PSTAT 5A: Lecture 20

Post-MT2 Review, Part II

Ethan P. Marzban

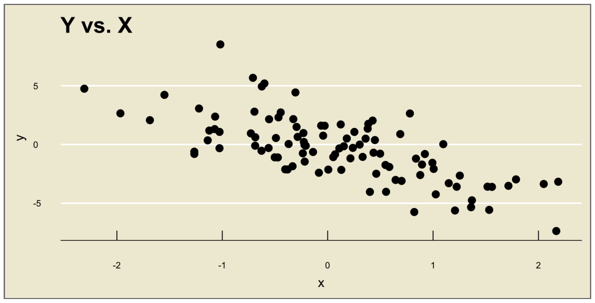
6/12/23

Correlation and Regression

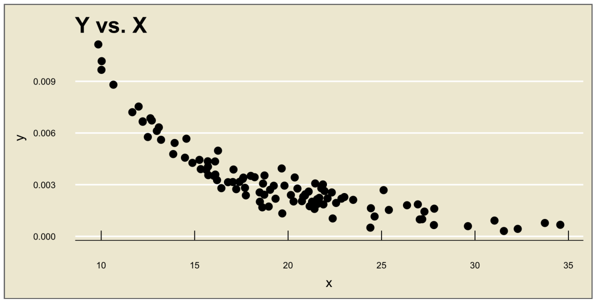
Associations and Correlations

- Recall, from Week 1, that a **scatterplot** is a good way to visualize the relationship between two numerical variables x and y.
- Two variables can have either a **positive** or a **negative** relationship/association, along with a **linear** or **nonlinear** one.
  - “Positive” means a one-unit increase in x translates to an increase in y
  - “Negative” means a one-unit increase in x translates to an degrease in y
  - “Linear” means the rate of change is fixed (i.e. constant)
  - “Nonlinear” means the rate of change depends on x

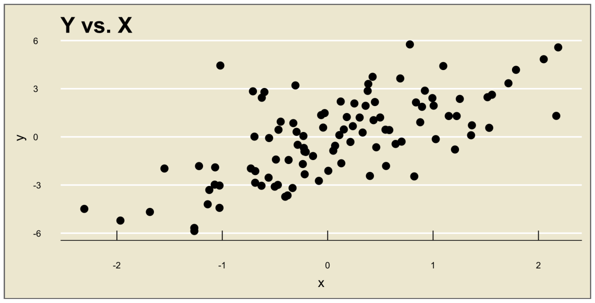
- Linear **Negative** Association:



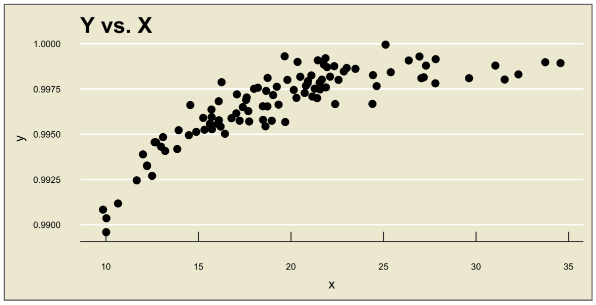
- Nonlinear **Negative** Association:



- Linear **Positive** Association:

