Making informed predictions

Intro to Regression and Linear Prediction

Data $\rightarrow$ **Model** $\rightarrow$ Predicted Outcome

Intro to Regression and Linear Prediction

Data $\rightarrow$ **Model** $\rightarrow$ Predicted Outcome

Consider polls. We know, or can approximate, the underlying processes by which data (opinion) are translated into an outcome (electoral winner)

Intro to Regression and Linear Prediction

Data $\rightarrow$ **Model** $\rightarrow$ Predicted Outcome

Consider polls. We know, or can approximate, the underlying processes by which data (opinion) are translated into an outcome (electoral winner)

We need to predict votes based on what we know (public opinion data)

Prediction

Data $\longrightarrow$ **Model** $\longrightarrow$ Predicted Outcome

Prediction

Data $\rightarrow$ **Model** $\rightarrow$ Predicted Outcome

Independent Variable $\rightarrow$ **f(x)** $\rightarrow$ Dependent Variable

Prediction

Data $\rightarrow$ **Model** $\rightarrow$ Predicted Outcome

Independent Variable $\rightarrow$ $\mathbf{f(x)}$ $\rightarrow$ Dependent Variable

$X \rightarrow \mathbf{f(X)} \rightarrow Y$

Prediction

Data $\rightarrow$ **Model** $\rightarrow$ Predicted Outcome

Independent Variable $\rightarrow$ **f(x)** $\rightarrow$ Dependent Variable

X $\rightarrow$ **f(X)** $\rightarrow$ Y

We assume that the dependent variable (Y) is a function of (or depends upon) the value of the independent variable (X)

Prediction

$$\text{Data} \rightarrow \textbf{Model} \rightarrow \text{Predicted Outcome}$$

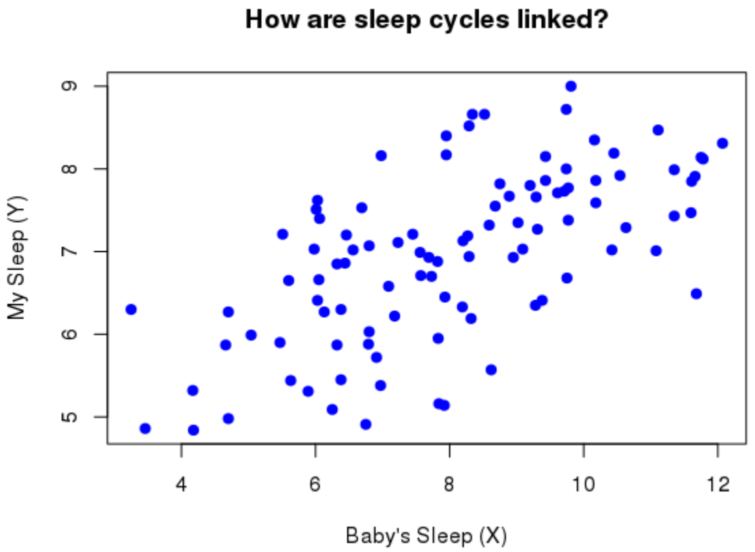
$$\text{Independent Variable} \rightarrow \mathbf{f(x)} \rightarrow \text{Dependent Variable}$$

$$X \rightarrow \mathbf{f(X)} \rightarrow Y$$

We assume that the dependent variable (Y) is a function of (or depends upon) the value of the independent variable (X)

We can rewrite this relationship, by convention, as: $\mathbf{Y = f(X)}$

Prediction (using Jeremy's life choice)



Prediction

$$Y = \quad f(X)$$

Prediction

$$\text{Y} = \text{f}(\text{X})$$

$$\text{My Sleep} = \text{f}(\text{Baby's Sleep})$$

Prediction

$$\text{Y} = \text{f}(\text{X})$$

$$\text{My Sleep} = \text{f}(\text{Baby's Sleep})$$

If we can define the relationship *f()*, we will be able to translate every possible value of Baby's Sleep into a predicted value for My Sleep.

Prediction

$$Y = f(X)$$

$$\text{My Sleep} = f(\text{Baby's Sleep})$$

If we can define the relationship *f()*, we will be able to translate every possible value of Baby's Sleep into a predicted value for My Sleep.

$$? = f(7 \text{ hours})$$

Prediction

$$\text{Y} = \text{f(X)}$$

$$\text{My Sleep} = \text{f(Baby's Sleep)}$$

If we can define the relationship *f()*, we will be able to translate every possible value of Baby's Sleep into a predicted value for My Sleep.

