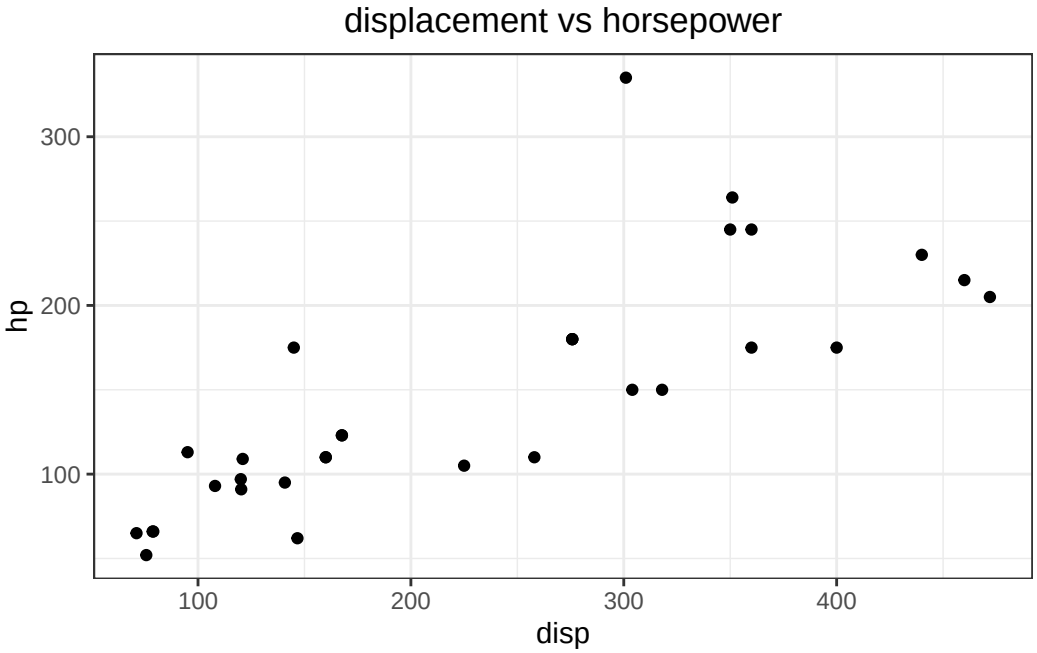
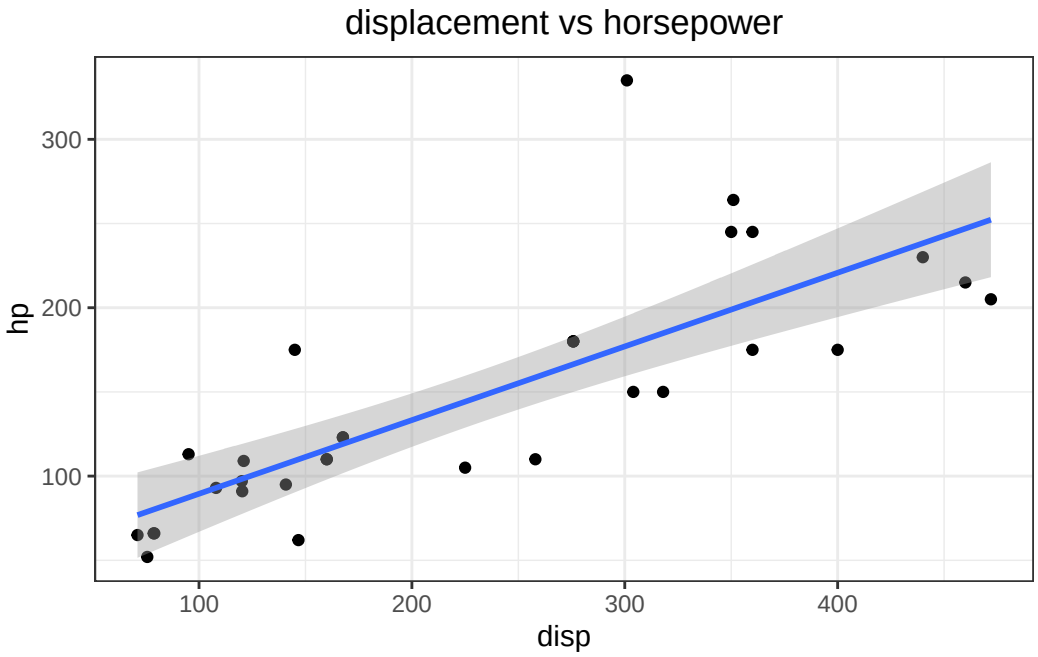
# Lecture 3

