Chapter 7: Simple linear regression
Learning Objectives

Reading: Section 7.1 of OpenIntro Statistics

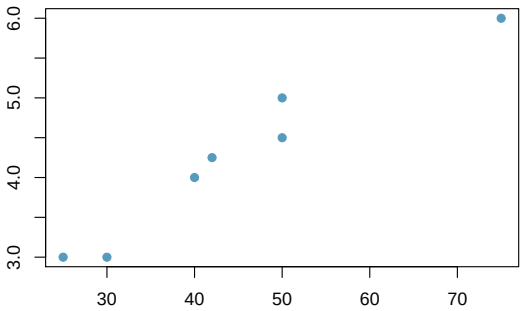
Video: Correlation vs. causation, YouTube (2:19)

Video: Intro to Linear Regression, YouTube (5:18) - very slow introduction to linear regression

1. Define the explanatory variable as the independent variable (predictor), and the response variable as the dependent variable (outcome).
2. Plot the explanatory variable (x) on the x-axis and the response variable (y) on the y-axis, and fit a linear regression model
$$y = \beta_0 + \beta_1 x,$$
where β_0 is the intercept, and β_1 is the slope.
   - Note that the point estimates (estimated from observed data) for β_0 and β_1 are b_0 and b_1, respectively.
3. When describing the association between two numerical variables, evaluate
   - direction: positive ($x \uparrow, y \uparrow$), negative ($x \downarrow, y \uparrow$)
   - form: linear or not
   - strength: determined by the scatter around the underlying relationship
4. Define correlation as the linear association between two numerical variables.
   - Note that a relationship that is nonlinear is simply called an association.
5. Note that correlation coefficient (R, also called Pearson's R) has the following properties:
   - the magnitude (absolute value) of the correlation coefficient measures the strength of the linear association between two numerical variables
   - the sign of the correlation coefficient indicates the direction of association
   - the correlation coefficient is always between -1 and 1, -1 indicating perfect negative linear association, +1 indicating perfect positive linear association, and 0 indicating no linear relationship
   - the correlation coefficient is unitless
   - since the correlation coefficient is unitless, it is not affected by changes in the center or scale of either variable (such as unit conversions)
   - the correlation of X with Y is the same as of Y with X
   - the correlation coefficient is sensitive to outliers
6. Recall that correlation does not imply causation.
7. Define residual (e) as the difference between the observed (y) and predicted ($\hat{y}$) values of the response variable.
$$e_i = y_i - \hat{y}_i$$

Test yourself:

1. Someone hands you the scatter diagram shown below, but has forgotten to label the axes. Can you calculate the correlation coefficient? Or do you need the labels?



2. A teaching assistant gives a quiz. There are 10 questions on the quiz and no partial credit is given. After grading the papers the TA writes down for each student the number of questions the student got right and the number wrong. What is the correlation of the number of questions right and wrong?
Hint: Make up some data for number of questions right, calculate number of questions wrong, and plot them against each other.
3. Suppose you fit a linear regression model predicting score on an exam from number of hours studied. Say you've studied for 4 hours. Would you prefer to be on the line, below the line, or above the line? What would the residual for your score be (0, negative, or positive)?

Reading: Section 7.2 of OpenIntro Statistics

8. Define the least squares line as the line that minimizes the sum of the squared residuals, and list conditions necessary for fitting such a line:

(1) linearity
(2) nearly normal residuals
(3) constant variability

Also be cautious about cases where consecutive observations are evidently related to each other.

9. Define an indicator variable as a binary explanatory variable (with two levels).

10. Calculate the estimate for the slope (b_1) as

$$b_1 = R\frac{s_y}{s_x},$$

where R is the correlation coefficient, s_y is the standard deviation of the response variable, and s_x is the standard deviation of the explanatory variable.

11. Generic interpretation of the slope:

- "For each unit increase in x, we would expect y to increase/decrease on average by $|b_1|$ units" when x is numerical.
- "The average increase/decrease in the response variable when between the baseline level and the other level of the explanatory variable is $|b_1|$." when x is categorical.
- If the data come from an observational study, these interpretations are meant in a way that is not causal.

12. Note that the least squares line always passes through the average of the response and explanatory variables $(\bar{x}, \bar{y})$.

13. Use the above property to calculate the estimate for the slope (b_0) as

$$b_0 = \bar{y} - b_1\bar{x},$$

where b_1 is the slope, $\bar{y}$ is the average of the response variable, and $\bar{x}$ is the average of explanatory variable.

14. Generic interpretation of the intercept:

- "When $x = 0$, we would expect y to equal, on average, b_0." when x is numerical.
- "The expected average value of the response variable for the reference level of the explanatory variable is b_0." when x is categorical.
- In some applications, the interpretation of the intercept will have no practical meaning.

