MA3206

Statistics I

Spring 2022

Satvik Saha
19MS154

Indian Institute of Science Education and Research, Kolkata,
Mohanpur, West Bengal, 741246, India.

Contents

1 Introduction

We are interested in two types of data: *categorical* and *numerical.* Categorical data used named qualities to describe a particular observation. This can be further categorized into *nominal* and *ordinal*; the latter admit a natural ordering. Numerical data uses numbers, and can be further categorized into *discrete* and *continuous.*

2 Measures of central tendency

2.1 Arithmetic mean

Suppose that we have been given a collection of n numeric observations, denoted $x_1, x_2, \ldots, x_n$. These may be concentrated around some specific point, or spread out over some range; regardless, we wish to identify one particular point around which our observations are 'balanced' or aggregate in some sense. In other words, we want to identify a point $\bar{x}$ such that the net deviation $|x_i - \bar{x}|$ is minimized. For convenience, we consider the square deviations $(x_i - \bar{x})^2$; thus, we wish to minimize the loss function defined by

$$t \mapsto \sum_{i=1}^{n} (x_i - t)^2.$$

It is easy to check that our loss function attains its minimum at

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This quantity $\bar{x}$ is called the *arithmetic mean* of our data. Note that this is not the only choice of loss function measuring central tendency, but it is certainly quite convenient.

If our data is summarized in terms of frequencies, i.e. each x_i has been recorded f_i times, we may write

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i, \qquad N = \sum_{i=1}^{n} f_i.$$

The quantities f_i/N are often referred to as the *weights* of the observations x_i. The arithmetic mean can thus be interpreted as their 'centre of mass'.

Now suppose that our data values have not been explicitly presented: instead, we have been given the data classes $(x_{i-1}, x_i]$ and the number of observations f_i falling within each class. We can make an estimate of the true mean by identifying each data class with some value, say $(x_{i-1}, x_i]$ gets associated with $x_i^* = (x_{i-1} + x_i)/2$. Then we calculate the usual arithmetic mean using these values. This gives us the estimate

$$\bar{x}^* = \frac{1}{N} \sum_{i=1}^{n} f_i x_i^*, \qquad N = \sum_{i=1}^{n} f_i.$$

Note that the true mean must lie within the bounds

$$\frac{1}{N} \sum_{i=1}^{n} f_i x_{i-1} \ \leq \bar{x} \leq \ \frac{1}{N} \sum_{i=1}^{n} f_i x_i.$$

Suppose that each data class has width h. We may estimate the error in our mean by observing that within a particular class $(x_{i-1}, x_i]$ with frequency f_i, the deviation between any of the true data points and x_i^* is at most $h/2$. Thus, the net deviation accumulated over a particular class is at most $f_i h/2$, and the net deviation overall is at most $Nh/2$. Putting everything together, we have

$$|\bar{x} - \bar{x}^*| \leq \frac{h}{2}.$$

2.2 Geometric mean

Another measure of central tendency is the geometric mean G, calculated

$$G = \sqrt[n]{x_1 x_2 \cdots x_n}.$$

Note that

$$\log G = \frac{1}{n}\sum_{i=1}^{n} \log x_i.$$

Consider k sets of observations, with n_i observations in each set. Then, the geometric mean of the combined data is related with the geometric means G_I of the sets as

$$\log G = \frac{1}{N}\sum_{i=1}^{k} n_i \log G_i, \qquad N = \sum_{i=1}^{k} n_i.$$

2.3 Harmonic mean

Another measure of central tendency is the harmonic mean G, calculated

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}.$$

The Harmonic means of combined data and sets of data are related as

$$\frac{N}{H} = \sum_{i=1}^{k} \frac{n_i}{H_i}, \qquad N = \sum_{i=1}^{k} n_i.$$

Exercise 2.1. Given two positive numbers, their arithmetic, geometric, and harmonic means all lie between them.