**Correlation: Displacement vs Horsepower**

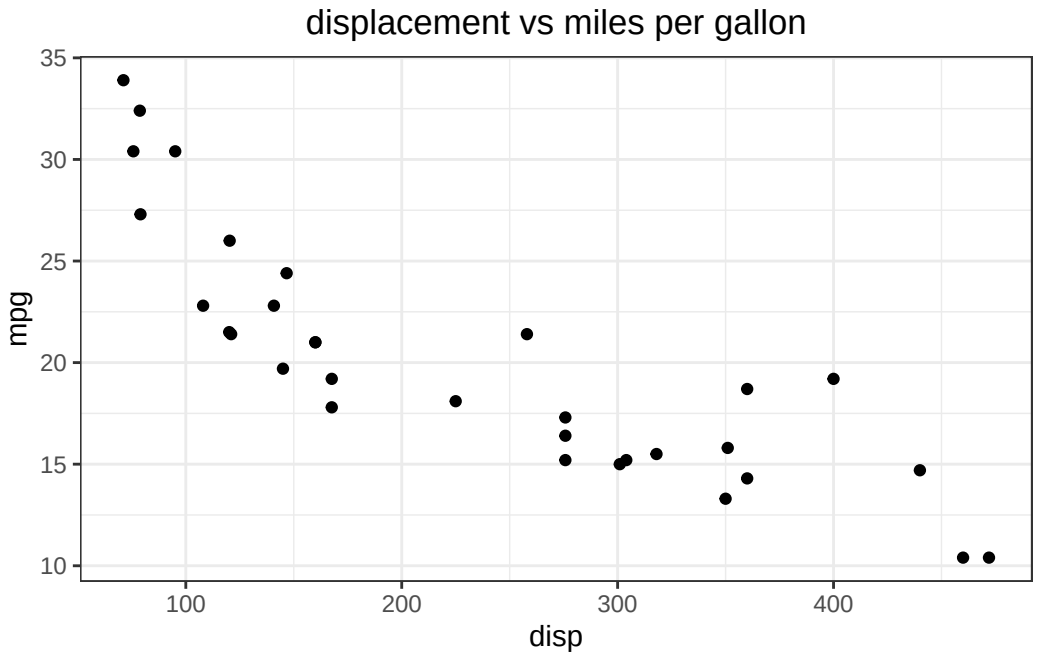


**Correlation: Answer**

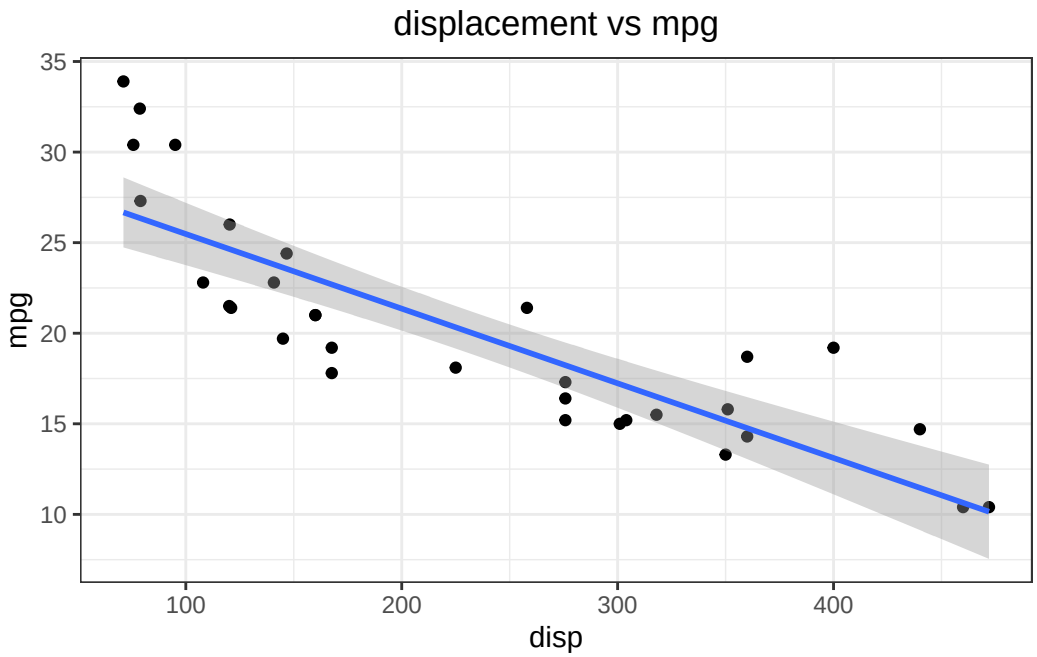



```
[1] 0.7909486
```

**Correlation: Displacement vs Miles per Gallon**

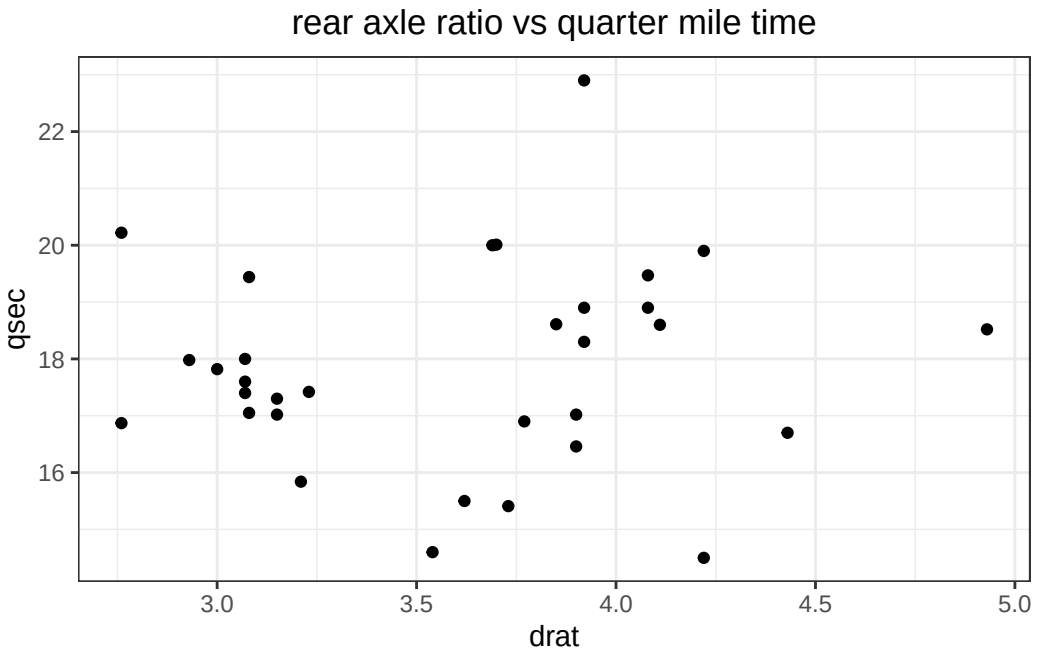



