

COMPUTER LAB#3: MONTE-CARLO SIMULATIONS

Mean prediction, confidence interval

The goals of this lab are to

- explore aspects of inference based on regression modeling;
- understand the shape of confidence bands around regression lines;

In this lab, we

- use real data from a newspaper article
- compute a regression line and perform hypothesis testing

Runway work to illuminate safety hazard by Michael Sangiacomo, The Plain Dealer, 8/24/1997, Section B, pages 1 and 4, reports about number of incidents (accidents) at 10 airports in the USA and the number of arrivals and departures at those airports for the period 1990 to 1996. The Cleveland Hopkins has a disproportionately large number of incidents.

Enter the data found in the article: let the number of incidents be the dependent variable, and the number of departures and arrivals be the independent variable. Prepare below two arrays INCIDENTS and ARRIVDEPT for 9 other airports (2 vectors, or matrices, each with 1 column and 9 rows; use the Matrix toolbar)

$$\text{INCIDENTS} := \begin{pmatrix} 63 \\ 50 \\ 62 \\ 47 \\ 53 \\ 48 \\ 44 \\ 44 \\ 40 \end{pmatrix} \quad \text{ARRIVDEPT} := \begin{pmatrix} 6020212 \\ 5633187 \\ 4886539 \\ 3655929 \\ 3292425 \\ 2966632 \\ 2187357 \\ 1757460 \\ 1750872 \end{pmatrix}$$

What type of dataset is this? (time-series, cross-section, panel?)

Compute the Pearson's correlation coefficient between incidents and arrivals-departures:

$$r := \text{corr}(\text{ARRIVDEPT}, \text{INCIDENTS}) \quad r = 0.826$$

Does the correlation coefficient indicate that there might be a significant linear relationship between incidents and departures-arrivals?

Compute the regression intercept and slope, and r square:

$$\beta_1 := \text{intercept}(\text{ARRIVDEPT}, \text{INCIDENTS})$$

$$\beta_2 := \text{slope}(\text{ARRIVDEPT}, \text{INCIDENTS})$$

$$\beta_1 = 35.594$$

$$\beta_2 = 4.064 \times 10^{-6}$$

$$r^2 = 0.682$$

What does the intercept mean here? The slope?

How well does this model fit the data? Why do you think that?

Compute the homoscedastic standard error:

$$\sigma_{\text{hat}} := \sqrt{\frac{9}{7} \cdot \text{var}(\text{INCIDENTS}) \cdot (1 - r^2)}$$

$$\sigma_{\text{hat}} = 4.795$$

What does this figure mean in the context of this problem?

Compute the conditional standard error of estimate (to use for confidence bands around the regression line):

$$\text{seINCIDENTS}(x) := \sqrt{\frac{\sigma_{\text{hat}}^2}{9} \cdot \left[1 + \frac{(x - \text{mean}(\text{ARRIVDEPT}))^2}{\text{var}(\text{ARRIVDEPT})} \right]}$$

Hypothesis test

H_0 : the number of accidents in Cleveland is in line with the national standards for similarly sized airports.

H_1 : the number of accidents in Cleveland is larger than the number of accidents at similarly sized airports.

One way to test this hypothesis is to

- seek figures for Cleveland (number of incidents and number of arrivals and departures)
- plot the Cleveland values on a graph of the computed regression line (representing mean number of incidents predicted when the number of arrivals and departures is given) surrounded by a confidence band.

If Cleveland's number of incidents falls

- within the confidence band, then we can conclude that the Cleveland airport has a safety record comparable to that of other airports of with the same activity level;
- Outside the confidence band, then it is significantly different (either safer, of the value is below the regression line, or less safe if the value is above the regression line)

We begin by selecting a confidence level of 95% for our test.

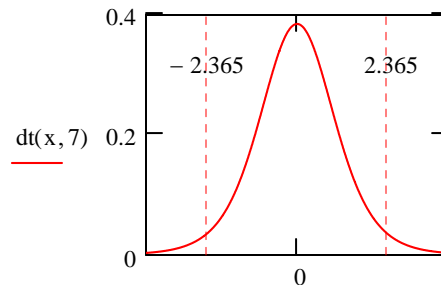
The critical t value for 7df and 95% confidence level is 2.365 (see table in the Gujarati textbook). Mathcad computes this value by using the (canned) function qt:

$$qt(0.975,7) = 2.365$$

We can graph the t distribution for 7df

(the variable denoted x represents the value of the t statistic in the range -10, 10):

$$x := -10, -9.99 .. 10$$



For verification, the confidence level we seek (.95) should be the surface under the curve between the two critical points, which is computed by integrating (summing continuously) the t probability function between the critical values:

$$\int_{-2.365}^{2.365} dt(x,7) dx = 0.95$$

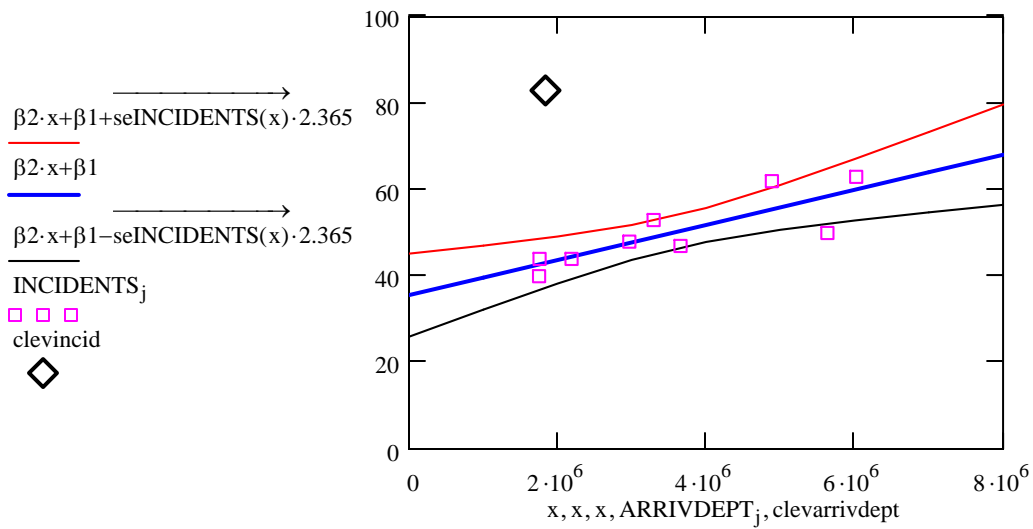
We enter the values for incidents (clevincid) and for arrivals and departures (clevarrivdept) corresponding to Cleveland's Hopkins airport:

$$\text{clevincid} := 83 \qquad \text{clevarrivdept} := 1837061$$

We graph the estimated regression line (number of incidents as a function of number of departures-arrivals) and build 95% confidence bands around the estimated regression line;

$$j := 0..8$$

$$x_j := j \cdot 10^6$$



Look at the graphic results:

How should we interpret the position of the Cleveland data with respect to the regression line and its confidence band?

- is this observation "usual" or "as expected"? Why or why not?
- what does it mean in terms of number of incidents related to departures and arrivals at the Hopkins airport?
- what do you conclude about the null hypothesis?
- what do you conclude about your traveling plans from Cleveland?