Proof. Without loss of generality, let $x \geq y > 0$. Then for any a, b, we have

$$x = \frac{ax + bx}{a + b} \geq \frac{ax + by}{a + b} \geq \frac{ay + by}{a + b} = y.$$

Setting $a = b = 1/2$ give the result for the arithmetic mean. Now, the logarithm function is monotonic for positive reals, so $\log x \geq \log y$. Applying the above gives

$$\log x \geq \frac{1}{2}(\log x + \log y) \geq \log y,$$

and taking exponentials yields

$$x \geq \sqrt{xy} \geq y.$$

Finally, applying the result to $1/y \geq 1/x$, we have

$$\frac{1}{y} \geq \frac{a/y + b/x}{a + b} \geq \frac{1}{x},$$

which we can rearrange and set $a = b = 1/2$ to get

$$x \geq \frac{2}{1/x + 1/y} \geq y. \qquad \square$$

Remark. The same proof applies for weighted means.

Theorem 2.1. *For n observations $x_1, \dots, x_n$, the arithmetic mean, geometric mean, and harmonic mean are in descending order, i.e.*

$$AM \geq GM \geq HM.$$

Proof. We assume that all $x_i > 0$. Consider the case $n = 2$. Then,

$$(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0, \qquad x_1 + x_2 \geq 2\sqrt{x_1 x_2}$$

is precisely $AM \geq GM$. Applying the same on the reciprocals,

$$\frac{1}{x_1} + \frac{1}{x_2} \geq 2\sqrt{\frac{1}{x_1 x_2}}, \qquad \sqrt{x_1 x_2} \geq \frac{2}{1/x_1 + 1/x_2}$$

is precisely $GM \geq HM$.

Suppose that the result holds for some n. Now consider a collection of $2n$ observations $x_1, \dots, x_{2n}$. Then, applying $AM \geq GM$ on both halves, then the two variable case gives

$$\sum_{i=1}^{2n} x_i \geq n\sqrt[n]{x_1 \cdot x_n} + n\sqrt[n]{x_{n+1} \cdots x_{2n}} \geq 2n\sqrt[2n]{x_1 \cdots x_n x_{n+1} \cdots x_{2n}}$$

which is precisely $AM \geq GM$ for $2n$ observations. Now suppose that $AM \geq GM$ holds for some $n + 1$. Consider a collection of n observations $x_1, \dots, x_n$, set $\bar{x} = (x_1 + \cdots + x_n)/n$, and note that

$$\sum_{i=1}^{n} x_i + \bar{x} \geq (n+1)\sqrt[n+1]{x_1 \cdots x_n \bar{x}}.$$

The left-hand side is simply $(n+1)\bar{x}$, so

$$\bar{x} \geq \sqrt[n+1]{x_1 \cdots x_n \bar{x}}, \qquad \bar{x}^{n/n+1} \geq (x_1 \cdots x_n)^{1/n+1}, \qquad \bar{x} \geq \sqrt[n]{x_1 \cdots x_n},$$

which is precisely $AM \geq GM$ for n observations. Therefore, $AM \geq GM$ holds for all $n \geq 2$ by induction.

Now that we have $AM \geq GM$ for n observations, use it on their reciprocals to get

$$\sum_{i=1}^{n} \frac{1}{x_i} \geq n\sqrt[n]{\frac{1}{x_1 \cdots x_n}}, \qquad \sqrt[n]{x_1 \cdots x_n} \geq \frac{n}{\sum_{i=1}^{n} 1/x_i}$$

which is precisely $GM \geq HM$. □

2.4 Median

The median of a collection of ordered observations $x_1 \leq x_2 \leq \cdots \leq x_n$ is defined to be their middle value: x_{k+1} if $n = 2k + 1$ is odd, and the mean $(x_k + x_{k+1})/2$ if $n = 2k$ is even.

