Let's boldly try to fit a linear model to the data, as if we were studying pairs of quantitative variables

```
hc_trump_model <- lm(hate_crimes ~ trump_support, data = hate_crimes1)
get_regression_table(hc_trump_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.460	0.053	8.692	0.000	0.353	0.567
trump_supportmedium	-0.238	0.079	-3.028	0.004	-0.396	-0.079
trump_supporthigh	-0.269	0.080	-3.363	0.002	-0.430	-0.108

- But we've seen some of those numbers before. . .

trump_support	mean_hate_crime	difference
low	0.460	0.000
medium	0.222	-0.238
high	0.191	-0.269

Linear Regression Unemployment

- A linear model for `hate_crimes` as a function of `unemployment`.

Linear Regression Unemployment

- A linear model for `hate_crimes` as a function of `unemployment`.

```
hc_un_model<- lm(hate_crimes ~ unemployment, data = hate_crimes1)
  get_regression_table(hc_un_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.289	0.052	5.549	0.000	0.184	0.394
unemploymentlow	0.031	0.074	0.421	0.676	-0.118	0.181

Linear Regression Unemployment

- A linear model for `hate_crimes` as a function of `unemployment`.

```
hc_un_model<- lm(hate_crimes ~ unemployment, data = hate_crimes1)
  get_regression_table(hc_un_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.289	0.052	5.549	0.000	0.184	0.394
unemploymentlow	0.031	0.074	0.421	0.676	-0.118	0.181

unemployment	mean_hate_crime	difference
high	0.289	0.000
low	0.320	0.031

Linear Regression Income

- Consider the following model for `hate_crimes` as a function of `median_house_inc`

```
hc_in_model<- lm(hate_crimes ~ median_house_inc, data = hate_crimes1)
  get_regression_table(hc_in_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.389	0.052	7.548	0.000	0.285	0.493
median_house_inclow	-0.160	0.071	-2.268	0.028	-0.303	-0.018

Linear Regression Income

- Consider the following model for `hate_crimes` as a function of `median_house_inc`

```
hc_in_model<- lm(hate_crimes ~ median_house_inc, data = hate_crimes1)
  get_regression_table(hc_in_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.389	0.052	7.548	0.000	0.285	0.493
median_house_inclow	-0.160	0.071	-2.268	0.028	-0.303	-0.018

Based on the table. . .

Linear Regression Income

- Consider the following model for `hate_crimes` as a function of `median_house_inc`

```
hc_in_model<- lm(hate_crimes ~ median_house_inc, data = hate_crimes1)
  get_regression_table(hc_in_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.389	0.052	7.548	0.000	0.285	0.493
median_house_inclow	-0.160	0.071	-2.268	0.028	-0.303	-0.018

Based on the table...

- Which level is used as a baseline? What is this level's mean hate crime rate?

Linear Regression Income

- Consider the following model for `hate_crimes` as a function of `median_house_inc`

```
hc_in_model<- lm(hate_crimes ~ median_house_inc, data = hate_crimes1)
  get_regression_table(hc_in_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.389	0.052	7.548	0.000	0.285	0.493
median_house_inclow	-0.160	0.071	-2.268	0.028	-0.303	-0.018

Based on the table...

- Which level is used as a baseline? What is this level's mean hate crime rate?
- What is the difference between the baseline and the other level's mean hate crime rate?

Linear Regression Income

- Consider the following model for `hate_crimes` as a function of `median_house_inc`

```
hc_in_model<- lm(hate_crimes ~ median_house_inc, data = hate_crimes1)
  get_regression_table(hc_in_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.389	0.052	7.548	0.000	0.285	0.493
median_house_inclow	-0.160	0.071	-2.268	0.028	-0.303	-0.018

Based on the table...

- Which level is used as a baseline? What is this level's mean hate crime rate?
- What is the difference between the baseline and the other level's mean hate crime rate?
- What is the other level's mean hate crime rate?

Indicator Functions

- In order to describe the best fitting "line", we need a way of converting categorical values into quantitative values.