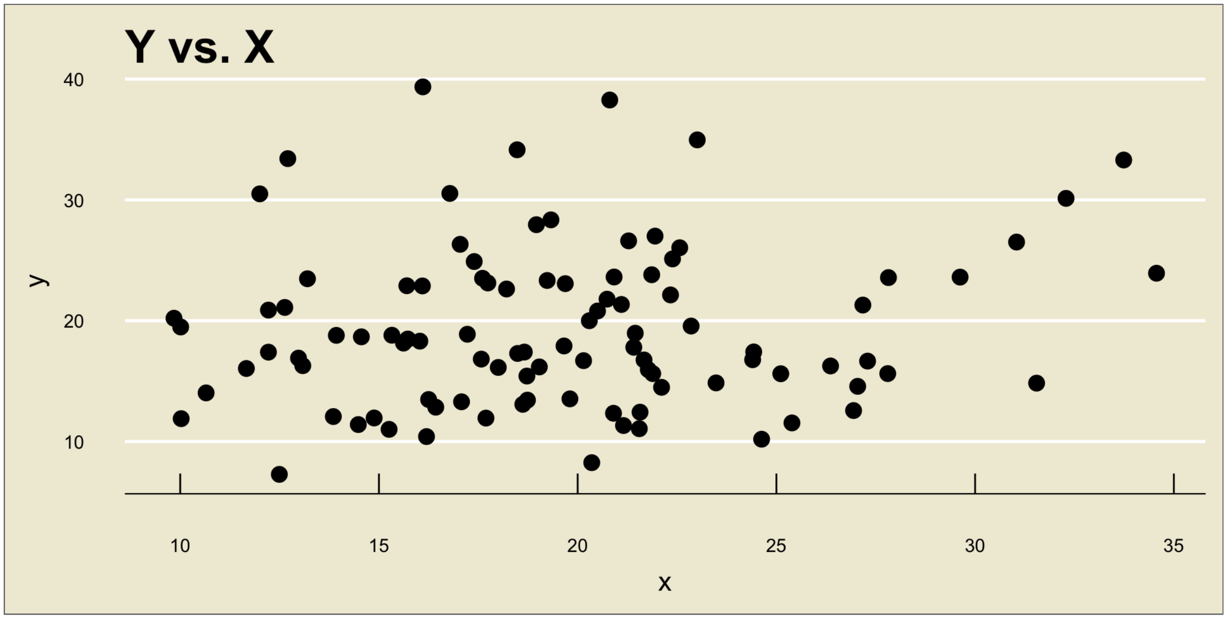


- Nonlinear **Positive** Association:

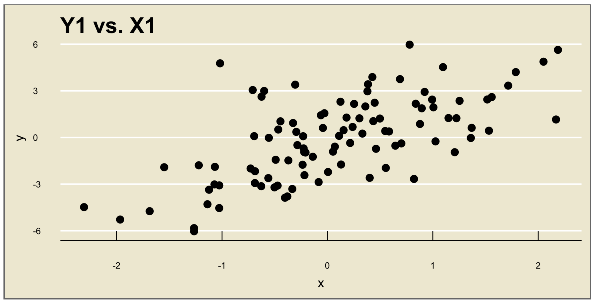


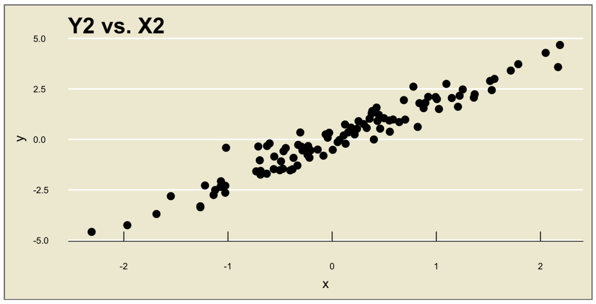
No Relationship

- Sometimes, two variables will have no relationship at all:



Strength of a Relationship





Pearson's *r*

- **Pearson's *r*** (or just the **correlation coefficient**) is a metric used to quantify the strength and direction of a linear relationship between two variables.
- Given variables x and y (whose elements are denoted using the familiar notation we've been using throughout this course), we compute *r* using

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right) \left(\frac{y_i - \overline{y}}{s_Y} \right)$$

- Recall that $-1 \leq r \leq 1$ for any two variables x and y.
  - Furthermore, r will only ever be -1 or 1 exactly when the points in the scatterplot fall perfectly on a line.

Regression

- We may also want to *model* the relationship between x and y.
- Specifically, given a **respone variable** y and an **explanatory variable** x, we typically assume x and y are related through the equation

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{noise}$$

  where f is some function.

  - By the way: on a scatterplot, the response variable will always appear on the vertical axis and the explanatory variable will appear on the horizontal axis.