For grouped data, we assume that the observations are evenly distributed over the median class $(l, u]$ with frequency f_m, width h. If the total frequency is denoted by N, we write

$$\frac{M - l}{h} = \frac{N/2 - n_l}{f_m}.$$

Here, n_l is the cumulative frequency of the preceding classes. This will give

$$M = l + \frac{N/2 - n_l}{f_m} \cdot h.$$

Another way of estimating the median of grouped data is by drawing the more than and less than ogives, and picking the abscissa of their intersection point. In the median class, the ogives have the equations

$$y = n_l + \frac{f_m}{h}(x - l), \qquad y = N - n_l - \frac{f_m}{h}(x - l).$$

Solving for their intersection, we recover our formula.

Theorem 2.2. *Let φ be a monotone function, and let two variables be related as $y = \varphi(x)$. Then their medians are related as $M_y = \varphi(M_x)$.*

Theorem 2.3. *The median of a combination of two sets of observations lies in between the individual medians.*

2.5 Mode

The mode of a collection of observations $x_1, \dots, x_n$ is the value with the highest frequency.

For grouped data, we pick the value with the highest frequency density. Let f_m denote the frequency of the modal class $(l, u]$. We approximate

$$M_0 = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \cdot h.$$

Theorem 2.4. *An empirical relation between these measures of central tendency is given by*

$$\textit{mean} - \textit{mode} \approx 3(\textit{mean} - \textit{median}).$$

3 Measures of dispersion

3.1 Range

The range is a simple way of measuring how *dispersed* or scattered a set of observations is. This is simply the difference between the maximum and the minimum value in the set.

Theorem 3.1. *If two variables are related by $y = a + bx$, then their ranges are related by*

$$R_Y = |b| \cdot R_X.$$

3.2 Mean deviation

The mean deviation about some value α is defined by

$$MD(\alpha) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \alpha|.$$

Theorem 3.2. *If two variables are related by $y = a + bx$, then*

$$MD_Y(\alpha) = |b| \cdot MD_X(\alpha).$$

Theorem 3.3. *The mean deviation about a point is minimized at the median.*

Exercise 3.1. The mean deviation is given by

$$n \cdot MD(\alpha) = S_2 - S_1 + (n_1 - n_2)\alpha.$$

Here, n_1 is the number of values less than α and S_1 is their sum, and n_2 is the number of values more than α and S_2 is their sum.

Proof. Calculate

$$\begin{aligned} n \cdot MD(\alpha) &= \sum_{i=1}^{n} |x_i - \alpha| \\ &= \sum_{x_i<\alpha} \alpha - x_i + \sum_{x_i \geq \alpha} x_i - \alpha \\ &= n_1\alpha - S_1 + S_2 - n_2\alpha \\ &= S_2 - S_1 + (n_1 - n_2)\alpha. \end{aligned}$$ □

By denoting n_α to be the number of values less than α, S_α to be their sum, and S to be the sum of all elements, we have

$$n \cdot MD(\alpha) = S - 2S_\alpha + (2n_\alpha - n)\alpha.$$

3.3 Root mean square deviation

The RMS deviation about some value α is defined by

$$RMS(\alpha) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2}.$$

We call $RMS(\bar{x})$ the standard deviation σ, and its square the variance. We can calculate

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Theorem 3.4. *The root mean square deviation about a point is minimized at the mean.*

Theorem 3.5. *The standard deviations of two sets of observations are related by*

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1+n_2}.$$

In general, for k sets of observations, we have

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} n_i\sigma_i^2 + n_i(\bar{x}_i - \bar{x})^2.$$