Indicator Functions

- In order to describe the best fitting "line", we need a way of converting categorical values into quantitative values.
- Suppose A is a subset of observations. We define an **indicator function** 1_A as the function from the set of observations to the real numbers by

$$1_A(x) = \begin{cases} 1, & \text{if x is in A,} \\ 0, & \text{if x is not in A.} \end{cases}$$

Indicator Functions

- In order to describe the best fitting "line", we need a way of converting categorical values into quantitative values.
- Suppose A is a subset of observations. We define an **indicator function** 1_A as the function from the set of observations to the real numbers by

$$1_A(x) = \begin{cases} 1, & \text{if x is in A,} \\ 0, & \text{if x is not in A.} \end{cases}$$

- For example, if Low denotes the subset of states with low Trump support, then for a state x,

$$1_{\mathrm{Low}}(x) = \begin{cases} 1, & \text{if x has low Trump support,} \\ 0, & \text{if x does not have low Trump support.} \end{cases}$$

Indicator Functions

- In order to describe the best fitting "line", we need a way of converting categorical values into quantitative values.
- Suppose A is a subset of observations. We define an **indicator function** 1_A as the function from the set of observations to the real numbers by

$$1_A(x) = \begin{cases} 1, & \text{if x is in A,} \\ 0, & \text{if x is not in A.} \end{cases}$$

- For example, if Low denotes the subset of states with low Trump support, then for a state x,

$$1_{\mathrm{Low}}(x) = \begin{cases} 1, & \text{if x has low Trump support,} \\ 0, & \text{if x does not have low Trump support.} \end{cases}$$

  - If $x = \mathrm{Oregon}$, then $1_{\mathrm{Low}}(\mathrm{Oregon}) = 1$

Indicator Functions

- In order to describe the best fitting "line", we need a way of converting categorical values into quantitative values.
- Suppose A is a subset of observations. We define an **indicator function** 1_A as the function from the set of observations to the real numbers by

$$1_A(x) = \begin{cases} 1, & \text{if x is in A,} \\ 0, & \text{if x is not in A.} \end{cases}$$

- For example, if Low denotes the subset of states with low Trump support, then for a state x,

$$1_{\mathrm{Low}}(x) = \begin{cases} 1, & \text{if x has low Trump support,} \\ 0, & \text{if x does not have low Trump support.} \end{cases}$$

  - If $x = \mathrm{Oregon}$, then $1_{\mathrm{Low}}(\mathrm{Oregon}) = 1$
  - While if $x = \mathrm{Texas}$, then $1_{\mathrm{Low}}(\mathrm{Texas}) = 0$

The Best Fitting "Line"

- Let $\hat{y}$ denote predicted hate crime rate. The linear regression equation for $\hat{y}$ in terms of trump support is

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_1 \cdot 1_{\text{Medium}}(x) + \beta_2 \cdot 1_{\text{High}}(x) \\ &= 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)\end{aligned}$$

The Best Fitting "Line"

- Let $\hat{y}$ denote predicted hate crime rate. The linear regression equation for $\hat{y}$ in terms of trump support is

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_1 \cdot 1_{\text{Medium}}(x) + \beta_2 \cdot 1_{\text{High}}(x) \\ &= 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)\end{aligned}$$

  which we can read off from the first column of the regression table:

The Best Fitting "Line"

- Let $\hat{y}$ denote predicted hate crime rate. The linear regression equation for $\hat{y}$ in terms of trump support is

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_1 \cdot 1_{\text{Medium}}(x) + \beta_2 \cdot 1_{\text{High}}(x) \\ &= 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)\end{aligned}$$

  which we can read off from the first column of the regression table:

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.460	0.053	8.692	0.000	0.353	0.567
trump_supportmedium	-0.238	0.079	-3.028	0.004	-0.396	-0.079
trump_supporthigh	-0.269	0.080	-3.363	0.002	-0.430	-0.108

- The intercept β_0 corresponds to the hate crime rate for low Trump support

The Best Fitting "Line"

- Let $\hat{y}$ denote predicted hate crime rate. The linear regression equation for $\hat{y}$ in terms of trump support is

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_1 \cdot 1_{\text{Medium}}(x) + \beta_2 \cdot 1_{\text{High}}(x) \\ &= 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)\end{aligned}$$

  which we can read off from the first column of the regression table:

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.460	0.053	8.692	0.000	0.353	0.567
trump_supportmedium	-0.238	0.079	-3.028	0.004	-0.396	-0.079
trump_supporthigh	-0.269	0.080	-3.363	0.002	-0.430	-0.108

- The intercept β_0 corresponds to the hate crime rate for low Trump support
- Each other level's parameter β is the difference between the baseline mean and the mean for that level.