15. Predict the value of the response variable for a given value of the explanatory variable, $x^\star$, by plugging in $x^\star$ in the linear model:

$$\hat{y} = b_0 + b_1 x^\star$$

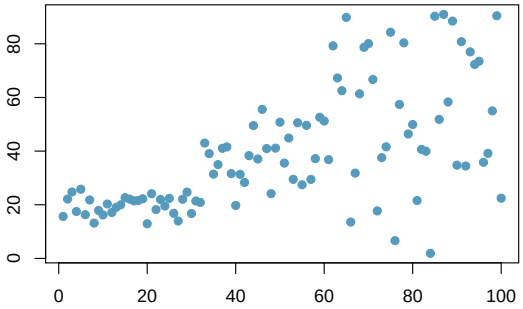
- Only predict for values of $x^\star$ that are in the range of the observed data.
- Do not extrapolate beyond the range of the data, unless you are confident that the linear pattern continues.

16. R^2 is the percentage of the variability in the response variable explained by the the explanatory variable.

- For a good model, we would like this number to be as close to 100% as possible.
- This value is calculated as the square of the correlation coefficient.

Test yourself:

1. We would not want to fit a least squares line to the data shown in the scatterplot below. Which of the conditions does it appear to violate?



2. Derive the formula for b_0 given the fact that the linear model is $\hat{y} = b_0 + b_1 \times x$ and that the least squares line goes through $(\bar{x}, \bar{y})$.
3. One study on male college students found their average height to be 70 inches with a standard deviation of 2 inches. Their average weight was 140 pounds, with a standard deviation of 25 pounds. The correlation between their height and weight was 0.60. Assuming that the two variables are linearly associated, write the linear model for predicting weight from height.
4. Is a male who is 72 inches tall and who weighs 115 pounds on the line, below the line, or above the line?
5. Describe what is an indicator variable, and what levels 0 and 1 mean for such variables.
6. The model below predicts GPA based on an indicator variable (0: not premed, 1: premed). Interpret the intercept and slope estimates in context of the data.

$$\widehat{gpa} = 3.57 - 0.01 \times premed$$

7. If the correlation between two variables y and x is 0.6, what percent of the variability in y does x explain?

Reading: Section 7.3 of OpenIntro Statistics

17. Define a leverage point as a point that lies away from the center of the data in the horizontal direction.

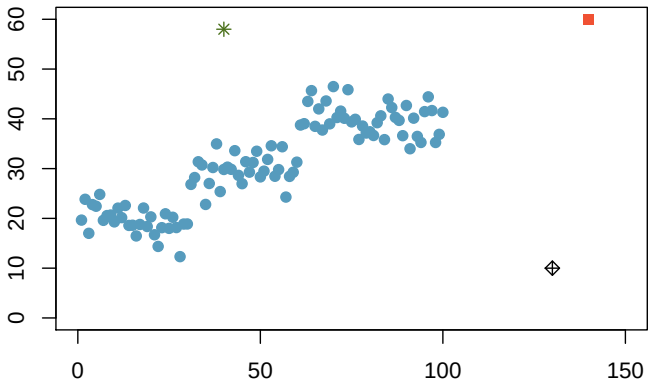
18. Define an influential point as a point that influences (changes) the slope of the regression line much more than the other points.

    - This is usually a leverage point that is away from the trajectory of the rest of the data.

19. Do not remove outliers from an analysis without good reason.

20. Be cautious about using a categorical explanatory variable when one of the levels has very few observations, as these may act as influential points.

    Test yourself: Determine if each of the three unusual observations in the plot below would be considered just an outlier, a leverage point, or an influential point.



Reading: Section 7.4 of OpenIntro Statistics

21. Determine whether an explanatory variable is a significant predictor for the response variable using the t-test and the associated p-value in the regression output.

22. Set the null hypothesis testing for the significance of the predictor as $H_0 : \beta_1 = 0$, and recognize that the standard software output yields the p-value for the two-sided alternative hypothesis.

    - Note that $\beta_1 = 0$ means the regression line is horizontal, hence suggesting that there is no relationship between the explanatory and the response variables.

23. Calculate the T score for the hypothesis test as

$$T_{df} = \frac{b_1 - \text{null value}}{SE_{b_1}}$$

    with $df = n - 2$.

    - Note that the T score has $n-2$ degrees of freedom since we lose one degree of freedom for each parameter we estimate, and in this case we estimate the intercept and the slope.

24. Note that a hypothesis test for the intercept is often irrelevant.

25. Calculate a confidence interval for the slope as

$$b_1 \pm t^\star_{df} SE_{b_1},$$

    where $df = n - 2$ and $t^\star_{df}$ is the critical score associated with the given confidence level at the desired degrees of freedom.

- Note that the standard error of the slope estimate SE_{b_1} can be found on the regression output.

Test yourself:

1. Given the regression output below for predicting y from x where $n = 100$, confirm the T score and the p-value, determine whether x is a significant predictor of y, and interpret the p-value in context.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.8581	6.1379	3.24	0.0017
x	0.2557	0.1055	2.42	0.0172

2. Calculate a 95% confidence interval for the slope given above.