Supplementary Note

Concepts of population and sample.

A random sample of n =38 cars was selected in 2009. The measured quantities are:

miles per gallon (mpg), gallons per 100 miles, weight (in 1000lbs), displacement, # of cylinders (4, 6 or 8), horsepower, acceleration, engine type (V(0) or regular(1)).

We will concentrate only on two variables: gallons per 100 miles (GPM, X) and weight (Y). Note that gallons per 100 miles is inversely proportional to miles per gallon.

The population here is all cars in the US in the year 2009, and the number of cars in the population is N, very large perhaps in many millions. Let us assume that the values of weight and GPM in the population are $(x_1, y_1), \ldots (x_N, y_N)$. We may never know all these N pairs of values. However, the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $n = 38$ may provide some estimates about some crucial population summaries such as the population means, variances, covariance and correlation. Note that here weight is the independent variable and GPM is the independent variable.

Population means, variances and standard deviations (SD).

The population means of weight and GPM are

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \mu_Y = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

The population variances of weights and GPM are

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2,\ \sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu_Y)^2.$$

Note that the populations standard deviations of weight and GPM are σ_X and σ_Y respectively (positive square roots of σ_X^2 and σ_Y^2).

If a car is randomly taken from the population and its weight and GPM are (X, Y), then we say that X and Y are random variables. We employ the following notations to describe the populations means and variances

$$E(X) = \mu_X, E(Y) = \mu_Y, Var(X) = \sigma_X^2, Var(Y) = \sigma_Y^2.$$

Sample means, variances and standard deviations.

The mean weight and mean GPM of the sample cars are

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 2.863,\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 4.331.$$

Statistics tells us $\bar{X}$ and $\bar{Y}$ estimate μ_X and μ_Y respectively. In many textbooks, $\bar{X}$ and $\bar{Y}$ are denoted by $\hat{\mu}_X$ and $\hat{\mu}_Y$ respectively. Note that $\bar{X}$ and $\bar{Y}$ vary from sample to sample, but μ_X and μ_Y are fixed as they do not depend on the sample.

The variance of weights and GPM of the sample cars are

$$\begin{aligned}
\hat{\sigma}_X^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2 = 0.4865167, \\
\hat{\sigma}_Y^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_Y)^2 = 1.301173, \\
\hat{\sigma}_X &= 0.6975,\ \hat{\sigma}_Y = 1.1407.
\end{aligned}$$

Note that $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ are estimate of σ_X^2 and σ_Y^2 respectively. Similarly, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are estimate of σ_X and σ_Y respectively.

[**An important note: In the textbooks, the divisor is usually $n-1$ instead of n in the formulas for $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$. And there is a good reason for it. Division by $n-1$ makes the sample variances unbiased estimates of the population variances. However, if n is large, division by n or $n-1$ makes little difference in the estimated values.**]

Population covariance and correlation.

Common sense tells us that there should be an increasing relation between weight and GPM, and a scatterplot of the data indicates this. Not only that, the relation seems to be linear. There are measures called 'covariance' and 'correlation' which capture this relationship

$$\begin{aligned}
\sigma_{XY} &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y) \text{ (covariance between weight and GPM)}, \\
\rho_{XY} &= \frac{\sigma_{XY}}{\sqrt{\sigma_X^2\sigma_Y^2}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \text{ (correlation between weight and GPM)}.
\end{aligned}$$

Note that $\sigma_{XY} = \sigma_{YX}$ and $\rho_{XY} = \rho_{YX}$.

Since GPM seems to increase with weight, both σ_{XY} and ρ_{XY} are positive. A decreasing relation will lead to negative values of σ_{XY} and ρ_{XY}. In other words, the sign of σ_{XY} and ρ_{XY} indicate the nature of relationship between weight and GPM. Note however that the covariance is not unit free, but the correlation is. If the values of weights and GPM are converted into kilograms and liters respectively, the value of the covariance will change, but the value of the correlation remain the same, ie, the correlation is unit free.

Variables X and Y are said to to be **uncorrelated** if $Cov(X,Y) = 0$ (and hence $Corr(X,Y) = 0$ assuming that $Var(X) > 0$ and $Var(Y) > 0$). If variables X and are independent, then $Cov(X,Y) = 0$ (and $Corr(X,Y) = 0$). But the converse is not true except when (X,Y) are jointly normally distributed. If (X,Y) are normally distributed, then the concepts of uncorrelatedness and independence are the same.

Here are some properties of the correlation.

Fact 1: (a) ρ_{XY} is unit free, (b)$-1 \le \rho_{XY} \le 1$,
(c) $\rho_{XY} = 1$ when and only when the plot of y's against x's would be exactly a straight line with a positive slope,
(d) $\rho_{XY} = -1$ when and only when the plot of y's against x's would be exactly a straight line with a negative slope,

(e) If $\rho_{XY} = 0$, then there is no linear relation between x's and y's. However, a nonlinear relation may still be present.

Population covariance and correlation.