$? = \text{f(7 hours)}$

$? = \text{f(3.5 hours)}$

Prediction

$$\text{Y} = \quad \text{f}(\text{X})$$

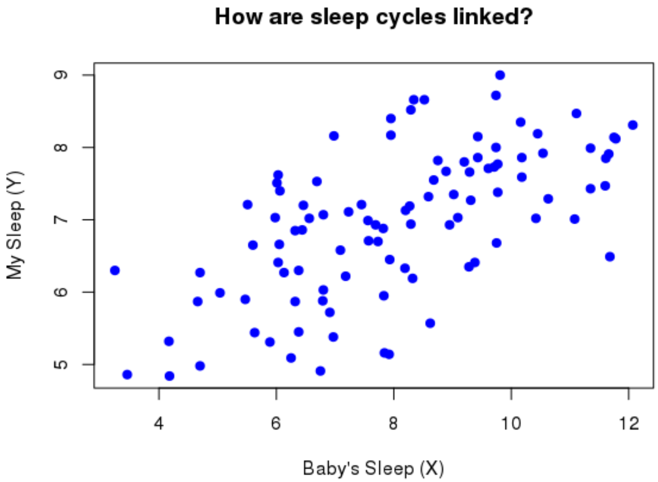
$$\text{My Sleep} = \text{f}(\text{Baby's Sleep})$$

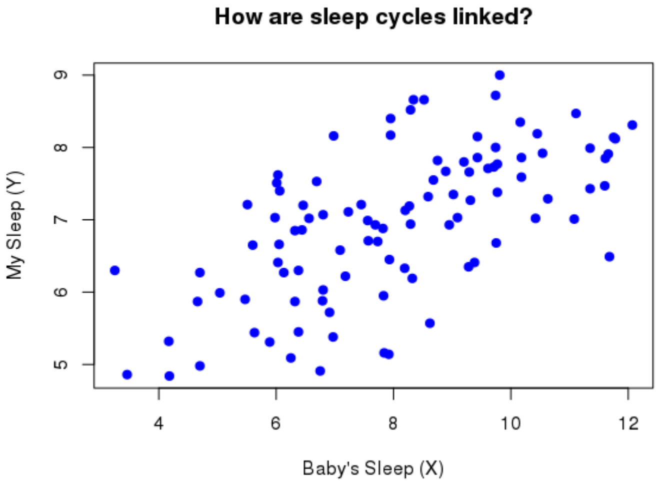
If we can define the relationship *f()*, we will be able to translate every possible value of Baby's Sleep into a predicted value for My Sleep.

$? = \text{f}(7 \text{ hours})$

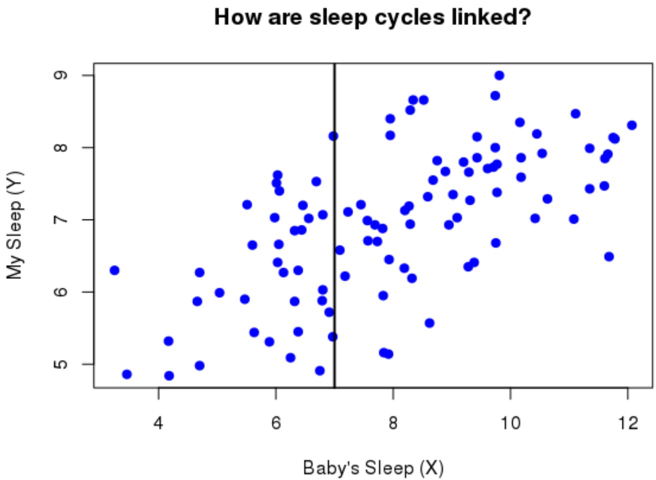
$? = \text{f}(3.5 \text{ hours})$

$? = \text{f}(8 \text{ hours})$

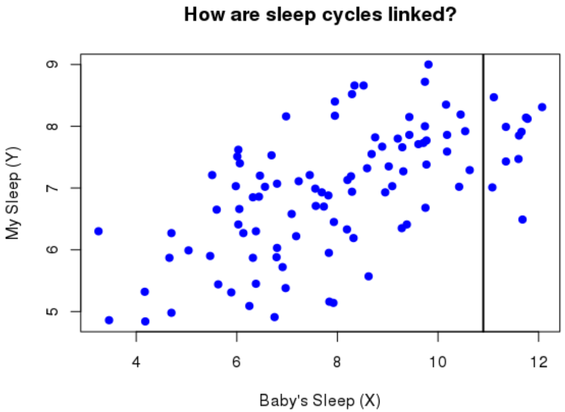

How are sleep cycles linked?
My Sleep (Y)
9
8
7
6
5
4
6
8
10
12
Baby's Sleep (X)



