MAT1372, Classwork26, Fall2025

8.1 Fitting a Line, Residuals, and Correlation

1. Linear regression.

   Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

   The values β_0 and β_1 represent the model's parameters (β is the Greek letter beta), and the error is represented by ε (the Greek letter epsilon).

2. The Purpose of Linear Regression.

   When we use x to predict y, we usually call x the explanatory or predictor variable, and we call y the response; we also often drop the ε term when writing down the model since our main focus is often on the prediction of the average outcome.

3. Residual: Difference Between Observed And Expected.

   The residual of the ith observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

   We typically identify $\hat{y}_i$ by plugging x_i into the model.

4. Correlation: Strength Of a Linear Relationship.

   Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

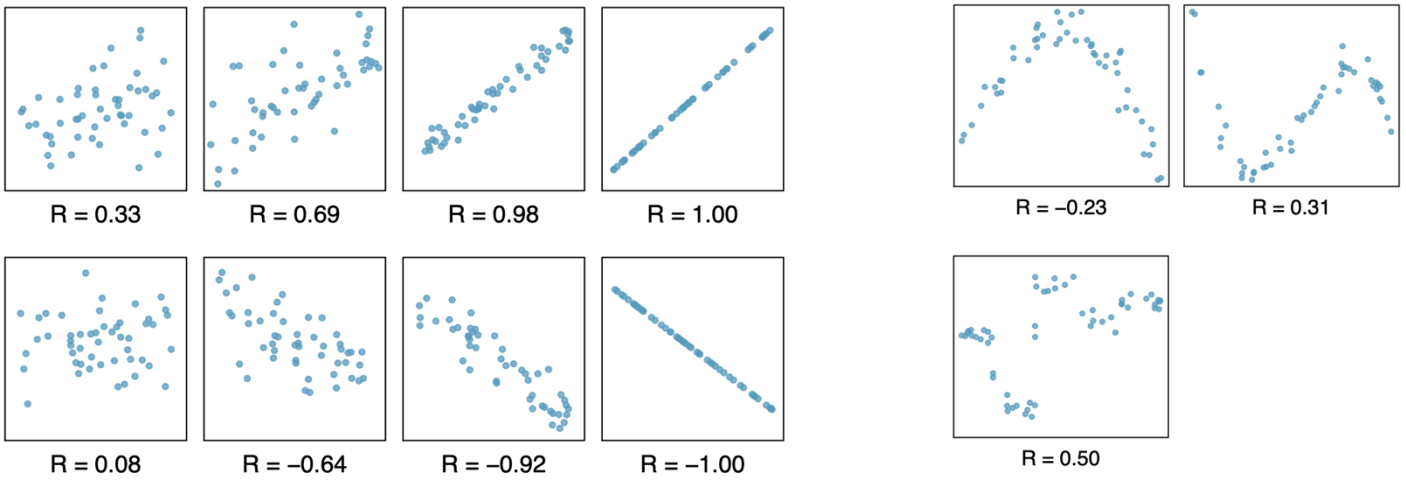
   While the correlation is generally calculated on a computer, here is the formula of correlation:

   To compute the correlation R for n observations (x_1, y_1), (x_2, y_2), …, (x_n, y_n), we have

$$R = \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

   where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

5. Examples of Correlation.



8.2 Least Squares Regression

1. An Objective Measure for Finding the Best Line.

Given n observations (x_1, y_1), (x_2, y_2), …, (x_n, y_n) and we want to find a line which fits with these data. Mathematically, we want a line that has small residuals. A more common practice is to choose the line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + e_3^2 + \cdots e_n^2$$

The line that minimizes this least squares criterion is called the least squares line.

2. Conditions for the Least Squares Line

When fitting a least squares line, we generally require

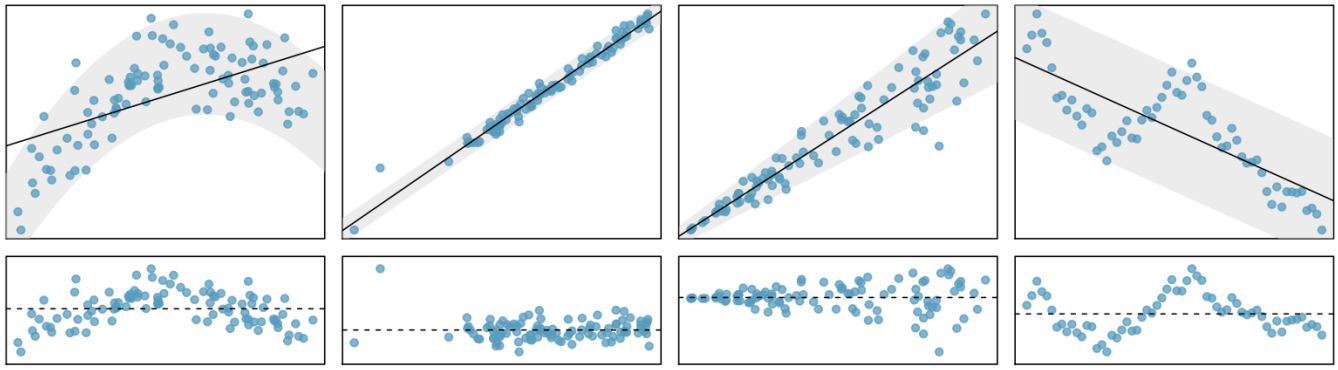
Linearity. The data should show a linear trend.

Nearly normal residuals. Generally, the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points.

Constant variability. The variability of points around the least squares line remains roughly constant.

Independent observations. Be cautious about applying regression to time series data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis.

3. Examples that fail to satisfy the conditions for Least Squares Line.



Linearity fails | Outliers | errors related to x | underlying structure

4. Finding the Least Squares Line

For a given data, we could write the equation of the least squares regression line as

$$\hat{y} - y_0 = b_1(x - x_0)$$

To identify the least squares line from summary statistics:

- Estimate the *slope* parameter, $b_1 = \frac{s_y}{s_x} R$.
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$, and $y_0 = \bar{y}$, with the point-slope equation:

$$\hat{y} - \bar{y} = b_1(x - \bar{x})$$