Theorem 3.6. *If two variables are related by $y = a + bx$, then*

$$\sigma_Y = |b| \cdot \sigma_X.$$

Exercise 3.2. If a single observation α is added to a set of n values, then the standard deviation increases only if

$$|\bar{x} - \alpha| > \sqrt{\frac{n+1}{n}} \cdot \sigma.$$

Proof. Without loss of generality, let $\bar{x} = 0$; this can be done by relabelling the data $x_i - \bar{x}$, putting the mean at zero without affecting the variance. Thus, we have

$$n\sigma^2 = \sum_{i=1}^{n} x_i^2.$$

Upon adding the point α to our data, the new mean is

$$\bar{x}_n = \frac{\alpha}{n+1},$$

so our new variance is related as

$$\begin{aligned}(n+1)\sigma_n^2 &= \sum_{i=1}^{n} x_i^2 + \alpha^2 - (n+1)\bar{x}_n^2 \\ &= n\sigma^2 + \alpha^2 - \frac{1}{n+1}\alpha^2, \\ (n+1)(\sigma_n^2 - \sigma^2) &= -\sigma^2 + \frac{n}{n+1}\alpha^2.\end{aligned}$$

For the standard deviation to increase, this must be positive, hence

$$\alpha^2 > \frac{n+1}{n}\sigma^2$$

as desired. □

Theorem 3.7. *The mean deviation about the mean cannot exceed the standard deviation.*

Theorem 3.8. *The difference between the mean and median cannot exceed the standard deviation.*

Exercise 3.3. The range and standard deviation obey

$$\frac{R^2}{2n} \leq \sigma^2 \leq \frac{R^2}{4}.$$

Proof. Without loss of generality, let $\bar{x} = 0$. Then,

$$n\sigma^2 = \sum_{i=1}^{n} x_i^2, \qquad R = x_n - x_1.$$

Set $\alpha = (x_n + x_1)/2$. Since the RMS deviation is minimized at the mean, we have

$$\begin{aligned} n\sigma^2 &\leq \sum_{i=1}^{n} (x_i - \alpha)^2 \\ &= \sum_{x_i < \alpha} (x_i - \alpha)^2 + \sum_{x_i \geq \alpha} (x_i - \alpha)^2 \\ &\leq \sum_{x_i < \alpha} (x_1 - \alpha)^2 + \sum_{x_i \geq \alpha} (x_n - \alpha)^2 \\ &= \sum_{x_i < \alpha} \frac{R^2}{4} + \sum_{x_i \geq \alpha} \frac{R^2}{4} \\ &= \frac{nR^2}{4}. \end{aligned}$$

Finally note that $RMS \geq AM$ gives

$$n\sigma^2 \geq x_n^2 + x_1^2 \geq \frac{(|x_n| + |x_1|)^2}{2} = \frac{R^2}{2}.$$ □

Lemma 3.9. *The standard deviation is given by*

$$\sigma^2 = \frac{1}{2n^2} \sum_{i,j} (x_i - x_j)^2.$$

Proof. Observe that

$$\sum_{ij} (x_i - x_j)^2 = 2\sum_{ij} x_i^2 - \sum_{ij} 2x_j x_j = 2n\sum_i x_i^2 - 2\sum_i x_i \sum_j x_j = 2n\sum_i x_j^2 - 2n^2\bar{x}^2.$$ □

3.4 Quartile deviation

A *quantile* of order p is such a value of the variable such that a proportion p of all the values are less than or equal to it. For grouped data, we estimate

$$z_p = l + \frac{np - n_l}{f_m} \cdot h.$$

The quartile deviation, or semi-interquartile range is defined

$$Q = \frac{Q_3 - Q_1}{2} = \frac{z_{3/4} - z_{1/4}}{2}.$$

3.5 Coefficient of variation

Unlike the previous measures, the coefficient of variation is a *relative* measure of dispersion, expressed as a percentage.

$$CV = \frac{\sigma}{\bar{x}}.$$

A variable having a lower coefficient of variation is considered to be more stable. Similar coefficients are

$$CV(\alpha) = \frac{MD(\alpha)}{\alpha}$$

where α is the mean or median.

Exercise 3.4. Suppose that the deviations $x_i - \bar{x}$ are small, so that $((x_i - \bar{x})/\bar{x})^3$ and higher powers can be neglected. Then,

1. $GM \approx \bar{x}(1 - \sigma^2/2\bar{x}^2)$.
2. $HM \approx \bar{x}(1 - \sigma^2/\bar{x}^2)$.
3. $\bar{x}^2 - GM^2 \approx \sigma^2$.
4. $\bar{x} - 2GM + HM \approx 0$.
5. $E(\sqrt{X}) \approx \sqrt{\bar{x}}(1 - \sigma^2/8\bar{x}^2)$.