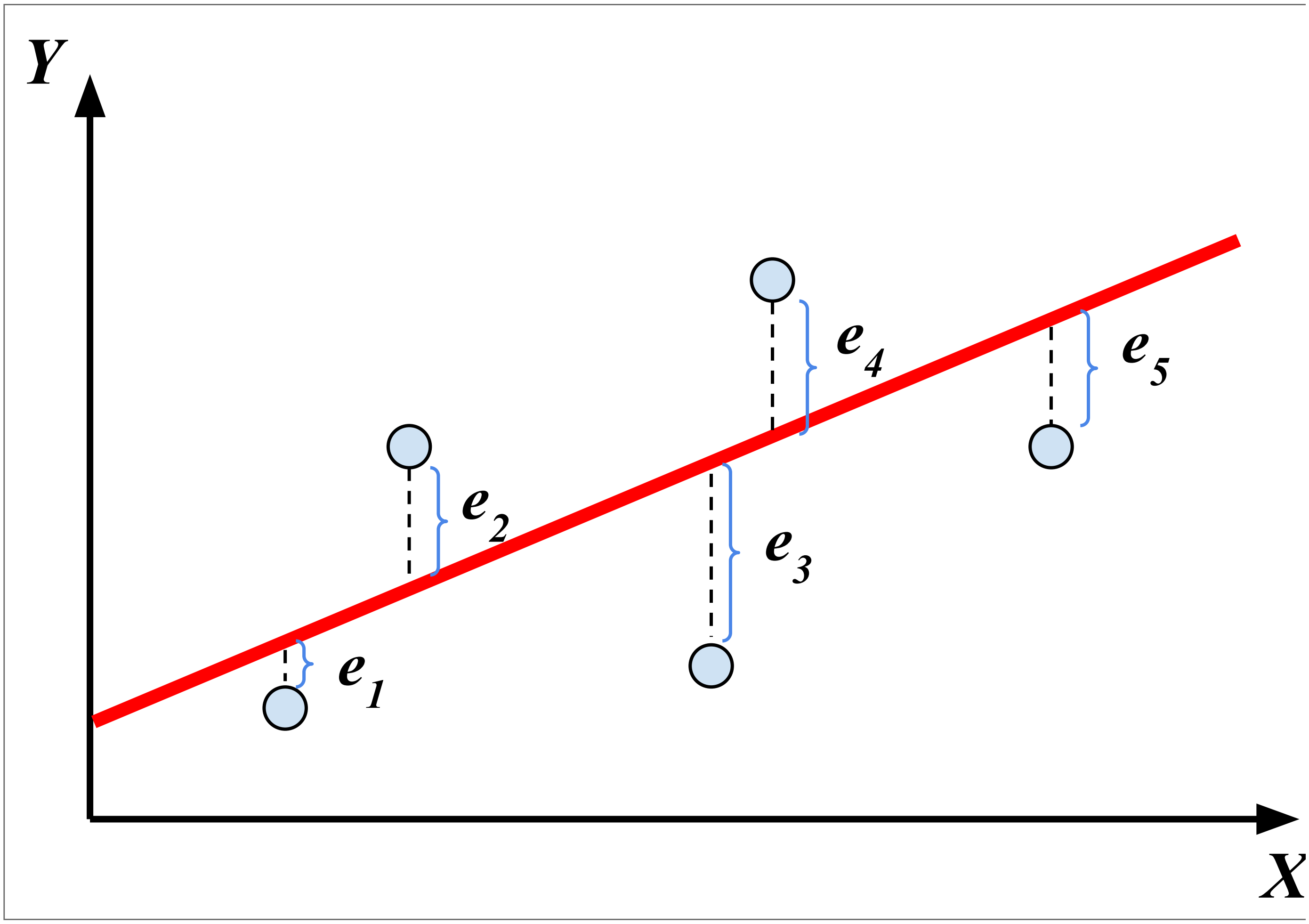
- Of particular interest to us in this class is when f takes the form of a linear equation: i.e. when our model is of the form

$$\mathbf{y} = \beta_0 + \beta_1 \cdot \mathbf{x} + \mathbf{noise}$$

Regression

- Now, the noise part of our model makes it impossible to know the true values of β_0 and β_1.
  - In this way, we can think of them as population parameters.
- As such, we seek to find point estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that best estimate β_0 and β_1, respectively.
- To quantify what we mean by “best”, we employed the condition of minimizing the **residual sum of squares**.
  - Effectively, this means finding the line $\widehat{\beta_0} + \widehat{\beta_1} \cdot \mathbf{x}$ that minimizes the average distance between the points in the dataset and the line.