**Correlation: Answer**

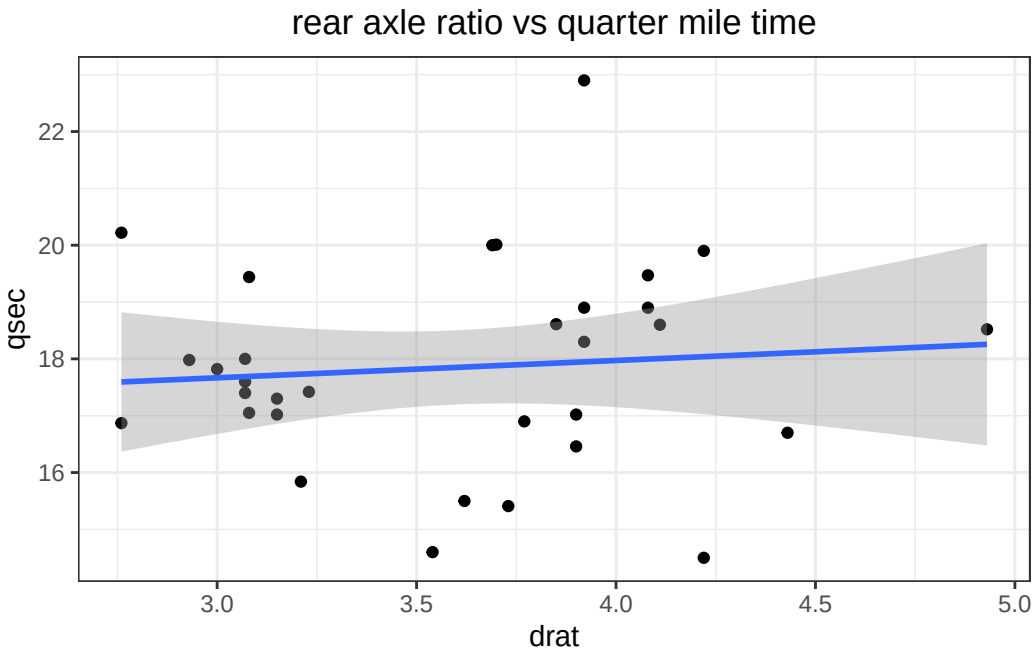


```
[1] -0.8475514
```

**Correlation: Real Axle Ratio vs Quarter Mile Time**

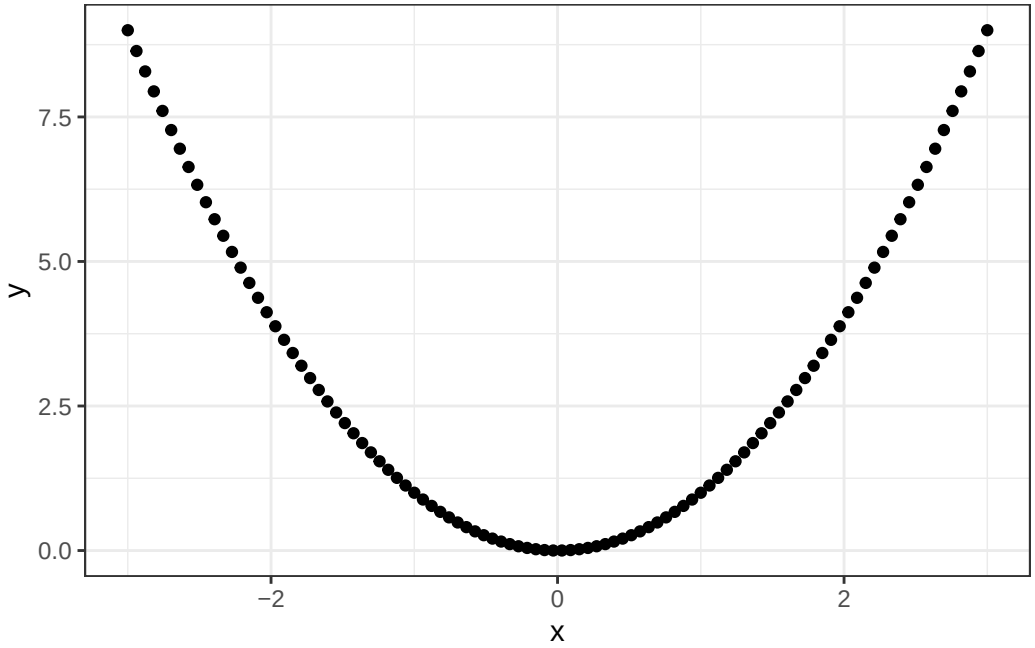


**Correlation: Answer**

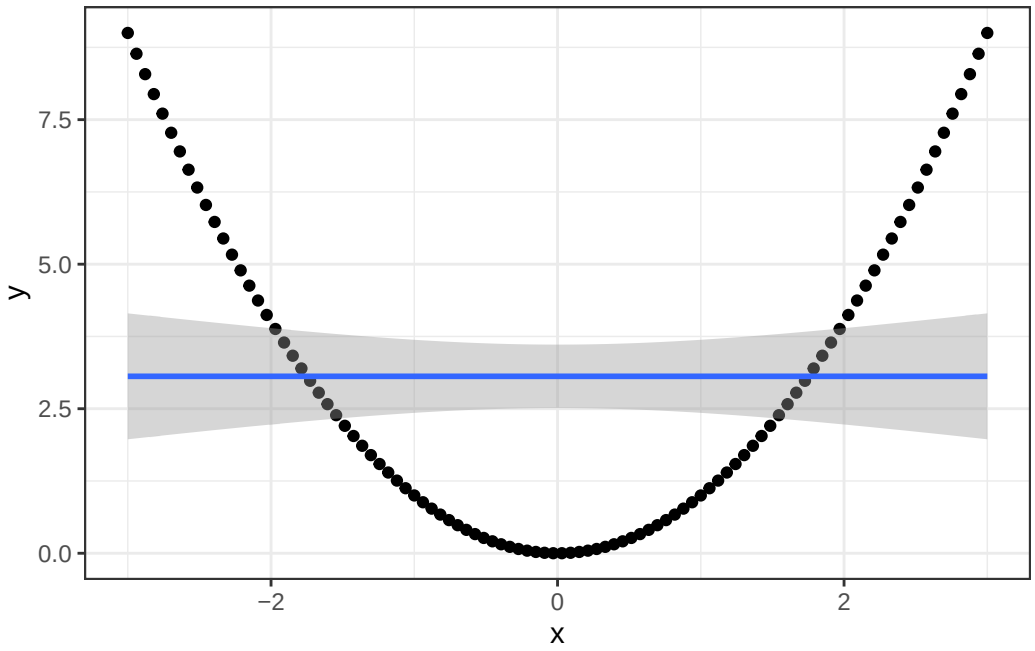



```
[1] 0.09120476
```

**Correlation Example: nonlinear**




**Correlation Example: nonlinear, answer**



```
[1] 0.000000000000000001239961
```

## OLS Regression

- Our goal: to estimate the causal effect of some treatment
  - e.g. how does welfare reform impact poverty rates?
  - Does getting the flu vaccine improve your health?
  - Does increasing teacher salary improve students outcomes?
  - Do cats respond to cat music?
    - This is a real academic paper.

