1.4 Regression and Correlation

People in Atlanta may want to know how quickly the temperature has risen in recent years. Regression is a technique to show a statistical relationship between variables, and such as time and temperature. In the linear regression analysis, the linear relationship is derived by fitting a line through the data points in Figure 1.1. With that information, people can quantify the rate of warming over period of time when we had the measurements.

Least Squares

Consider a generic linear equation,

$$\tilde{y} = a_0 + a_1 x. \tag{1.21}$$

In our example, x can be the time and $\tilde{y}$ can be the modeled annual mean temperature of Atlanta. The goal of the regression analysis is to find the coefficients (a_0, a_1) that give us the best fit to the observation.

One way to measure the model-data misfit is to take the difference between the model ($\tilde{y}$) and observation (y), square it, and sum them up.

$$J = \frac{1}{2}\sum_i^N (\tilde{y}_i - y_i)^2 = \frac{1}{2}\sum_i^N (a_0 + a_1 x_i - y_i)^2 \tag{1.22}$$

The observations (y_i) are taken at different times (x_i). If we were to include all data points from 1879 to 2015, we would have 137 data points ($N = 137$). We can also choose to limit the analysis within a narrower time range, and fit the line for the specific period.

The method of least squares estimates the coefficients (a_0, a_1) by minimizing the square of the error. Eq. 1.22 shows that J is a quadratic equation of a_0 and a_1. J is often called as the "cost" function, which is something we would want to minimize. It is equivalent of looking for zeros of the gradient of J with respect to a_0 and a_1.

$$\frac{\partial J}{\partial a_0} = \textstyle\sum_i^N \left(a_0 + a_1 x_i - y_i\right) = 0 \tag{1.23}$$

$$\frac{\partial J}{\partial a_1} = \textstyle\sum_i^N \left(a_0 + a_1 x_i - y_i\right) x_i = 0 \tag{1.24}$$

Solving for a_0 and a_1, we find a beautiful result:

$$a_1 = cov(x, y)/var(x) \tag{1.25}$$

$$a_0 = \overline{y} - a_1\overline{x} \tag{1.26}$$

The slope (a_1) is the ratio between co-variance of y and x and the variance of x where

$$cov(y, x) = \frac{1}{N}\sum_i^N \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right) = \overline{xy} - \overline{x}\,\overline{y}. \tag{1.27}$$

Co-variance can take positive or negative values depending on the relationship between x and y. Looking at

Eq. 1.27, a positive co-variance indicates that when x is higher than average, y tends to be higher than average too. It makes sense that the co-variance essentially determines the sign of the slope of the line.

Let's apply linear regression the annual mean temperature of Atlanta. Using all data available, we get $(a_0, a_1) = (37.52°F, 0.0125°F/yr)$, and Figure 1.11 shows the regression line. a_1 is called *regression coefficient* and its positive value indicates that the temperature tends to increase with time.

There are additional two regressions in Figure 1.11 using two 41-year periods (1930-1970, 1970-2010). These examples show that the regression coefficient strongly depends on the starting and ending point. From 1930 to 1970, the regression coefficient is negative $(-0.053°F/yr)$, and it becomes positive from 1970 to 2010 $(+0.051°F/yr)$. The regression coefficients are sensitive to the period especially when the starting/ending points are close to local maximum or minimum.

Correlation

Correlation coefficient is closely related to the regression coefficient. Mathematically the correlation coefficient (r) between x and y is defined as

$$r = \frac{\sum_i^N (y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_i^N (x_i - \overline{x})^2}\sqrt{\sum_i^N (y_i - \overline{y})^2}}. \qquad (1.28)$$

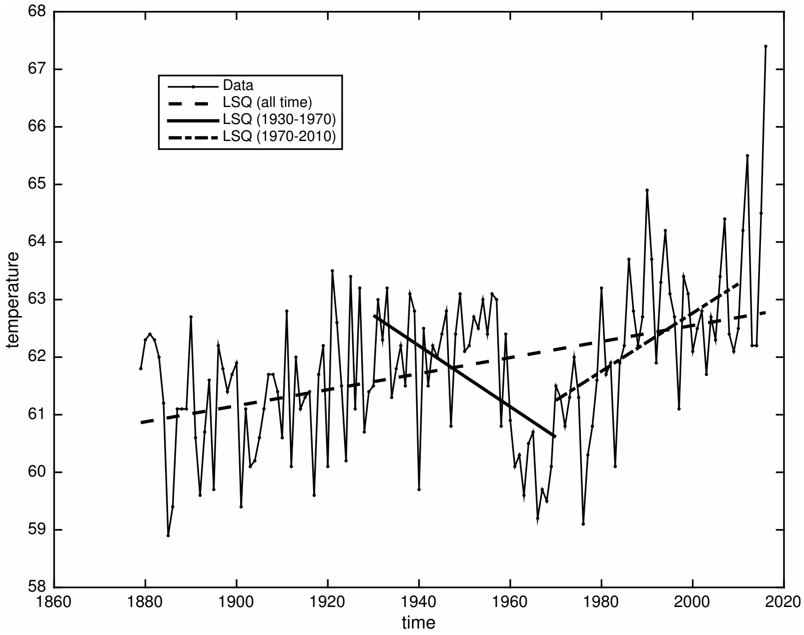

Figure 1.11: Least square fit to the annual mean temperature of Atlanta. Three different period is used for regression. Dash line uses the whole period (1879 to 2015), solid line covers 1930 to 1970, and dash dot line covers 1970 to 2010.