Regression



Regression

- Such estimators (i.e. those that minimize the RSS) are said to be **ordinary least squares** (OLS) estimates.
  - The resulting line $\widehat{\beta_0} + \widehat{\beta_1} \cdot \mathbf{x}$ is thus called the **OLS Regression Line**
- It turns out that the OLS estimates of β_0 and β_1 are:

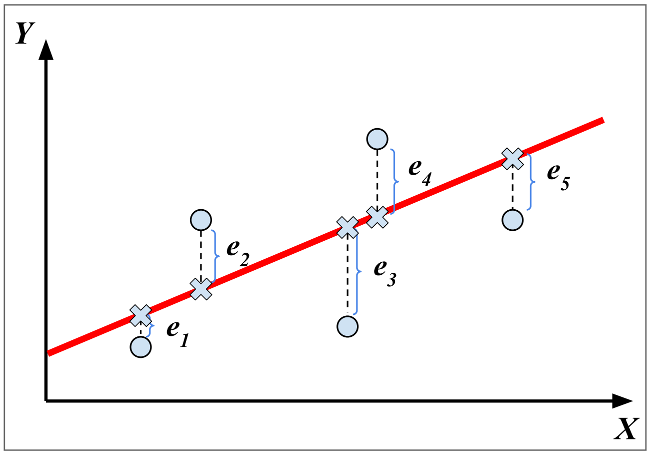
$$\begin{aligned}
\widehat{\beta_1} &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x - \overline{x})^2} = \frac{s_Y}{s_X} \cdot r \\
\widehat{\beta_0} &= \overline{y} - \widehat{\beta_1} \cdot \overline{x}
\end{aligned}$$

where *r* denotes **Pearson's Correlation Coefficient**

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right) \left(\frac{y_i - \overline{y}}{s_Y} \right)$$

Regression

- The values along the OLS regression line corresponding to x values observed in the dataset are called **fitted values**:



- In a sense, the fitted values represent guess/estimate of the *de-noised* value of y

Regression

- We can use the OLS regression line to perform **prediction**; i.e. to infer response values associated with explanatory values that were not included in the original dataset.

Example

Worked-Out Example 3 (formerly a “Your Turn” Exercise)

An airline is interested in determining the relationship between flight duration (in minutes) and the net amount of soda consumed (in oz.). Letting `x` denote `flight duration` (the explanatory variable) and `y` denote `amount of soda consumed` (the response variable), a sample of size 100 yielded the following results:

$$\sum_{i=1}^{100} x_i = 10{,}211.7; \qquad \sum_{i=1}^{100} (x_i - \overline{x})^2 = 38{,}760.68$$

$$\sum_{i=1}^{100} y_i = 14{,}3995.8; \qquad \sum_{i=1}^{100} (y_i - \overline{y})^2 = 87.23984$$

$$\sum_{i=1}^{100} (x_i - \overline{x})(y_i - \overline{y}) = 379.945$$

a. Find the equation of the OLS Regression line.

b. If a particular flight has a duration of 110 minutes, how many ounces of soda would we expect to be consumed on the flight?

Solutions

a.

$$\begin{aligned}\widehat{\beta_0} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{379.945}{38{,}760.68} \approx \boxed{0.0098} \\ \widehat{\beta_0} &= \overline{y} - \widehat{\beta_1}\cdot\overline{x} = \boxed{1438.957}\end{aligned}$$

Therefore,

$$(\widehat{\texttt{\textbf{amt. of soda}}}) = 1436.159 + (0.0098)\cdot(\texttt{\textbf{flight duration}})$$

b. $\hat{y}^{(110)} = 1438.957 + (0.0098)(110) = \boxed{1440.035 \text{ oz.}}$

Extrapolation

- Remember that it is dangerous to try and use the OLS regression line to predict response values for explanatory variables that are far outside of the scope of the original data.
- For example, if the dataset in the previous example only included flights between 100 minutes and 230 minutes, it would be dangerous to try to predict the amount of soda that would be consumed on a 13-hr flight (780 mins) using the OLS regression line, as we cannot be certain that the relationship between `amt. of soda` and `flight duration` remains linear for larger values of `flight duration.`
- Recall that this relates to **extrapolation**.