## OLS Regression: Causal Relationships

- Even simple causal models can have very complicated formulas. Think physics calculations
- We need to simplify this model to get anywhere. We assume a linear relationship
- e.g. student grades increase (or decrease) linearly with teacher salary
  - Going from $20,000/year to $21,000/year has the same effect on grades as going from $150,000/year to $151,000/year
  - We relax this assumption later

## A Roadmap of What's to Come

- Start with mechanics of OLS and how to interpret
  - Focus is on describing data
- Once we get to multivariate OLS we can control for issues we know cause bias
- We finish by engineering specific controls to create "quasi-experiments"

## OLS Regression: "The Core Model"

- We arrive at what the book calls the core model, though you will not see this terminology elsewhere
  - It's also called the regression equation or estimating equation
- It is a generic formula that applies to any relationship between x and y:
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

## OLS Regression: Example

- Suppose we use the teacher salary example: how do student's test scores relate to teacher salaries on average?
- We can rename our "core model":
  - $grade_i = \beta_0 + \beta_1 salary_i + \varepsilon_i$
- How do we interpret $\beta_1$? $\beta_0$?

## OLS Regression: iClicker

- $grade_i = \beta_0 + \beta_1 salary_i + \varepsilon_i$
- Salary is in dollars. Grade is in GPA scale (1-4). Interpret $\beta_1$
- A A 1 unit increase in GPA is associated with a $\beta_1$ dollar increase in teacher salary, on average
- B A 1 dollar increase in teacher salary is associated with a $\beta_1$ unit increase in student GPA, on average
- C A $\beta_1$ unit increase in GPA is associated with a 1 dollar increase in teacher salary, on average
- D A $\beta_1$ dollar increase in teacher salary is associated with a 1 unit increase in student GPA, on average

## OLS Regression: "The Core Model"

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - $y_i$ is our outcome (dependent) variable for individual i. Here it's student i's grade
  - $\beta_0$ is the intercept. If we take our model seriously it's the **average** grade for a student whose teacher has 0 salary
  - $x_i$ is the independent variable for individual i. Here it is student i's teacher's salary
  - $\beta_1$ is the slope. This is the actual (average) causal effect of increasing x by 1 unit on y

## OLS Regression: "The Core Model"

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $\varepsilon_i$ is the error term. It captures every factor not included in our model (which is a lot of things!)

  - Examples: student intelligence. Other student characteristics (e.g. demographics). Whether the student is interested in a subject. Whether the student slept in for an exam.
  - It has mean value of 0

## Sources of variation

- Suppose we talk about "the" causal effect of increasing teacher salary on grades. Consider the following ways that teacher salary can increase:
  - A more qualified teacher is hired
  - A bonus is paid based on performance
  - Teachers with low pay are laid off and class sizes increase
  - There is a shortage of teachers, driving up salaries
  - Salaries must be increased due to bad working conditions
  - Hours are increased, so salary is also increased

## What we're measuring

- We're measuring the average association between the two variables in the data
  - We're getting a mix of all of the possible reasons why different teachers have different salarys (and students have different grades)
  - The weightings are based on the population and sample we use
- This is called a reduced form estimate.
- Just describing this data is often useful

## OLS Regression: Estimating the Core Model

- We don't know the true values of $\beta_0, \beta_1$, so we need to estimate
- We end up estimating $y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\varepsilon}_i$
- Hats are used to indicate estimates. We know the actual value of x and y, but not anything else

## OLS Questions

- $grade_i = \beta_0 + \beta_1 salary_i + \varepsilon_i$
  - $salary_i$ is salary (in thousands of dollars), $grade_i$ is final grade in percent (e.g. 100)
- Suppose $\beta_0 = 60, \beta_1 = 1$ how do we interpret this?
- Student 1 has a teacher who is paid \$20000. Calculate $\hat{g}rade_1$
- Student 1's actual grade in class was a 65. What is $\varepsilon_1$?

## OLS Questions

- $grade_i = \beta_0 + \beta_1 salary_i + \varepsilon_i$
  - $salary_i$ is salary (in thousands of dollars), $grade_i$ is final grade in percent (e.g. 100)
  - What is included in $\varepsilon$?
  - Suppose we estimate $\widehat{\beta}_0 = 40, \widehat{\beta}_1 = 2$. What is $\hat{g}rade_1, \hat{\varepsilon}_1$?
  - What is included in $\hat{\varepsilon}$ that is not included in $\varepsilon$?

