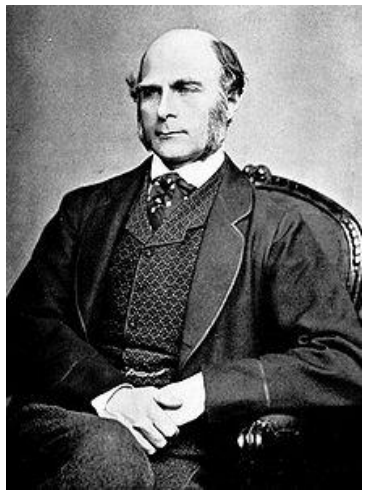
Regression and Correlation

1. Regression analysis
2. Confidence intervals in linear regression
3. Correlation analysis
4. In-class exercise

Regression and Correlation

Regression Analysis



Sir Francis Galton
1894

Introduction

- Perform experiments and collect data for two variables x and y
- Dataset: (x_1,y_1), (x_2,y_2), …, (x_n,y_n)
- Regression analysis
  - » x is a non-random, independent variable
  - » y is a random, dependent variable
  - » We want to statistically quantify the dependence of y on x
- Correlation analysis
  - » x and y are both random variables
  - » We want to statistically determine if there is a relationship between x and y

Regression Analysis

- Problem
  - » Sample two variables: (x_1,y_1), (x_2,y_2),…, (x_n,y_n)
  - » Determine dependence of random output variable Y on non-random input variable x
- Linear regression
  - » Assume the mean μ of Y depends linearly on x:

$$\mu(x) = \kappa_0 + \kappa_1 x \quad \Rightarrow \quad y = k_0 + k_1 x$$

  - » We would like to minimize the difference between the measured values $(y_1,\ldots,y_n)$ and the corresponding predicted values $(\hat{y}_1,\ldots,\hat{y}_n)$
  - » Least-squares objective to be minimized:

$$q = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 = \sum_{j=1}^{n} (y_j - k_0 - k_1 x_j)^2$$

Linear Regression

- Normal equations

$$\frac{\partial q}{\partial k_0} = -2\sum_{j=1}^{n}(y_j - k_0 - k_1 x_j) = 0 \quad \Rightarrow \quad k_0 n + k_1 \sum x_j = \sum y_j$$

$$\frac{\partial q}{\partial k_1} = -2\sum_{j=1}^{n} x_j (y_j - k_0 - k_1 x_j) = 0 \quad \Rightarrow \quad k_0 \sum x_j + k_1 \sum x_j^2 = \sum x_j y_j$$

- First normal equation

$$k_0 n + k_1 \sum x_j = \sum y_j \quad \Rightarrow \quad k_0 = \frac{1}{n}\sum y_j - \frac{k_1}{n}\sum x_j = \bar{y} - k_1 \bar{x}$$

Linear Regression

- Solution of normal equations:

$$k_0 = \frac{\sum_j x_j^2 \sum_j y_j - \sum_j x_j \sum_j x_j y_j}{n\sum_j x_j^2 - \left(\sum_j x_j\right)^2}$$

$$k_1 = \frac{n\sum x_j y_j - \sum x_j \sum y_j}{n\sum x_j^2 - \left(\sum x_j\right)^2} = \frac{s_{xy}}{s_x^2}$$

- Alternative equation for calculating k_0:

$$k_0 = \bar{y} - k_1 \bar{x}$$

Linear Regression

- Sample variances

$$s_x^2 \equiv \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right]$$

$$s_y^2 \equiv \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2\right]$$

- Sample covariance

$$s_{xy} \equiv \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)\right]$$

UMASS.

