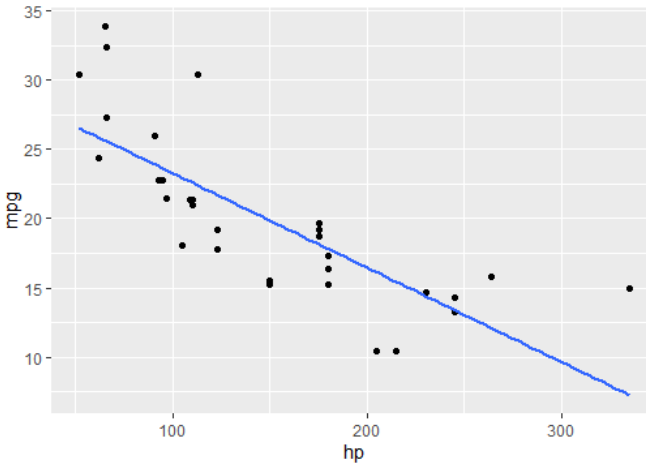
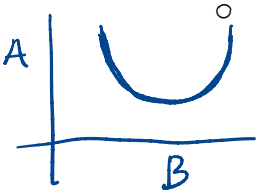
Chapter 8: Linear Regression

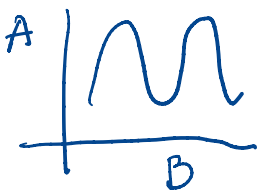
Introduction to Regression

- What is "Regression"?
  - Up until this point, our discussion of inferential investigation has focused on hypothesis testing to decide if we have evidence for an effect/relationship (p-values) as well as quantifying the magnitude of the effect/relationship (effect sizes)
  - Regression is a type of statistical analysis that goes beyond simply determining whether there is some effect, or even how strong the effect is. It is the process of using one (or more) variables to **model the relationship** of another variable.
  - **Statistical Modeling:** Using data to define the general relationship of multiple variables in the form of an equation.
  - While a two-sample test is investigating the relationship between a continuous variable (e.g., height) with a binary variable (e.g., sex) and determining if there is a difference/effect, regression is especially helpful when we have two (or more) numeric variables to model together:
    - Example: How does a car's miles per gallon (mpg) rate compare to its horsepower (hp)?



  - …Or potentially discrete variables as well
    - Example: How does a car's mpg relate to the number of cylinders it has?
  - While a regression model could include categorical variables as possible predictors of our response variable, they usually only do so if complementing other numeric predictor variables, rather than as a lone predictor. We discuss that more in Chapters 14 and 15!
    - Example: Can we predict someone's testosterone levels based on their sex, bmi, and muscle mass?
- What is Simple Linear Regression?
  - **Simple Linear Regression (SLR)** is restricted to comparing no more than two continuous/discrete variables to determine if there is a linear relationship between them.
  - This is in contrast to…
    - Multiple Regression, where we may have two or more predictor variables used to model a response variable.
    - Polynomial Regression, where we may identify quadratic (curved) relationships or even more complex relationships.
    - Logistic Regression, where our response variable is actually a binary variable (think prevalence of having or not having a particular disease or condition).
    - A host of other regression situations that go *well* beyond this course!





- When doing SLR, we're often interested in 1) Determining if there is a linear relationship between these two variables and (if there is indeed a relationship), 2) Modeling their relationship to see how raising the value of one variable might affect the other variable.
- The two variables are denoted as X (Predictor variable) and Y (Response variable) and are plotted against each other on a scatterplot. Sometimes, it doesn't matter much which is which, but in experiments, the predictor (X) variable is what we think *might* be the causal agent.

- An Example! Time Outside vs. Time on Social Media
  - Open up the PhET least squares regression 1.1.4 simulation (try a Google search).
  - Play with the sim for a few minutes.
  - After playing around, think about the following scenario: 15 college students are asked how many hours a week they spend on social media and how many hours a week they spend doing outdoor activities (walking, playing sports, getting some sun, etc.). Use the scatterplot to predict some of the responses you might get.

  - Do you think there is a relationship between these two variables? If so, how would you describe the kind of relationship these two variables have?

  - How strong a relationship do you think they have?

  - Reveal the "correlation coefficient" box. Try moving dots around and see if you can figure out what the correlation coefficient measures. Why does it change? What is the highest and lowest value it can be?

- Defining some terms
  - **Positive Correlation**: As one variable increases, the other variable also increases (directly proportional).
  - **Negative Correlation**: As one variable increases, the other variable decreases (indirectly proportional).
  - **No correlation**: There is no relationship between the variables (for this section, we are specifically looking at linear relationships.
  - **Correlation Coefficient**: Measures the strength of the LINEAR relationship between two variables (from -1 to +1).