## OLS Regression: Some Observations

- Core model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Key question 1: If the model is this simplified, is it even useful?
- From a **predictive analytics** perspective this is very weak. But as long as some simple assumptions are satisfied (covered later) this efficiently measures an **average causal effect** .
- This is important for policy evaluation. If a union negotiates a salary increase what will happen to the average grade? This is the causal effect.

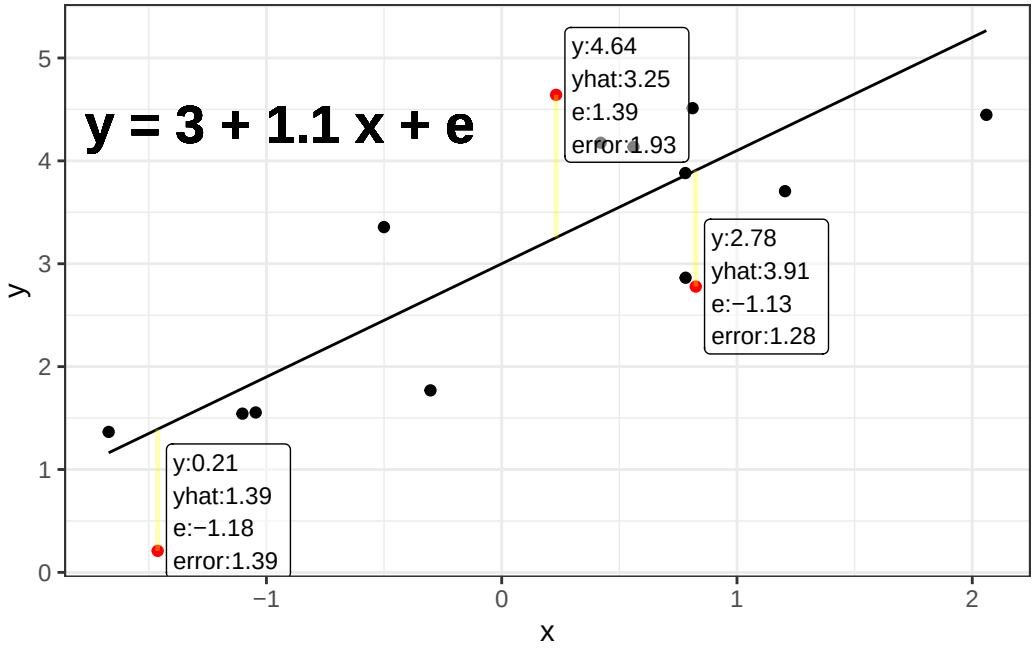
## OLS Regression: Some Observations

- Key Question 2: Given our model, how do we calculate $\beta_1$ (and $\beta_0$)?
- We can never observe $\beta_1$, but we can estimate it as $\widehat{\beta}$ using ordinary least squares regression
- $\widehat{\beta}_1$ is a sample statistic (we'll calculate later)
- We then have to ask if $\widehat{\beta}_1$ is close to $\beta_1$

## OLS regression: ideas

- Our model is a line, and we have data. We estimate $\beta_0, \beta_1$ by finding the best fit line to the observed data
- We can measure the fit using sum of squared errors or mean squared error
- $SSE = \sum \varepsilon_i^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2$, $MSE = SSE/n$

## OLS Regression: Graph




## OLS Regression: Fitting

- Here we have a scatterplot of data, and the line $y = 3 + 1.1x$. For each point we can calculate the error term, then take the average to get the mean square error.
- How do we know what the best fit line is?
- Naive: for every possible $\beta_0, \beta_1$ compute the MSE (or SSE), then choose the parameters that give the lowest value (best fit)
  - This is actually how many machine learning models work, but they use methods from calculus to make it fast