Proof.

1. Write
$$\log GM = \frac{1}{n}\sum \log x_i = \frac{1}{n}\sum \log\left[\bar{x}\left(1 + \frac{x_i - \bar{x}}{\bar{x}}\right)\right].$$
Using the series expansion of the logarithm, this is approximately
$$\log \bar{x} + \frac{1}{n}\sum \frac{x_i - \bar{x}}{\bar{x}} + \frac{(x_i - \bar{x})^2}{2\bar{x}^2} = \log \bar{x} - \frac{\sigma^2}{2\bar{x}^2}.$$
Finally, use $e^\alpha \approx 1 + \alpha$ to write
$$GM \approx \bar{x}\left(1 - \frac{\sigma^2}{2\bar{x}^2}\right).$$

2. Write
$$\frac{1}{HM} = \frac{1}{n}\sum \frac{1}{x_i} = \frac{1}{n\bar{x}}\sum \left[1 + \frac{x_i - \bar{x}}{\bar{x}}\right]^{-1}.$$
Using the series expansion of $1/(1+x)$, this is approximately
$$\frac{1}{n\bar{x}}\sum 1 - \frac{x_i - \bar{x}}{\bar{x}} + \frac{(x_i - \bar{x})^2}{\bar{x}^2} = \frac{1}{\bar{x}}\left(1 + \frac{\sigma^2}{\bar{x}^2}\right).$$
Taking the reciprocal and approximating $(1+\alpha)^{-1} \approx 1 - \alpha$ gives
$$HM \approx \bar{x}\left(1 - \frac{\sigma^2}{\bar{x}^2}\right).$$

3. Use the first approximation to estimate

$$\bar{x}^2 - GM^2 \approx \bar{x}^2 \left[1 - \left(1 - \frac{\sigma^2}{2\bar{x}^2}\right)^2\right].$$

Use $(1-x)^2 \approx 1 - 2x$ to write

$$\bar{x}^2 - GM^2 \approx \bar{x}^2 \left[1 - 1 + \frac{\sigma^2}{\bar{x}^2}\right] = \sigma^2.$$

4. Use the first two approximations to write

$$\bar{x} - 2GM + HM \approx 0.$$

5. Write

$$E[\sqrt{X}] = \frac{1}{n}\sum \sqrt{x_i} = \frac{\sqrt{\bar{x}}}{n}\sum \left[1 + \frac{x_i - \bar{x}}{\bar{x}}\right]^{1/2}.$$

Using the series expansion of the square root, this is approximately

$$E[\sqrt{X}] \approx \frac{\sqrt{\bar{x}}}{n}\sum 1 + \frac{x_i - \bar{x}}{2\bar{x}} - \frac{(x_i - \bar{x})^2}{8\bar{x}^2} = \sqrt{\bar{x}}\left(1 - \frac{\sigma^2}{8\bar{x}^2}\right).$$

□

3.6 Moments

The rth order moment about α is given by

$$m_r(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(x - \alpha)^r.$$

The corresponding central moment is simply

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^r.$$

The rth order raw moment is simply $m_r' = m_r(0)$.

Lemma 3.10. *If two variables are related by $y = a + bx$, then*

$$m_{r,Y} = b^r \cdot m_{r,X}.$$

Lemma 3.11. *The central moments can be expressed in terms of raw moments as*

$$m_r = \sum_{k=0}^{r}\binom{n}{k}(-1)^k m_{r-k}'(m_1')^k.$$

Definition 3.1. Define

$$b_1 = \frac{m_3^2}{m_2^3}, \qquad b_2 = \frac{m_4}{m_2^2}.$$

Theorem 3.12. *For any frequency distribution, $b_1 \geq 1$, $b_2 > b_1$, $b_2 \geq b_1 + 1$. Equality holds only when the variable takes two values with equal frequency.*

The absolute rth order moment about α is given by

$$\nu_r(\alpha) = \frac{1}{n} \sum_{i=1}^{n} |x - \alpha|^r.$$

For absolute central moments, we have $\nu_r = \nu_r'(\bar{x})$.