Inference on the Slope

- We also talked about how we can perform inference on the slope β_1 of the OLS regression line.
- Specifically, we may want to test

$$\left[\begin{array}{ll} H_0: & \beta_1 = 0 \\ H_A: & \beta_1 \neq 0 \end{array}\right.$$

  - The reason we want to test this is that, if we have reason to believe that β_1 *could* be zero, then there might not be a linear relationship between y and x at all!
- Under normality conditions,

$$\frac{\widehat{\beta_1} - \beta_1}{\mathrm{SD}(\widehat{\beta_1})} \overset{H_0}{\sim} t_{n-2}$$

Example

Worked-Out Example 4

The results of regressing a variable y onto another variable x are shown below:

	Estimate	Std. Error	*t*-value	Pr(>|t|)
Intercept	-0.05185	0.24779	-0.209	0.836
Slope	0.08783	0.07869	1.116	0.272

Is it possible that there exists no linear relationship between y and x? (Use a 5% level of significance wherever necessary.) Explain.

a. Since the _p_value of testing $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$ is 0.272, which is greater than a significance level of 5%, we would fail to reject the null; that is, it **is** possible that there exists no linear relationship between y and x.

Sampling, and the Structure of Studies

Sampling Procedures

- Finally, last lecture, we returned to the basics- data!
- Specifically, we discussed different ways data can be collected; i.e. the different **sampling procedures** that are available to us.
- In a **simple random sample**, every individual in the population has an equal chance of being included in the sample.
    - This can sometimes be costly, or even lead to biased samples.
- In a **stratified sampling** scheme, the population is first divided into several *strata* (groups), and an SRS is taken from each stratum.
    - This has the benefit of creating a potentially more representative sample, though can still be quite costly. Results are also heavily dependent on the strata that were created.
- A **cluster sampling** scheme again divides the population into groups (now called *clusters*), takes an SRS of clusters, and then takes an SRS from the selected clusters.
    - This has the benefit of being (potentially) cheaper, but can again lead to biased samples and is also heavily dependent on the clusters that were created.

Sampling Procedures

- A **convenience sample** is one in which individuals are included (or excluded) from the sample based on convenience; e.g. people who are nearby (geographically) are included whereas people who are farther away are not.
    - Convenience Samples are cheap and, well, convenient, but can lead to very skewed or biased results.
- Speaking of bias, there was another form of bias we discussed: **non-response bias**.
    - This occurs when certain individuals (or potentially even demographics, genders, etc.) do not participate in a survey, despite having been included in the sample of surveyed individuals.

Other Distinctions

- In an **observational study**, treatment is neither administered nor withheld from subjects.
- In an **experiment**, treatment *is* administered (or possibly withheld) from subjects.
- In a **longitudinal study**, subjects are tracked over a period of time. (Observations are therefore **correlated**)
- In a **cross-sectional study**, there is no tracking of subjects over time.

Example

Example (1.20 from *OpenIntro*)

On a large college campus first-year students and sophomores live in dorms located on the eastern part of the campus and juniors and seniors live in dorms located on the western part of the campus. Suppose you want to collect student opinions on a new housing structure the college administration is proposing and you want to make sure your survey equally represents opinions from students from all years.

a. What type of study is this?

b. Suggest a sampling strategy for carrying out this study.

Solutions

a. Treatment has neither been administered nor withheld, meaning this is an **observational study**.

b. **Stratified sampling** seems like the way to go, with `western campus` and `eastern campus` being the two strata.

- Specifically, we should take an SRS from both `western campus` and `eastern campus` students, to ensure that students across all years are (somewhat) equally represented.

Error

×