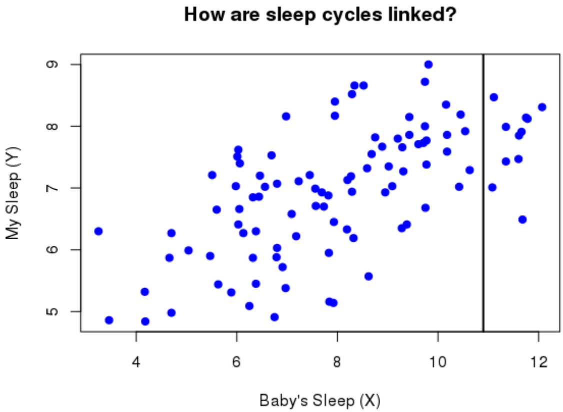
What is my predicted level of sleep if Baby's Sleep = 7?



What is my predicted level of sleep if Baby's Sleep = 7?

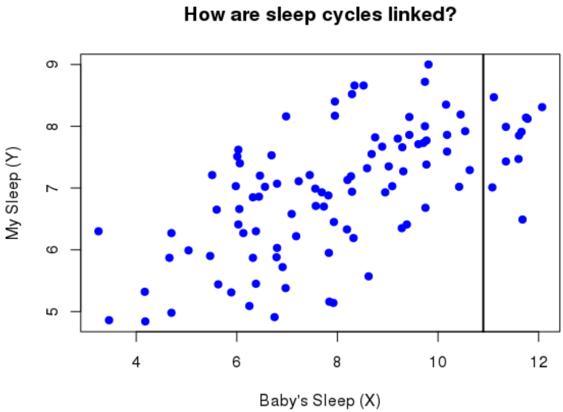


What is my predicted level of sleep if Baby's Sleep = 10.9?



What is my predicted level of sleep if Baby's Sleep = 10.9?

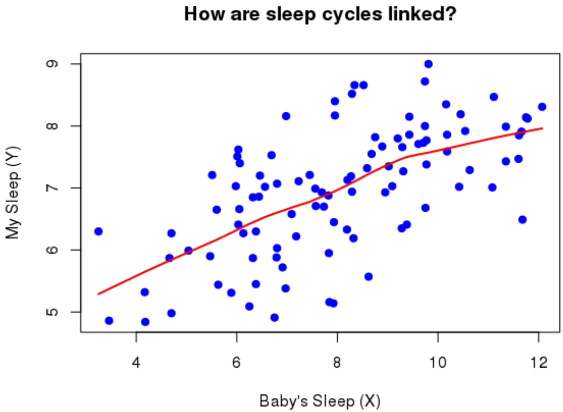
Option 1: We could divide the space into bins, and take the average in each bin



What is my predicted level of sleep if Baby's Sleep = 10.9?

Option 1: We could divide the space into bins, and take the average in each bin

Problem: Bins are arbitrary

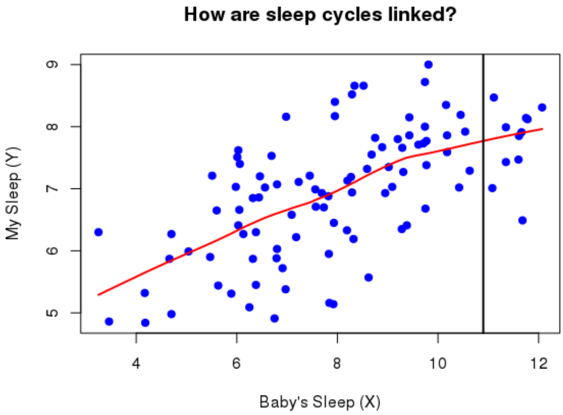


Option 2: Use a moving average

Where we smooth over time

Problem: windows are arbitrary

But with a moving average we can make predictions for any value in the range of our data



A better option

Linear Prediction

Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

Linear Prediction

Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

- We have used averages to calculate an *underlying linear relationship*

Linear Prediction

Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

- We have used averages to calculate an *underlying linear relationship*
- We call this approach **non-parametric estimation**

Linear Prediction

Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

- We have used averages to calculate an *underlying linear relationship*
- We call this approach **non-parametric estimation**
- This can be used to predict the value of Y for possible values of X