Typically the symbol, r, is used to indicate the correlation coefficient. The correlation measures how closely the two variables (x and y) are related.

Correlation coefficient can be contrasted to the regression coefficient,

$$a_1 = \frac{\sum_i^N (y_i - \overline{y})(x_i - \overline{x})}{\sum_i^N (x_i - \overline{x})^2}. \tag{1.29}$$

In short, we can also write the regression and correlation coefficients as $\overline{x'y'}/\sigma_x^2$ and $\overline{x'y'}/(\sigma_x \sigma_y)$ respectively, where x' and y' are the anomalies of x and y about their respective averages. In the previous section, a linear trend of temperature is calculated using the regression analysis where x is time and y is temperature. It can be generalized that the regression/correlation analysis can be applied to any two variables.

Sometime we wish to calculate what is the fraction of variance in y that is explained by the linear fit from the least square solution. This ratio can be calculated as,

$$\begin{aligned} \frac{a_1^2 \sigma_x^2}{\sigma_y^2} &= \left(\frac{\overline{x'y'}}{\sigma_x^2} \right)^2 \frac{\sigma_x^2}{\sigma_y^2} \\ &= \left(\frac{\overline{x'y'}}{\sigma_x \sigma_y} \right)^2 = r^2. \end{aligned} \tag{1.30}$$

Thus, the square of correlation coefficient is equal to the fraction of variance explained by the linear regression model.

Matrix-Vector formulation of the linear regression

Here, an alternative expression of the linear regression is provided. The annual mean temperature is expressed as a $N \times 1$ vector $\mathbf{y}$, and the linear model of the temperature can be expressed as a matrix-vector product.

$$\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} &= \begin{pmatrix} a_0 + a_1 x_1 \\ a_0 + a_1 x_2 \\ \vdots \\ a_0 + a_1 x_N \end{pmatrix} \\
\mathbf{y} &= \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}}_{\mathbf{E}} \underbrace{\begin{pmatrix} a_0 \\ a_1 \end{pmatrix}}_{\mathbf{a}} \\
\mathbf{y} &= \mathbf{Ea} \qquad (1.31)
\end{aligned}$$

The cost function becomes

$$\begin{aligned}
J &= \frac{1}{2}\left(\mathbf{y} - \mathbf{y_{obs}}\right)^T \left(\mathbf{y} - \mathbf{y_{obs}}\right) \\
&= \frac{1}{2}\left(\mathbf{Ea} - \mathbf{y_{obs}}\right)^T \left(\mathbf{Ea} - \mathbf{y_{obs}}\right) \qquad (1.32)
\end{aligned}$$

Then taking $\partial J / \partial \mathbf{a} = 0$ gives the least square fit solution for $\mathbf{a}$.

$$\mathbf{E}^T \left(\mathbf{Ea} - \mathbf{y_{obs}}\right) = 0$$

$$\begin{aligned}\mathbf{E}^T\mathbf{E}\mathbf{a} - \mathbf{E}^T\mathbf{y_{obs}} &= 0\\ \mathbf{a} &= \left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{y_{obs}}\end{aligned} \tag{1.33}$$

This results in exactly the same answer as Eqs. 1.25 and 1.26. The matrix product, $\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T$, is known as the pseudo-inverse of $\mathbf{E}$.

Exercises

1. Using the annual mean temperature of Atlanta from 1879 to 2015, reproduce Figure 1.11 including the three regression lines.

2. Compare the two methods of calculating the regression coefficients and make sure that you get the same result.

3. Calculate the fraction of variance explained by the linear trend for the three regression lines.

4. Make your own MATLAB function that calculates regression and correlation coefficient between two variables. Input is the two variable of equal length, and the output is the regression and correlation coefficients.

Advanced exercises

1. Download the monthly surface air temperature (*air.sig995.mon.mean.nc* or *air.mon.mean.nc*) from

NCEP reanalysis `https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surface.html`. Calculate the monthly climatology of the surface air temperature for the 68-year period of 1948-2015.

2. Subtract the monthly climatology from the temperature to remove the mean seasonal cycle. Generate a longitude-latitude map of the temperature trend for the 68 year period.

3. Generate a longitude-latitude map of (r^2) the fraction of variance explained by the linear trend.

4. Publish the MATLAB script that performs all of the activity above, and submit it as a report in the PDF format.