For the car data we have $n = 38$, and we can calculate the sample covariance and sample correlation as

$$\begin{aligned} \hat{\sigma}_{XY} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y) \ = 0.7369742, \\ \hat{\rho}_{XY} &= \frac{\hat{\sigma}_{XY}}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}} = \frac{0.7369742}{\sqrt{(0.4865167)(1.301173)}} = 0.9263. \end{aligned}$$

The sample correlation is quite high, close to 0.93, and this is consistent with the scatterplot. The sample covariance and sample correlation are estimates of the population covariance and population correlation respectively. The sample correlation also unit-free and its value is between -1 and 1.

[**An important note: In the textbooks, the divisor for $\hat{\sigma}_{XY}$ is usually $n-1$ instead of n in the formula above. Division by $n-1$ makes the sample covariance an unbiased estimates of the population covariance. However, if n is large, division by n or $n-1$ makes little difference in the estimated values.**]

Mathematical Notations and some results.

If a car is selected randomly and its weight and GPM are (X, Y), then we can take X and Y to be random variables. Then σ_{XY} is denoted by $Cov(X, Y)$, and ρ_{XY} is denoted by $Corr(X, Y)$. Thus we have

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

Let X, Y, Z etc. denote random variables, and let $c_1, c_2, c_3, d_1, d_2, d_3$ etc. denote constants, say $c_1 = 1.5, c_2 = 0.9, c_3 = -1.1$ etc. The following mathematical results are useful.

Fact 2: Properties of the mean

$$\begin{aligned} E(c_1X + d_1) &= c_1E(X) + d_1, \\ E(c_1X + c_2Y) &= c_1E(X) + c_2E(Y), \text{ and in general} \\ E\left(\sum c_iU_i\right) &= \sum c_iE(U_i), \end{aligned}$$

where $U_1, U_2, \ldots$ are random variables.

Fact 3: Properties of variances:
(a) $Cov(X, X) = Var(X)$,
(a) $Var(X) = 0$ when and only when X is a constant,
(b) $Var(c_1X + d_1) = c_1^2Var(X)$,
(c) $Var(c_1X + c_2Y) = c_1^2Var(X) + c_2^2Var(Y) + 2c_1c_2Cov(X, Y)$,

(d) Analogue of (c):

$$\begin{aligned} & Var(c_1X + c_2Y + c_3Z) \\ = \; & c_1^2 Var(X) + c_2^2 Var(Y) + c_3^2 Var(Z) + \\ & 2c_1c_2 Cov(X,Y) + 2c_1c_3 Cov(X,Z) + 2c_2c_3 Cov(Y,Z). \end{aligned}$$

(e) A general formula for variances: If $U_1, U_2, \ldots$ are random variables, then

$$\begin{aligned} Var\left(\sum c_iU_i\right) &= \sum_i \sum_j c_ic_j Cov(U_i, U_j) \\ &= \sum c_i^2 Var(U_i) + \sum_i \sum_{j<i} c_ic_j Cov(U_i, U_j), \end{aligned}$$

(f) If $U_1, U_2, \ldots$ are mutually uncorrelated, then it follows from part (e) that

$$Var\left(\sum c_iU_i\right) = \sum c_i^2 Var(U_i).$$

Fact 4: Properties of covariances:
(a) $Cov(X, d_1) = 0$,
(b) $Cov(c_1X + d_1, c_2Y + d_2) = c_1c_2Cov(X,Y)$,
(c) $Cov(X + Y, Z) = Cov(X,Z) + Cov(Y,Z)$,
(d) A generalization of the formula in part (c) is

$$\begin{aligned} & Cov(X + Y, Z + U) \\ = \; & Cov(X + Y, Z) + Cov(X + Y, U) \\ = \; & Cov(X,Z) + Cov(Y,Z) + Cov(X,U) + Cov(Y,U), \end{aligned}$$

(e) A general formula for covariances: If $U_1, U_2, \ldots$ and $V_1, V_2, \ldots$ are random variables, then

$$Cov\left(\sum c_iU_i, \sum d_jV_j\right) = \sum_i \sum_j c_id_j Cov(U_i, V_j).$$

Examples.
1. Let $c_1 = 1.1, c_2 = 0.5$, and let U_1, U_2 be random variables with $E(U_1) = 5, E(U_2) = 10, Var(U_1) = 2, Var(U_2) = 3$ and $Cov(U_1, U_2) = 1.5$. Then

$$\begin{aligned} E(c_1U_1 + c_2U_2) &= c_1E(U_1) + c_2E(U_2) = (1.1)(5) + (0.5)(10) = 10.5, \\ Var(c_1U_1 + c_2U_2) &= \sum c_i^2 Var(U_i) + 2\sum_i \sum_{j<i} c_ic_j Cov(U_i, U_j) \\ &= c_1^2 Var(U_1) + c_2^2 Var(U_2) + 2c_1c_2 Cov(U_1, U_2) \\ &= (1.1)^2(2) + (0.5)^2(3) + 2(1.1)(0.5)(1.5) = 4.82. \end{aligned}$$