Theorem 3.13.

$$\nu_b^{a-c} \leq \nu_c^{a-b}\,\nu_a^{b-c}, \qquad a > b > c \geq 0.$$

Corollary 3.13.1.

$$\nu_{k+l}^2 < \nu_{2k}\,\nu_{2l}, \qquad m_{k+l}^2 < m_{2k}\,m_{2l}.$$

Theorem 3.14 (Liapunov)**.** *$\nu_x^{1/x}$ is increasing in x.*

4 Skew and kurtosis

4.1 Skew

Skewness is a measure of lack of symmetry in a frequency distribution. A positively skewed distribution has a longer tail to the right.

The odd central moments of a symmetric distribution are all zero for a symmetric distribution, positive for a positively skewed distribution. Thus, one measure of skewness is

$$g_1 = \frac{m_3}{m_2^{3/2}}.$$

Lemma 4.1. *For a positively skewed distribution, we have*

$$\textit{mean} > \textit{median} > \textit{mode}.$$

Thus,

$$s_k = \frac{\text{mean} - \text{mode}}{\sigma}$$

is considered a measure of skewness. We have seen the empirical relation between mean, median, mode, and standard deviation, hence we typically have

$$-3 \leq s_k \leq 3.$$

This measure s_k is called Pearson's coefficient of skewness.

For a positively skewed distribution, Q_1 is nearer to Q_2 than Q_3. Thus, Bowley's coefficient of skewness is

$$s_k = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \in [-1, 1].$$

4.2 Kurtosis

Kurtosis is a measure of the peakedness of a frequency distribution. We look at m_4, normalized as $b_2 = m_4/m_2^2$ as a measure of kurtosis.

Lemma 4.2. *For a normal distribution, $b_2 = 3$.*

Thus, $g_2 = b_2 - 3$ is also a measure of kurtosis. A distribution with $g_2 = 0$ is called *mesokurtic*, $g_2 > 0$ is called *leptokurtic*, and $g_2 < 0$ is called *platykurtic*.

5 Bivariate data

Here, out data items are in the form of points (x_i, y_i). We are typically interested in predicting the values of one of these variables (called the *study* variable) given knowledge of the other (called the *auxiliary* variable).

5.1 Correlation

Correlation is a measure of how change in one variable is associated with change in the other. Here, we only examine linear correlation. Two variables are said to be *positively* correlated if one variable increases with average increase in the other, or *negatively* correlated if one variable decreases with average increase in the other.

We define the covariance between two variables as

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

The Pearson's product moment correlation coefficient is now defined as

$$r(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \cdot \sigma_y}.$$

Lemma 5.1.

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Lemma 5.2.

$$\operatorname{Cov}(ax+by, cx+dy) = ac\sigma_x^2 + bd\sigma_y^2 + (ad+bc)\operatorname{Cov}(x,y).$$

Lemma 5.3. *The numerical value of r is invariant under shifting and scaling.*

Lemma 5.4. *We have*

$$-1 \le r(x,y) \le 1.$$

Furthermore, equality holds if and only if $(x_i - \bar{x}) = k(y_i - \bar{y})$, $k = \pm\sigma_x/\sigma_y$.

5.2 Linear regression

Given two variables X, Y, we want to relate them as

$$Y = \phi(X) + \epsilon = \beta_1 + \beta_2 X + \epsilon.$$

Here, our choice of the regression function ϕ is linear. X is called the *auxiliary* variable, and Y is called the *response* variable.