Linear Regression Example

- Reaction rate data

Experiment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Reactant Concentration	0.1	0.3	0.5	0.7	0.9	1.2	1.5	2.0
Rate	2.3	5.7	10.7	13.1	18.5	25.4	32.1	45.2

- Sample means:

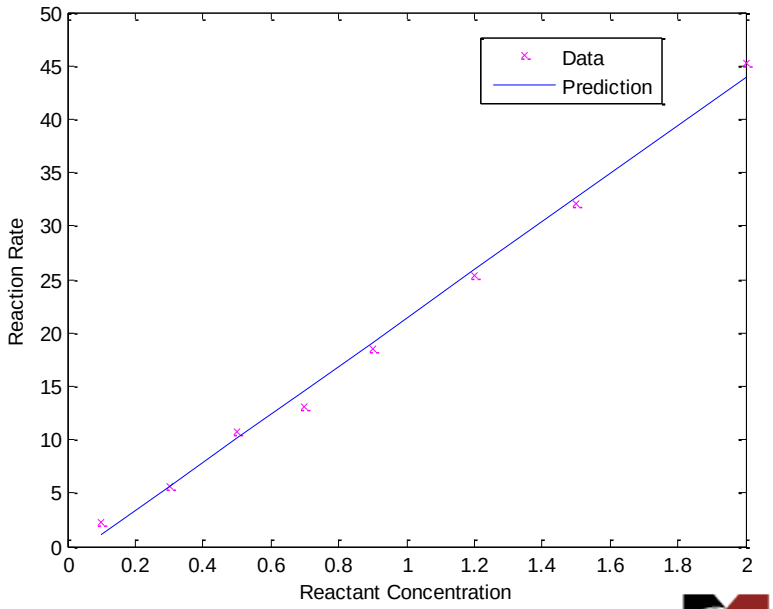
$$\bar{x} = 0.9 \quad \bar{y} = 19.125$$

- Sample variances and covariance:

$$s_x^2 = 0.4086 \quad s_y^2 = 208.4 \quad s_{xy} = 9.206$$

- Linear regression model:

$$k_0 = -1.15 \quad k_1 = 22.5 \quad y = 22.5x - 1.15$$



UMASS

Regression and Correlation

Confidence Intervals in Linear Regression Analysis

Jerzy Neyman 1937

Confidence Intervals in Linear Regression

- Assumptions
  - » The random variable Y is normal with mean $\mu(x)$ and variance σ^2 independent of x
  - » The n experiments are independent
- Confidence interval on the slope κ_1
  - » Choose confidence level γ
  - » Determine c from the t-distribution with $m = n$-2 degrees of freedom:

$$F(c) = \tfrac{1}{2}(1+\gamma)$$

  - » Compute:

$$q_0 = (n-1)(s_y^2 - k_1^2 s_x^2) \quad K = c\sqrt{\frac{q_0}{(n-2)(n-1)s_x^2}}$$

  - » Confidence interval:

$$CONF_\gamma\{k_1 - K \le \kappa_1 \le k_1 + K\}$$

Confidence Interval Example

- Reaction rate data problem: $n = 8$, $m = 6$
- Choose $\gamma = 0.95 \rightarrow F(c) = 0.975 \rightarrow c = 2.45$
- Calculate q_0 and K using results from linear regression:

