Reviewing Basic Statistics

Armine Hayrapetyan

Assumptions of linear regression

When doing a simple regression model, we make the (often reasonable!) assumptions that

- the errors are normally distributed and, on average, zero
- the errors all have the same variance (they are homoscedastic)
- the errors are unrelated to each other (they are independent across observations)

Assumptions of linear regression

Written mathematically (with the third assumption relaxed somewhat) for normally distributed errors

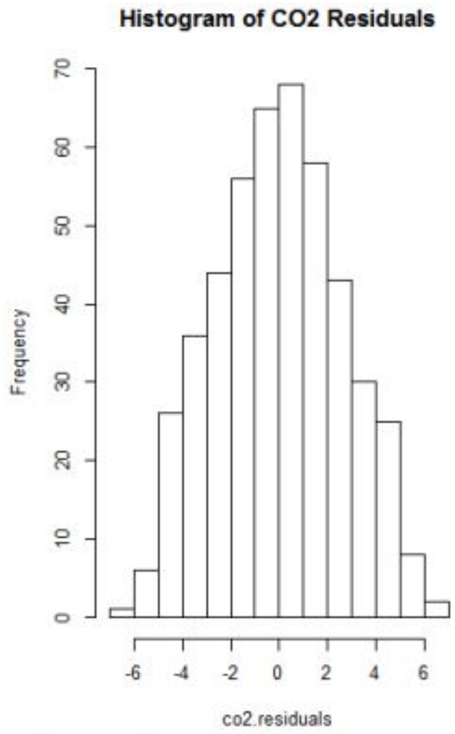
- $E[\epsilon_i] = 0$
- $Var[\epsilon_i] = \sigma^2\{\epsilon_i\} = constant = \sigma^2$
- $Cov[\epsilon_i, \epsilon_j] = \sigma\{\epsilon_i, \epsilon_j\} = 0, \ \forall \ i \neq j$

Even more compactly

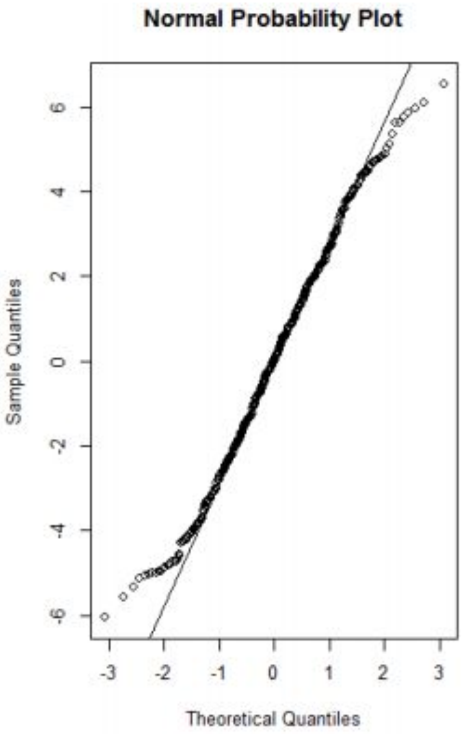
$$\boldsymbol{\epsilon_i \ iid \sim N(\mu = 0, \sigma^2 = constant)}$$

Normality of residuals

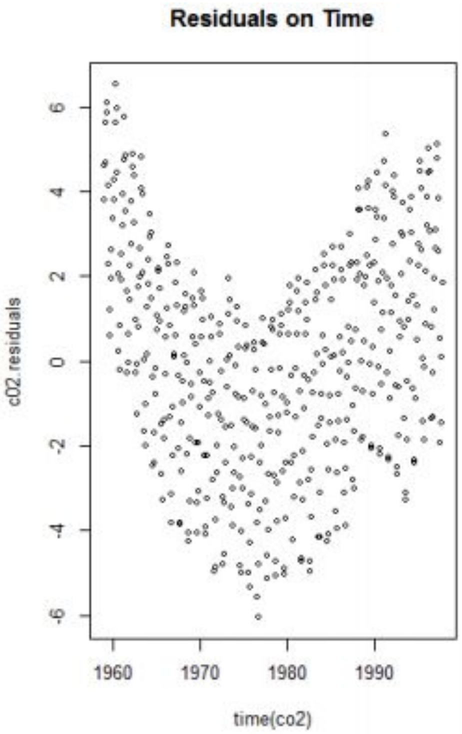
When we have a large data set we can look at a **histogram**. When a data set is smaller, we can look at a **normal probability plot**.



Histogram of CO2 Residuals



Normal Probability Plot



Residuals on Time

Normal probability plot

The normal probability plot is a graphical technique for assessing whether or not a data set is approximately normally distributed.