In order to determine β_1, β_2, we minimize the sum of squares of errors $\epsilon_i = y_i - (\beta_1 + \beta_2 x_i)$. This yields

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}, \qquad \hat{\beta}_2 = r_{xy}\cdot\frac{\sigma_y}{\sigma_x} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}.$$

The regression line of Y on X thus is

$$\hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

The slope b_{yx} is called the regression coefficient of Y on X. Note that we assume that X is free of error.

The regression line of X on Y is

$$\hat{x} = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y}).$$

Lemma 5.5. *If $u = (x-a)/c$, $v = (y-b)/d$, then*

$$b_{yx} = \frac{d}{c}b_{vu}, \qquad b_{xy} = \frac{c}{d}b_{uv}.$$

Lemma 5.6. *The mean of the predicted values $\hat{y}$ is equal to the mean $\bar{y}$ of the values.*

The residuals are

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \bar{y} - b_{yx}(x_i - \bar{x}).$$

Lemma 5.7. *The sum of residuals is zero.*

Lemma 5.8.
$$|r_{xy}| = \frac{\sigma_{\hat{y}}}{\sigma y}, \qquad \sigma_{\hat{\epsilon}} = \sigma_y \sqrt{1 - r^2}.$$

Thus, the coefficient of determination $r^2 = b_{yx}b_{xy}$ is a measure of the usefulness of the linear regression. If $r = 0$, then $\sigma_{\hat{\epsilon}} = \sigma_y$, making the regression pointless.

Lemma 5.9.
$$\operatorname{Cov}(\hat{y}, \hat{\epsilon}) = 0.$$

Lemma 5.10. *The angle between both regression lines satisfies*
$$\tan\theta = \left| \frac{1 - r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \right|.$$

Now suppose that our variables are empirically related as
$$y = f(x, \beta_1, \beta_2, \dots, \beta_n) = f(x; \beta).$$
Here, $\beta = (\beta_1, \dots, \beta_n)$ are unknown constants. Make an initial guess $\beta^{(0)}$, and note that we wish to minimize $\beta - \beta^{(0)}$. Taylor's theorem gives the approximation
$$y = f(x; \beta) \approx f(x; \beta^{(0)}) + (\beta - \beta^{(0)}) \cdot \left[\frac{\partial f}{\partial \beta_i}(\beta^{(0)}) \right].$$
Since this is now linear, the values $\beta - \beta^{(0)}$ can be determined using the least squares method as before. This gives us a new approximation $\beta^{(1)}$; we can repeat this process getting better and better solutions.

Suppose that both variables are subject to error. To perform a linear regression
$$ax + by + 1 = 0,$$
we minimize the sum of squares of distances from this line to the points (x_i, y_i), i.e. we want to minimize
$$\sum_{i=1}^{n} \frac{(ax_i + by_i + 1)^2}{a^2 + b^2}.$$
Setting the partial derivatives to zero yields
$$ab^2 \sum_i (x_i^2 - y_i^2) + b(b^2 - a^2) \sum_i x_i y_i + (b^2 - a^2) \sum_i x_i - 2ab \sum_i y_i = na,$$
$$a^2 b \sum_i (y_i^2 - x_i^2) + a(a^2 - b^2) \sum_i x_i y_i + (a^2 - b^2) \sum_i y_i - 2ab \sum_i x_i = nb,$$
$$a\bar{x} + b\bar{y} + 1 = 0.$$

5.3 Orthogonal polynomials

A sequence of polynomials $\{p_i\}$ is said to be orthogonal if

$$\sum_x p_i(x) \cdot p_j(x) \begin{cases} = 0, & \text{if } i \neq j \\ \neq 0, & \text{if } i = j \end{cases}.$$

Consider the polynomials

$$p_n(x) = \sum_{k=0}^{n} c_{nk} x^k.$$

When considering the first $N+1$ such polynomials, we have $(N+1)(N+2)/2$ unknown constants c_{nk}. For these to be orthogonal, we have $N(N+1)/2$ equations from the orthogonality equations. Thus, we need $N+1$ additional constraints to fully determine the coefficients. Typically, we take the coefficients of the highest power, $c_{nn} = 1$. Thus, we can calculate such orthogonal polynomials, given our data $x_1, \ldots, x_n$.