Regression Line Examples

- The regression line predicts that the hate crime rate for a state is equal to the mean rate for that state's Trump support:

$$\hat{y} = 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)$$

Regression Line Examples

- The regression line predicts that the hate crime rate for a state is equal to the mean rate for that state's Trump support:

$$\hat{y} = 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)$$

- We can verify this for a few cases:
  - Oregon has low Trump support, so $1_{\text{Medium}}(\text{Oregon}) = 0$ and $1_{\text{High}}(\text{Oregon}) = 0$, and so

$$\hat{y} = 0.460 - 0.238 \cdot 1_{\text{Medium}}(\text{Oregon}) - 0.269 \cdot 1_{\text{High}}(\text{Oregon}) = 0.46$$

Regression Line Examples

- The regression line predicts that the hate crime rate for a state is equal to the mean rate for that state's Trump support:

$$\hat{y} = 0.460 - 0.238 \cdot 1_{\text{Medium}}(x) - 0.269 \cdot 1_{\text{High}}(x)$$

- We can verify this for a few cases:
  - Oregon has low Trump support, so $1_{\text{Medium}}(\text{Oregon}) = 0$ and $1_{\text{High}}(\text{Oregon}) = 0$, and so
  $$\hat{y} = 0.460 - 0.238 \cdot 1_{\text{Medium}}(\text{Oregon}) - 0.269 \cdot 1_{\text{High}}(\text{Oregon}) = 0.46$$
  - Texas has high Trump support, so $1_{\text{Medium}}(\text{Texas}) = 0$ and $1_{\text{High}}(\text{Texas}) = 1$, and so
  $$\hat{y} = 0.460 - 0.238 \cdot 1_{\text{Medium}}(\text{Texas}) - 0.269 \cdot 1_{\text{High}}(\text{Texas}) = 0.46 - 0.269 = 0.191$$

Residuals

- Of course, we can also get residuals for each observation:

```
regression_points <- get_regression_points(hc_trump_model, ID = "state")
```

Residuals

- Of course, we can also get residuals for each observation:

```
regression_points <- get_regression_points(hc_trump_model, ID = "state")
```

```
## # A tibble: 47 x 5
##    state                hate_crimes trump_support hate_crimes_hat residual
##    <chr>                      <dbl> <fct>                   <dbl>    <dbl>
##  1 Alabama                    0.126 high                    0.191   -0.065
##  2 Alaska                     0.144 medium                  0.222   -0.078
##  3 Arizona                    0.225 medium                  0.222    0.003
##  4 Arkansas                   0.069 high                    0.191   -0.122
##  5 California                 0.256 low                     0.46    -0.204
##  6 Colorado                   0.391 low                     0.46    -0.069
##  7 Connecticut                0.335 low                     0.46    -0.125
##  8 Delaware                   0.323 low                     0.46    -0.137
##  9 District of Columbia       1.52  low                     0.46     1.06
## 10 Florida                    0.188 medium                  0.222   -0.034
## # ... with 37 more rows
```

Residuals

- Of course, we can also get residuals for each observation:

```
regression_points <- get_regression_points(hc_trump_model, ID = "state")
```

```
## # A tibble: 47 x 5
##    state                hate_crimes trump_support hate_crimes_hat residual
##    <chr>                      <dbl> <fct>                   <dbl>    <dbl>
##  1 Alabama                    0.126 high                    0.191   -0.065
##  2 Alaska                     0.144 medium                  0.222   -0.078
##  3 Arizona                    0.225 medium                  0.222    0.003
##  4 Arkansas                   0.069 high                    0.191   -0.122
##  5 California                 0.256 low                     0.46    -0.204
##  6 Colorado                   0.391 low                     0.46    -0.069
##  7 Connecticut                0.335 low                     0.46    -0.125
##  8 Delaware                   0.323 low                     0.46    -0.137
##  9 District of Columbia       1.52  low                     0.46     1.06
## 10 Florida                    0.188 medium                  0.222   -0.034
## # ... with 37 more rows
```

- Recall, residuals tell us the difference between the observed and predicted values

Residuals

- Of course, we can also get residuals for each observation:

```
regression_points <- get_regression_points(hc_trump_model, ID = "state")
```

```
## # A tibble: 47 x 5
##    state                hate_crimes trump_support hate_crimes_hat residual
##    <chr>                      <dbl> <fct>                   <dbl>    <dbl>
##  1 Alabama                    0.126 high                    0.191   -0.065
##  2 Alaska                     0.144 medium                  0.222   -0.078
##  3 Arizona                    0.225 medium                  0.222    0.003
##  4 Arkansas                   0.069 high                    0.191   -0.122
##  5 California                 0.256 low                     0.46    -0.204
##  6 Colorado                   0.391 low                     0.46    -0.069
##  7 Connecticut                0.335 low                     0.46    -0.125
##  8 Delaware                   0.323 low                     0.46    -0.137
##  9 District of Columbia       1.52  low                     0.46     1.06
## 10 Florida                    0.188 medium                  0.222   -0.034
## # ... with 37 more rows
```

- Recall, residuals tell us the difference between the observed and predicted values
- In this case, the residual is the difference between the actual hate crime rate, and the mean hate crime rate for that state's level of Trump support.