The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

Variation

Recall that the variance of a single random variable is written, for the random variable X as

$$\sigma^2 \equiv V[X] \equiv E[\ (X - \mu_x\ )(X - \mu_X)\ ]$$

For a data set we'd estimate this as

$$s^2 \equiv (1\ /\ (n - 1))\ \sum(x_i - \bar{x})(x_i - \bar{x})$$

Covariance

Now, if we have two random variables, we think about measuring their linear relationship with

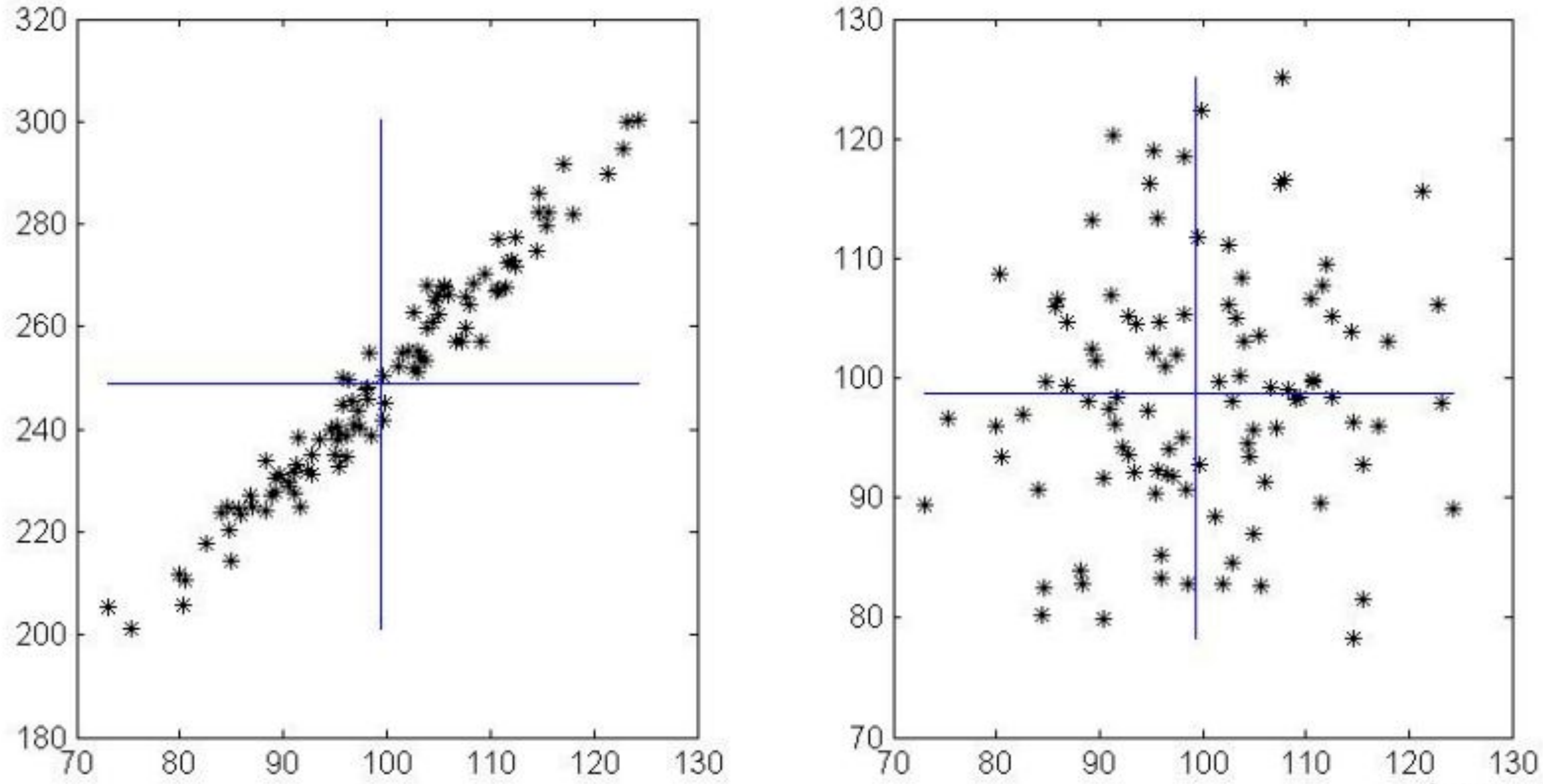
$$COV[X, Y] \equiv E[\ (X - \mu_X)(Y - \mu_Y\ )]$$

And, for data, we form the analogous estimator

$$cov \equiv ((1\ /\ n - 1))\ \sum(x_i - \bar{x})(y_i - \bar{y})$$

For motivation, we look at what happens “on average” (that’s the expected value operator, $E[\ ]$) when we center the random variables, and then multiply these quantities together. Let’s think through graphically why this is a good idea.