5.4 Polynomial regression

Here, our regression function is the polynomial

$$\phi(X) = \beta_0 + \beta_1 X + \cdots + \beta_k X^k.$$

Naturally, we wish to minimise the least square error

$$\sum_i (y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_k x_i^k)^2.$$

This yields $k+1$ normal equations. The estimated errors are

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \cdots - \hat{\beta}_k x_i^k.$$

These sum to zero, and their sum of squares is

$$\sum_i \hat{\epsilon}_i^2 = \sum_i y_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \cdots - \hat{\beta}_k x_i^k) = \sum_i y_i \hat{y}_i.$$

Lemma 5.11. *This residual sum of squares is a non-increasing with k.*

As usual, we will find $\operatorname{Cov}(\hat{y}, \hat{\epsilon}) = 0$, and $r^2 = \operatorname{Var}(\hat{y}) / \operatorname{Var}(y)$.

Lemma 5.12. *For different degrees k, we have $0 \leq r_k^2 \leq r_{k+1}^2 \leq 1$.*

Rewrite our kth degree polynomial in terms of orthogonal polynomials,

$$\phi = \beta_0 p_0 + \beta_1 p_1 + \cdots + \beta_k p_k.$$

Again, we solve the normal equations

$$\sum_i p_j(x_i)(y_i - \hat{\beta}_0 p_0(x_i) - \cdots - \hat{\beta}_k p_k(x_i)) = 0.$$

This immediately simplifies to

$$\sum_i y_i p_j(x_i) = \hat{\beta}_j \sum_i p_j^2(x_i).$$

It turns out that when increasing the degree of our polynomial regression, the first k coefficients $\hat{\beta}_i$ do not change!

6 Rank correlation

Suppose that u_i is the rank of the ith data point with respect to character A, and v_i is its rank with respect to character B. Rank correlation is a measure of how well these two rankings agree with each other.

6.1 Spearman's coefficient

Spearman's rank correlation coefficient is given by

$$r_R = 1 - \sum_{i=1}^{n} \frac{6d_i^2}{n(n^2-1)},$$

with $d_u = u_i - v_i$. Note that in case of ties, the group with the tie gets the average of their ranks assuming no tie.

Lemma 6.1. *Assuming no tie,*

$$\bar{u} = \bar{v} = \frac{1}{2}(n+1), \qquad \sigma_u^2 + \sigma_v^2 = \frac{1}{12}(n^2-1).$$

Lemma 6.2. *Assuming no tie,*

$$r_R = r_{uv} = \frac{\operatorname{Cov}(u,v)}{\sigma_u \sigma_v}.$$

Lemma 6.3.

$$-1 \le r_R \le 1.$$

Lemma 6.4. *For a tie of length k, the variance lowers by*

$$\frac{k(k^2-1)}{12n}.$$

6.2 Kendall's coefficient

Here, each pair of data points with indices (i,j), $i<j$, is assigned a score

$$s_{ij} = \operatorname{sgn}\left[(u_i - u_j)(v_i - v_j)\right].$$

Then, Kendall's rank correlation coefficient is

$$\tau = \frac{\text{Total score}}{\text{Maximum possible score}}.$$

In the case where there are no ties, the maximum possible score is $\binom{n}{2}$.

Note that by writing

$$a_{ij} = \operatorname{sgn}(u_i - u_j), \qquad b_{ij} = \operatorname{sgn}(v_i - v_j),$$

we have

$$\tau = \frac{\sum_{i<j} a_{ij} b_{ij}}{\sqrt{\sum_{i<j} a_{ij}^2}\sqrt{\sum_{i<j} b_{ij}^2}}.$$

Lemma 6.5. *τ may be regarded as a product moment correlation coefficient.*