Linear Prediction

Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

- We have used averages to calculate an *underlying linear relationship*
- We call this approach **non-parametric estimation**
- This can be used to predict the value of Y for possible values of X

Two remaining issues:

Linear Prediction

Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

- We have used averages to calculate an *underlying linear relationship*
- We call this approach **non-parametric estimation**
- This can be used to predict the value of Y for possible values of X

Two remaining issues:

1. Interpretability

Linear Prediction

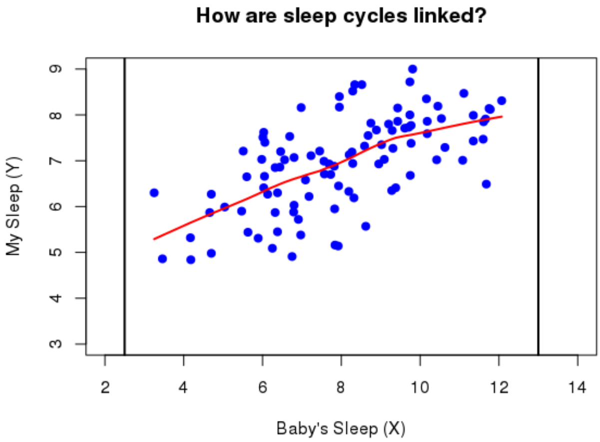
Use existing data we have on the relationship between our independent variable (X) and our dependent variable (Y):

- We have used averages to calculate an *underlying linear relationship*
- We call this approach **non-parametric estimation**
- This can be used to predict the value of Y for possible values of X

Two remaining issues:

1. Interpretability
2. Out of sample prediction

Out of sample prediction



f(X) is undefined outside of our sample. To predict, we need a model that allows for any feasible value of X.

Regression

First we need to make some assumptions:

Regression

First we need to make some assumptions:

Let's assume f(X) is linear (*in other words, the relationship between X and Y is constant for every unique value of X*)

Regression

First we need to make some assumptions:

Let's assume f(X) is linear (*in other words, the relationship between X and Y is constant for every unique value of X*)

$$Y = f(X)$$

Regression

First we need to make some assumptions:

Let's assume f(X) is linear (*in other words, the relationship between X and Y is constant for every unique value of X*)

$$Y = f(X)$$

This means we can use some helpful algebra:

Regression

First we need to make some assumptions:

Let's assume f(X) is linear (*in other words, the relationship between X and Y is constant for every unique value of X*)

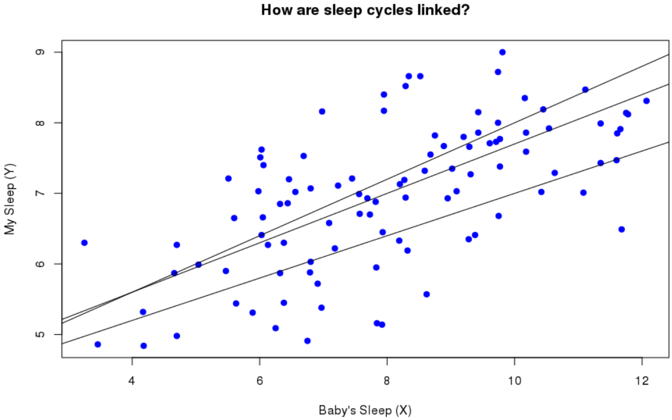
$$Y = f(X)$$

This means we can use some helpful algebra:

$$Y = mX + b \qquad (\text{slope} + \text{intercept})$$

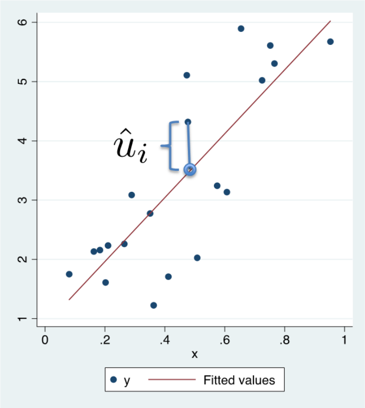
But how do we fit the slope and the intercept?