Covariance


320
300
280
260
240
220
200
180
70
80
90
100
110
120
130
130
120
110
100
90
80
70
70
80
90
100
110
120
130

Covariance

If you look at the first graph you will see that most of the above average x values go along with the above average y values. It is the same thing with the below average x values and the below average y values. Think about the deviations (distance from the mean). This means that

- When $x_i - \overline{x} > 0$ it is pretty common to find $y_i - \overline{y} > 0$ as well. A positive times a positive is positive, so this means that $(x_i - \overline{x})(y_i - \overline{y}) > 0$ is also greater than 0
- On the other side, when $x_i - \overline{x}< 0$ it is pretty common to find $y_i - \overline{y} < 0$ as well. A negative times a negative is negative, so this means that $(x_i - \overline{x})(y_i - \overline{y}) > 0$

Covariance

- There aren't many positive x values associated with negative y values. So there aren't many terms where $x_i - \overline{x} > 0$ and $y_i - \overline{y} < 0$. That means there aren't many terms where $(x_i - \overline{x})(y_i - \overline{y}) < 0$
- Just to be complete, there aren't many negative x values associated with positive y values. So there aren't many terms where $x_i - \overline{x} < 0$ and $y_i - \overline{y} > 0$. Again, that means there aren't many terms where $(x_i - \overline{x})(y_i - \overline{y}) < 0$

When we move on to correlation, we are really just expressing the covariance concept in standard units. The motivation for this might be obvious- if we are measuring strength of linear association, we shouldn't have to worry about whether we've measure in feet, in inches, or in miles.

Correlation

If we think about it in this way the defining formula for random variables should make sense:

$$\rho(X,Y) \equiv E\left[\left(\frac{X-\mu_X}{\sigma_x}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

If we have data instead of random variables, we estimate this term in the most direct way as

$$r \equiv \hat{\rho} \equiv \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_X}\right)\left(\frac{y_i-\bar{y}}{s_Y}\right)$$

Sum of squares notation

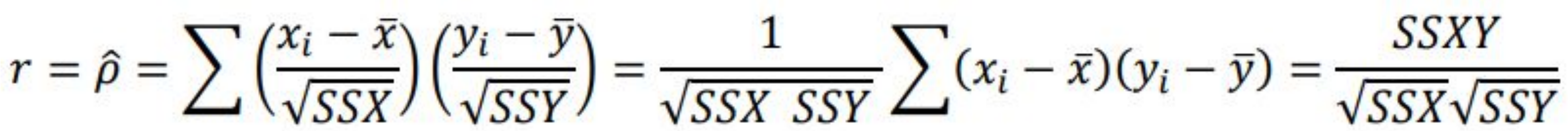
Remember the “sum of squares” notation as follows:

$$SSX \equiv \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n} \sum x_i \sum x_i$$

$$SSY \equiv \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n} \sum y_i \sum y_i$$

$$SSXY \equiv \sum x_i \, y_i - \frac{1}{n} \sum x_i \sum y_i$$

Correlation coefficient - r



$$r = \hat{\rho} = \sum \left(\frac{x_i - \bar{x}}{\sqrt{SSX}}\right)\left(\frac{y_i - \bar{y}}{\sqrt{SSY}}\right) = \frac{1}{\sqrt{SSX\ SSY}} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{SSXY}{\sqrt{SSX}\sqrt{SSY}}$$

References

- https://d3c33hcgiwev3.cloudfront.net/_9b250c9618ca8683ceba9343e0c0c83a_Introduction-to-R---Reviewing-Basic-Statistics.pdf?Expires=1581984000&Signature=hEsUB9Mg6ZF5FQPLn8yIuKBn9dH0EJyeuJnq2T6IYRAVyOgozc-iFsU7yKUDI4XvDVgxH2TAfSITQi3NE9TGT2804gEl~wSWXpaUdhP-lPZU~PQXBJftwlRaxc4yIndec~8prkGGVIIR5IMAGeXx9TXHhp3Ny9tdN5P-STkvw7U_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A
- https://d3c33hcgiwev3.cloudfront.net/_edba2673df58a3a2416f3cef12777fc1_Measuring-Linear-Association-with-the-Correlation-Function.pdf?Expires=1582070400&Signature=Pv-5Z~FdgsXH6YpmVtFcS5Fjs3CvYogxahXsDxGBNEfihmbKH6K-mQFZQ5q8-GwxN7rSkXFCVncxfcrU3eheurV6fQFfvFiZU5fQIlPpggQbNXJIjPBBLkvr5VAlJTC5MH2QcQUVOKZpbTr0BAHFkBioz6UI5CckzQpUVf9QYRw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A