2. Let U_1, U_2, V_1, V_2 be random variables with $Cov(U_1, V_1) = Cov(U_2, V_2) = 1$, and $Cov(U_1, V_2) =$

$Cov(U_2, V_1) = 0.3$. Then

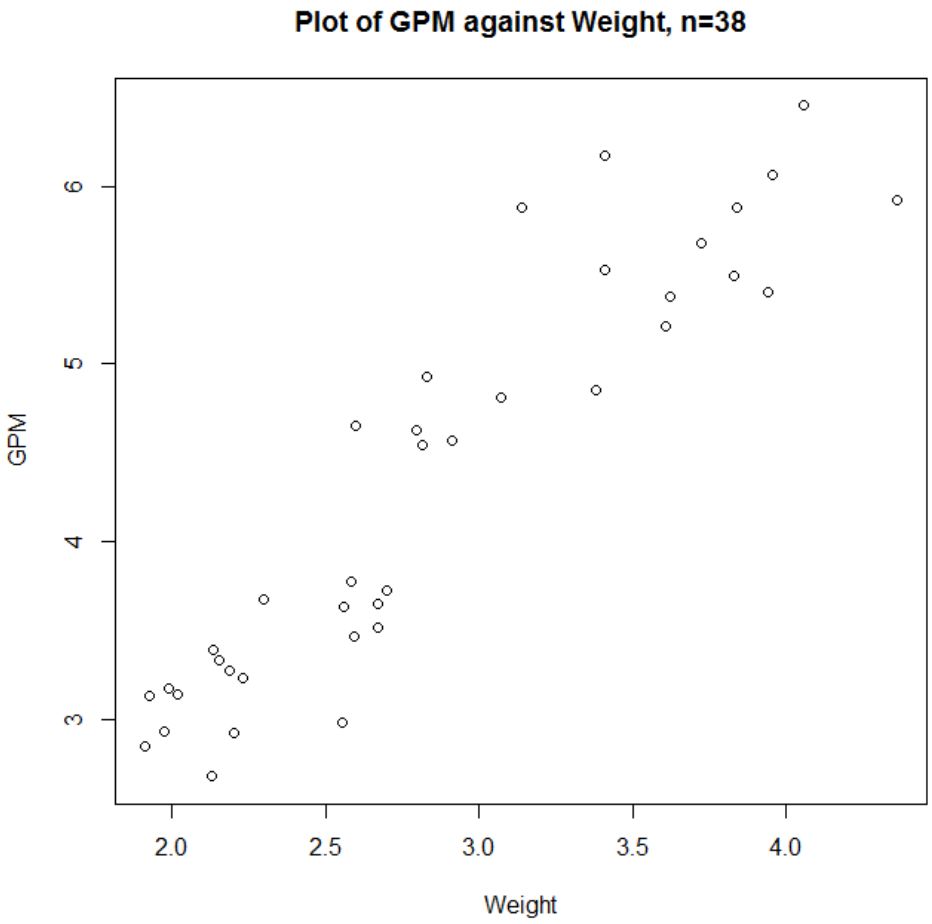
$$\begin{aligned} & Cov(U_1 + U_2, V_1 + V_2) \\ = & \sum_i \sum_j c_i d_j Cov(U_i, V_j) \text{ [with } c_i \equiv 1,\ d_j \equiv 1] \\ = & Cov(U_1, V_1) + Cov(U_1, V_2) + Cov(U_2, V_1) + Cov(U_2, V_2) \\ = & 1 + 0.3 + 0.3 + 1 = 2.6. \end{aligned}$$

3. Let U_1, U_2, V_1 and V_2 be the same as in Example 2. Let $c_1 = 0.5, c_2 = 0.5, d_1 = 1, d_2 = -1$. Then

$$\begin{aligned} & Cov(c_1 U_1 + c_2 U_2, d_1 V_1 + d_2 V_2) \\ = & \sum_i \sum_j c_i d_j Cov(U_i, V_j) \\ = & c_1 d_1 Cov(U_1, V_1) + c_1 d_2 Cov(U_1, V_2) + c_2 d_1 Cov(U_2, V_1) + c_2 d_2 Cov(U_2, V_2) \\ = & (0.5)(1)(1) + (0.5)(-1)(0.3) + (0.5)(1)(0.3) + (0.5)(-1)(1) = 0. \end{aligned}$$

Car Data: We have a random sample of $n = 38$ cars in 2009. Only a part of the data is given below.

MPG	GPM	Weight	Displacement	# of Cylinders	HP	Accelaration	EngineType
16.9	5.917	4.360	350	8	155	14.9	1
15.5	6.452	4.054	351	8	142	14.3	1
19.2	5.208	3.605	267	8	125	15.0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21.5	4.651	2.600	121	4	110	12.8	0
31.9	3.135	1.925	89	4	71	14.0	0


Plot of GPM against Weight, n=38
GPM
Weight
3
4
5
6
2.0
2.5
3.0
3.5
4.0