$$s_x^2 = 0.4086 \quad s_y^2 = 208.4 \quad s_{xy} = 9.206$$

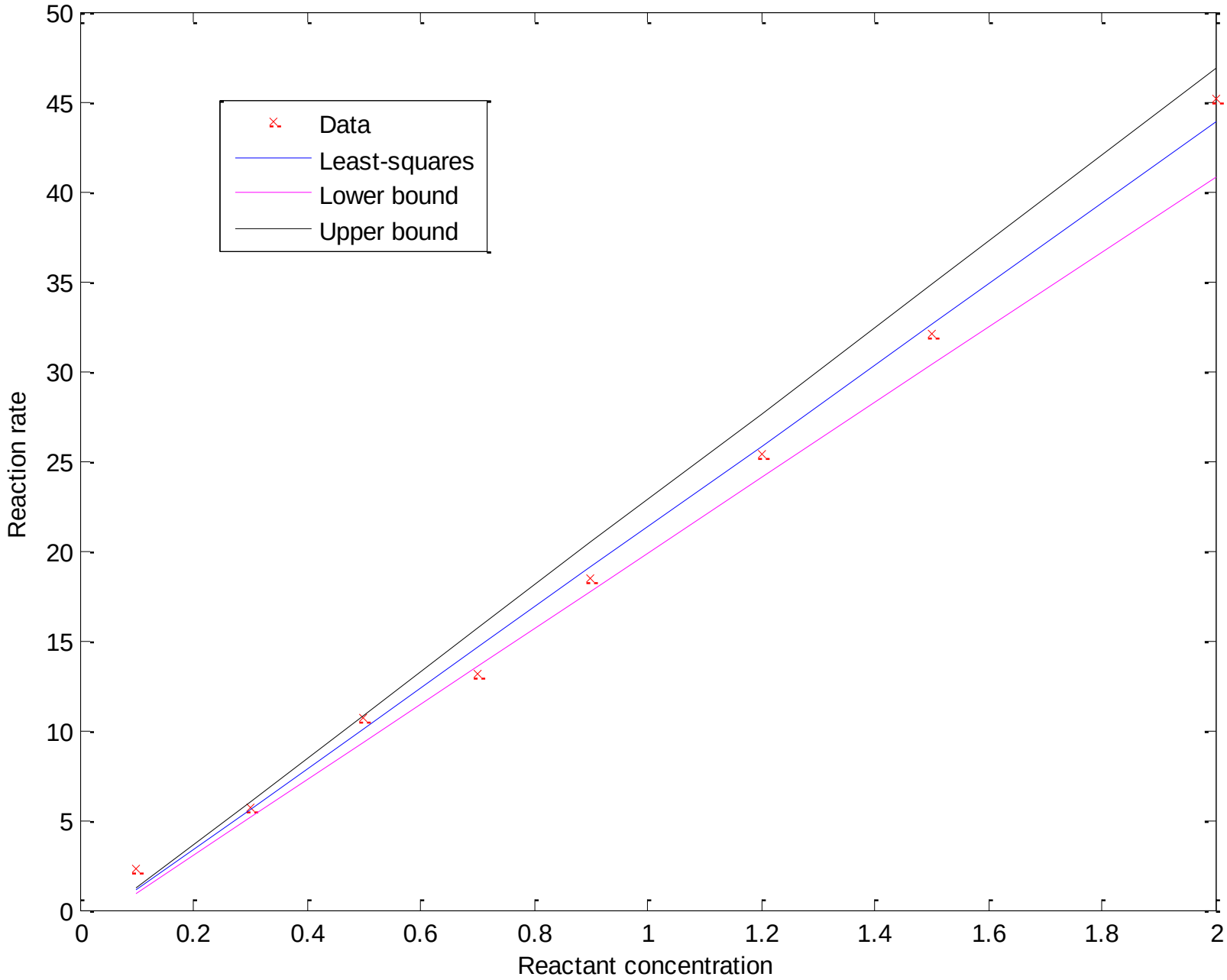
$$k_0 = -1.15 \quad k_1 = 22.5 \quad q_0 = 1140 \quad K = 1.53$$

- Confidence interval

$$CONF_{0.95}\{21.0 \le \kappa_1 \le 24.0\}$$

- Can compute confidence intervals on slope and intercept using MATLAB

Confidence Interval Example



Regression and Correlation

Correlation Analysis

Karl Pearson
1895

Correlation Analysis

- Problem
  - » Sample two variables (x_1,y_1), (x_2,y_2),…, (x_n,y_n)
  - » Determine relationship between random variables X and Y
- Sample correlation coefficient

$$r = \frac{s_{xy}}{s_x s_y} \quad -1 \le r \le 1$$

  - » $|r| = 1$ if sample values lie on a straight line
  - » $r = 0$ if the sample values are uncorrelated
- Theoretical correlation coefficient

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad -1 \le \rho \le 1$$

  - » If $|\rho| = 1$ then X and Y are linearly related: $Y = a+bX$
  - » If $\rho = 0$ then X and Y are uncorrelated