We have to optimize something



Regression: Ordinary Least Squares (OLS)

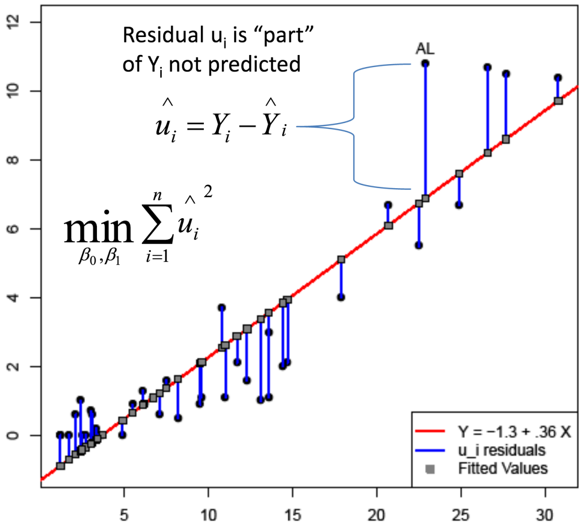
Mathematically, the least biased method selects the line that minimizes the squared differences between observed values of the dependent variable and its predicted values (the regression line)



$\hat{u}_i$ = the difference between the observed and predicted values

The subscript indicates each individual unit of observation (i.e., each country *i*, each person *i*, etc.)

Regression: Ordinary Least Square (OLS)



Note: Similar to RMSE

Regression

Bivariate Regression model components

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

Regression

Bivariate Regression model components

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- $\hat{\alpha}$: an estimated intercept coefficient

Regression

Bivariate Regression model components

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u_i}$$

- $\hat{\alpha}$: an estimated intercept coefficient
- $\hat{\beta}$: an estimated slope coefficient

Regression

Bivariate Regression model components

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- $\hat{\alpha}$: an estimated intercept coefficient
- $\hat{\beta}$: an estimated slope coefficient
- $\hat{u}_i$: an estimated error term, called a "residual"

Interpreting Regression Coefficients

The slope ("beta"):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

Interpreting Regression Coefficients

The slope ("beta"):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- Quantifies the estimated impact of the independent variable on the dependent variable.

Interpreting Regression Coefficients

The slope ("beta"):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- Quantifies the estimated impact of the independent variable on the dependent variable.
- More formally, it shows how many units y changes as x increases by one unit

Interpreting Regression Coefficients

The slope (“beta”):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- Quantifies the estimated impact of the independent variable on the dependent variable.
- More formally, it shows how many units y changes as x increases by one unit
  - Partial derivative of regression equation with respect to X

Interpreting Regression Coefficients

The intercept ("alpha"):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

Interpreting Regression Coefficients

The intercept (“alpha”):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- Measures the expected value of Y when the independent variable is 0

Interpreting Regression Coefficients

The intercept ("alpha"):

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- Measures the expected value of Y when the independent variable is 0
- Varies depending on the beta coefficient.

Interpreting Regression Coefficients

Residuals:

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

Interpreting Regression Coefficients

Residuals:

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- "What's left over" / "Error"

Interpreting Regression Coefficients

Residuals:

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u_i}$$

- "What's left over" / "Error"
- The difference between the observed value of Y (Y_i) and the predicted value of Y ($\hat{Y}_i$), based on the intercept and slope.

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

Performing regression

1. Obtain dataset with independent and dependent variables.

Performing regression

1. Obtain dataset with independent and dependent variables.
2. Run regression algorithm to minimize sum of squared residuals (RMSE)

Performing regression

1. Obtain dataset with independent and dependent variables.
2. Run regression algorithm to minimize sum of squared residuals (RMSE)
3. Obtain coefficients

Performing regression

1. Obtain dataset with independent and dependent variables.
2. Run regression algorithm to minimize sum of squared residuals (RMSE)
3. Obtain coefficients

For example:

$$MySleep_i = 4.48 + 0.31\ (Baby's\ Sleep)_i$$

Interpreting Regression Coefficients

1. Obtain dataset with independent and dependent variables.
2. Run regression algorithm to minimize sum of squared residuals (RMSE)
3. Obtain coefficients

$$\text{My Sleep}_i = 4.48 + 0.31(\text{Baby's Sleep})_i$$

Intercept: Expected level of my sleep if baby slept 0 hours

Interpreting Regression Coefficients

1. Obtain dataset with independent and dependent variables.
2. Run regression algorithm to minimize sum of squared residuals (RMSE)
3. Obtain coefficients

$$\text{My Sleep}_i = 4.48 + 0.31(\text{Baby's Sleep})_i$$

Intercept: Expected level of my sleep if baby slept 0 hours

$$MySleep_i = 4.48 + 0.31 * 0 = 4.48$$

Interpreting Regression Coefficients

1. Obtain dataset with independent and dependent variables.
2. Run regression algorithm to minimize sum of squared residuals (RMSE)
3. Obtain coefficients

$$\text{My Sleep}_i = 4.48 + 0.31(\text{Baby's Sleep})_i$$

Slope: Expected increase in my sleep (in hours)