Sample Correlation Coefficient

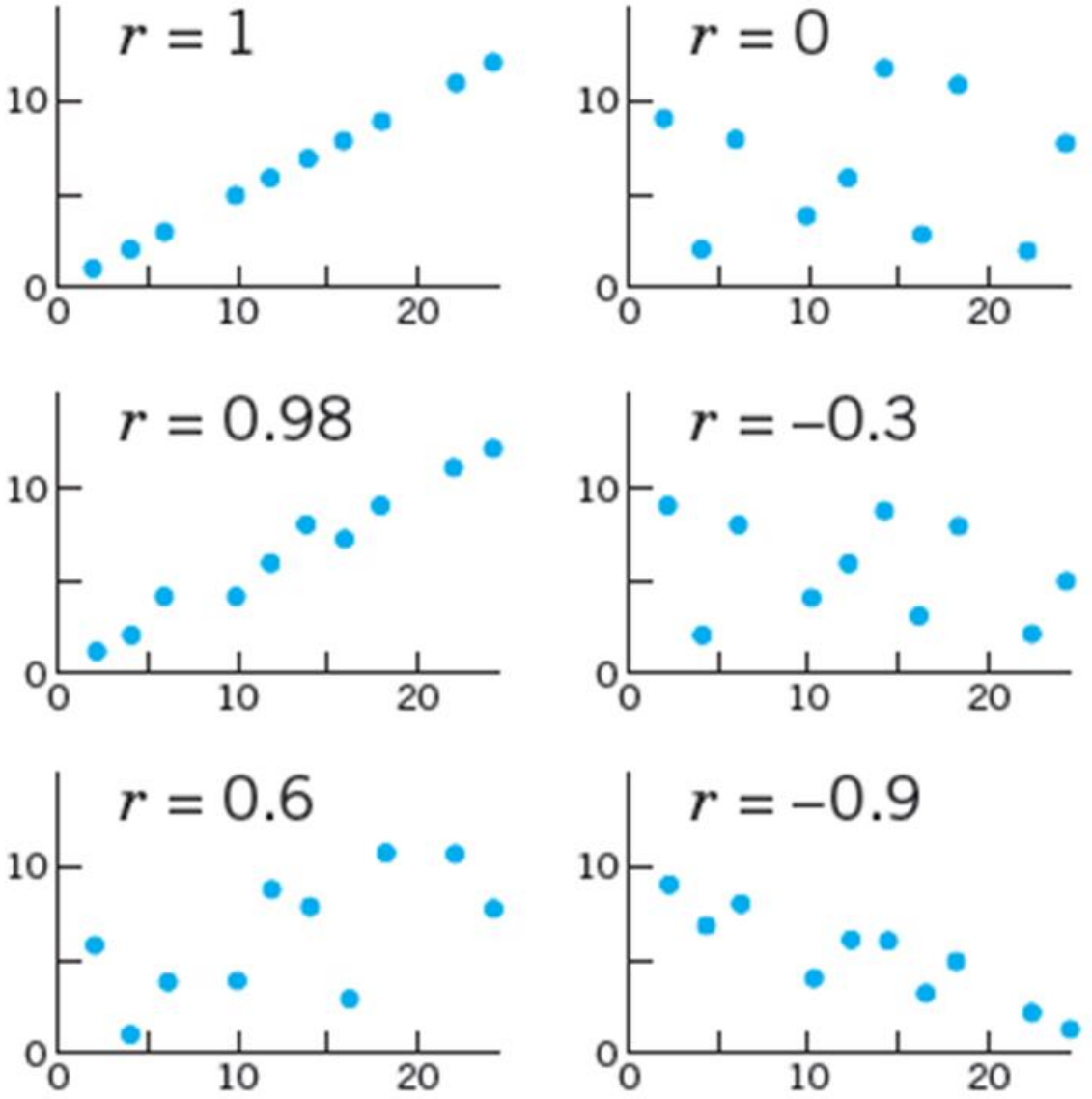


Fig. 544. Samples with various values of the correlation coefficient r

Correlation Hypothesis Test

- Test if two normal random variables are uncorrelated ($\rho = 0$) against the alternative they are correlated ($\rho > 0$)
- Choose significance level α
- Determine c from the t-distribution with $m = n$-2 degrees of freedom

$$P(T \le c) = 1 - \alpha$$

- Compute sample correlation coefficient and the t-statistic

$$r = \frac{s_{xy}}{s_x s_y} \quad -1 \le r \le 1 \quad t = r\sqrt{\frac{n-2}{1-r^2}}$$

- Accept hypothesis that X and Y are uncorrelated if $|t| \le c$

Correlation Analysis Example

- Polymerization rate data

Experiment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Hydrogen Concentration	0	0.1	0.3	0.5	1.0	1.5	2.0	3.0
Polymerization rate	9.7	9.2	10.7	10.1	10.5	11.2	10.4	10.8

- Correlation coefficient

$$s_x^2 = 1.11 \quad s_y^2 = 0.411 \quad s_{xy} = 0.421 \quad r = 0.624$$

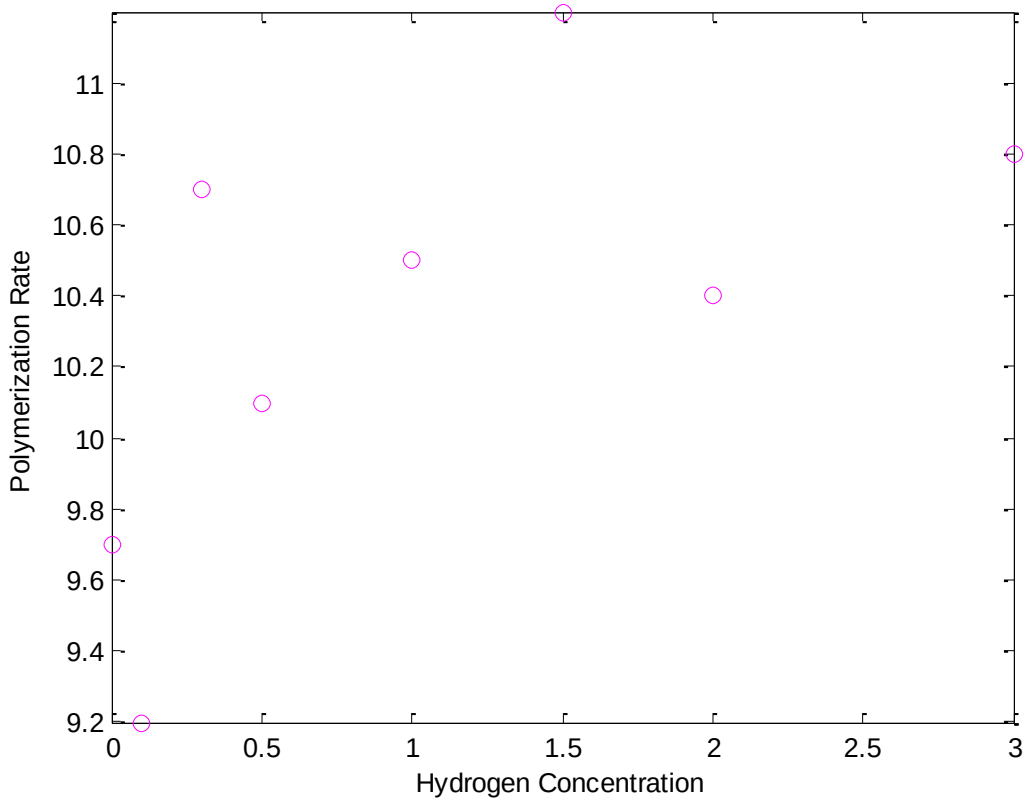
- $\alpha = 5\%$, $m = 6 \rightarrow c = 1.94$

Correlation Analysis Example

- Compute t-statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 1.95 > c = 1.94$$

- Reject the hypothesis that the hydrogen concentration and the polymerization rate are uncorrelated



Regression and Correlation

In-class Exercise