<u>Every additional hour of baby sleep is associated with an increase in my sleep of 0.31</u>

Practice: Interpreting Regression Coefficients

Y = Student Exhaustion Scale:

- 0 (Not Exhausted) $\rightarrow$ 100 (Exhausted)

Practice: Interpreting Regression Coefficients

Y = Student Exhaustion Scale:

- 0 (Not Exhausted) $\rightarrow$ 100 (Exhausted)

X = Hours Spent on PSET 4

$$Y = 50 \text{ -}2.2(X)$$

Practice: Interpreting Regression Coefficients

Y = Student Exhaustion Scale:

- 0 (Not Exhausted) $\rightarrow$ 100 (Exhausted)

X = Hours Spent on PSET 4

$$Y = 50 - 2.2(X)$$

$$Y = 52 + 3.5(X)$$

Practice: Interpreting Regression Coefficients

Y = Student Exhaustion Scale:

- 0 (Not Exhausted) $\rightarrow$ 100 (Exhausted)

X = Hours Spent on PSET 4

$$Y = 50 - 2.2(X)$$

$$Y = 52 + 3.5(X)$$

$$Y = -20 + 5(X)$$

Practice: Interpreting Regression Coefficients

Y = 2012 Vote Share for Obama (%)

- 0 to 1 interval; i.e., .15 = 15%

X = % of Registered Voters who are Republican:

- 0 to 1 interval; i.e., .15 = 15%

$$Y = 0.98 - 0.96(X)$$

Practice: Interpreting Regression Coefficients

Y = 2012 Vote Share for Obama (%)

- 0 to 1 interval; i.e., .15 = 15%

X = % of Registered Voters who are Republican:

- 0 to 1 interval; i.e., .15 = 15%

$$Y = 0.98 - 0.96(X)$$

The interpretation of a "1 unit change" depends on how the independent variable is measured

Practice: Interpreting Regression Coefficients

Y = 2008 Vote Share for Ralph Nader (%)

- 0 to 1 interval; i.e., .15 = 15 percent

X = % of Registered Voters who are Republican:

- 0 to 100 interval; ie, 15 = 15 percent

$$Y = 0.13 - 0.002(X)$$

OLS Regression as Linear Prediction

$$My\ Sleep_i = 4.48 + 0.31(Baby's\ Sleep)_i$$

OLS Regression as Linear Prediction

$$My\ Sleep_i = 4.48 + 0.31(Baby's\ Sleep)_i$$

Provides an expected value of the dependent variable, for **every value** of the independent variable (minimizing estimated prediction error)

OLS Regression as Linear Prediction

$$My\ Sleep_i = 4.48 + 0.31(Baby's\ Sleep)_i$$

Provides an expected value of the dependent variable, for **every value** of the independent variable (minimizing estimated prediction error)

Baby's Sleep $= 2$

My Sleep $= 4.48 + 0.31 * 2$

OLS Regression as Linear Prediction

$$My\ Sleep_i = 4.48 + 0.31(Baby's\ Sleep)_i$$

Provides an expected value of the dependent variable, for **every value** of the independent variable (minimizing estimated prediction error)

Baby's Sleep $= 2$

My Sleep $= 4.48 + 0.31 * 2$ My Sleep $= 5.1$

Baby's Sleep $= 14.4$

OLS Regression as Linear Prediction

$$My\ Sleep_i = 4.48 + 0.31(Baby's\ Sleep)_i$$

Provides an expected value of the dependent variable, for **every value** of the independent variable (minimizing estimated prediction error)

Baby's Sleep = 2

My Sleep = 4.48 + 0.31 * 2   My Sleep = 5.1

Baby's Sleep = 14.4

My Sleep = 4.48 + 0.31 * 14.4

OLS Regression as Linear Prediction

$$My\ Sleep_i = 4.48 + 0.31(Baby's\ Sleep)_i$$

Provides an expected value of the dependent variable, for **every value** of the independent variable (minimizing estimated prediction error)

Baby's Sleep $= 2$

My Sleep $= 4.48 + 0.31$ * 2 My Sleep $= 5.1$

Baby's Sleep $= 14.4$

My Sleep $= 4.48 + 0.31$ * 14.4 My Sleep $= 8.9$