CHAPTER 2
BASIC IDEAS OF LINEAR REGRESSION: THE TWO-VARIABLE MODEL

QUESTIONS

2.1. (***a***) It states how the population *mean* value of the dependent variable is related to one or more explanatory variables.

(***b***) It is the sample counterpart of the PRF.

(***c***) It tells how the individual *Y* are related to the explanatory variables and the stochastic error term, *u*, in the population as a whole.

(***d***) A model that is linear in the parameters, the *B*s.

(***e***) It is a proxy for all omitted or neglected variables that affect the dependent variable *Y*. The individual influence of each of these variables is random and small so that on average their influence on *Y* is zero.

(***f***) It is the sample counterpart of the stochastic error term.

(***g***) The expected value of *Y* conditional upon a given value of *X*. It is obtained from the conditional (probability) distribution of Y, given *X*.

(***h***) The expected value of an r.v. regardless of the values taken by other random variables. It is obtained from the unconditional, or marginal, probability distributions of the relevant random variables.

(***i***) The *B* coefficients in a linear regression model are called regression coefficients or regression parameters.

(***j***) The *b*s, which tell how to compute the *B*s, are called the estimators. Numerical values taken by the *b*s are known as estimates.

2.2. A stochastic SRF tells how Y_i in a randomly drawn sample from a *Y* population are related to the explanatory variables and the residuals e_i. A stochastic PRF tells how the individual Y_i are related to the explanatory variables and the stochastic error term u_i in the whole population.

2.3. The PRF is a theoretical, or idealized, model, just as the model of perfect competition is an idealized model. But such idealized models help us to see the essence of the problem.

2.4. **(*a*)** *False*. The residual e_i is an approximation (i.e., an estimator) of the true error term, u_i.

(*b*) *False*. It gives the *mean* value of the dependent variable, given the values of the explanatory variables.

(*c*) *False*. A linear regression model is *linear in the parameters* and not necessarily linear in the variables.

(*d*) *False*, generally. The cause and effect relationship between the *X*s and *Y* must be justified by theory.

(*e*) *False*, unless the "conditioned" and conditioning variables are independent.

(*f*) *False*. It is the other way around.

(g) *False*. It measures the change in the *mean* value of Y per unit change in *X*.

(*h*) *Uncertain*. There are many a phenomena which can be explained by the two-variable model. One example is the *Market Model* of portfolio theory which regresses the rate of return on a single security on the rate of return on a market index (e.g., S&P 500 stock index). The slope coefficient in this model, popularly known as the *beta coefficient*, is used extensively in portfolio analysis.

(*i*) *True*.

2.5. **(*a*)** b_1 is an estimator of B_1.

(*b*) b_2 is an estimator of B_2.

(*c*) e_i is an estimator of u_i.

We never observe B_1, B_2, and *u*. Once we have a specific sample, we can obtain their estimates via b_1, b_2 and *e*.

2.6. By simple algebra, we obtain:

$$X_t = 2.5 - 2.5Y_t$$

Sometimes Okun's model is run in this format, regressing percent growth in real output on the change in the unemployment rate.

2.7. (***a***) The answer will depend on how the various components of GDP (consumption expenditure, investment expenditure, government expenditure and expenditure on net exports) react to the higher interest rate. For instance, *ceteris paribus*, investment expenditure and the interest rate are inversely related.

(***b***) Positive. *Ceteris paribus*, the higher the interest rate is, the greater will be the incentive to save.

(***c***) Generally positive.

(***d***) Positive, to maintain at least the status quo.

(***e***) Probably positive.

(***f***) Probably negative; familiarity may breed contempt.

(***g***) Probably positive.

(***h***) Positive. Statistics is a major foundation of econometrics.

(***i***) Positive. As income increases, discretionary income is likely to increase, leading to an increased demand for more expensive cars. A large number of Japanese cars are expensive. In general, the income elasticity of demand for items like cars has been found to be not only positive but generally greater than 1.

PROBLEMS

2.8. (***a***) Yes (***b***) Yes (***c***) Yes (***d***) Yes (***e***) No (***f***) No.

2.9 (***a***) The conditional expected values are:

Value of *X*	***E*(*Y* | *X*)**	**Value of *X***	***E*(*Y* | *X*)**
80	65	180	125
100	77	200	137
120	89	220	149
140	101	240	161
160	113	260	173

(***b***) and (***c***). This is straightforward.

(***d***) The mean of *Y* increases with *X*. That may not be true of the individual *Y* values.

(***e***) PRF: $Y_i = B_1 + B_2 X_i + u_i$

SRF: $Y_i = b_1 + b_2 X_i + e_i$

(***f***) The scatter plot will show that the PRF is linear.

2.10. (***a***) This is straightforward.

(***b***) The relationship between the two is positive.

(***c***) SRF: $\hat{Y}_i = 24.4545 + 0.5091\, X_i$

The raw data give: $\sum Y_i = 1{,}110$; $\sum X_i = 1{,}700$; $\sum x_i^2 = 33{,}000$;

$\sum x_i y_i = 16{,}800$, where the small letters denote deviations from the mean values.

(***d***) This is straightforward.

(***e***) The two are close, but obviously they are not identical.

2.11. (***a***) From the time subscript *t*, it seems that this is a time series regression.

(***b***) The regression line is linear with a negative slope.

(***c***) The average number of cups of coffee consumed per person per day if the price of coffee were zero. Economically speaking, this may or may not make sense.

(***d***) *Ceteris paribus*, the *mean* consumption of coffee per day goes down by about 1/2 cup a day as the price of coffee per pound increases by a $1.

(***e***) No. But with the confidence interval procedure discussed in the next chapter, it is possible to tell, in probabilistic terms, what the PRF may be.

(***f***) We have information on the slope coefficient, but not on *X* and *Y*. Therefore, we cannot compute the price elasticity coefficient from the given information.

2.12. (***a***) and (***b***). The scattergram will show that the relationship between the S&P 500 index and the CPI is positive.

(***c***) $(\hat{S\&P})_t = -195.5149 + 3.8264\ \text{CPI}_t$

These results show that on average S&P goes up by about 3.8 points per unit increase in the CPI. The constant term suggests that if the value of the CPI were zero, the mean value of S&P would be about -195.

Note: This example is further examined in problem 6.15.

(***d***) The positive slope may make economic sense, but the negative intercept value may not.

(***e***) Most probably it was due to the October 1987 stock market crash.

2.13. (***a***) The scattergram will show a positive relationship between the nominal interest rate and the inflation rate, as per economic theory (the so-called *Fisher effect*). Notice that there is an extreme observation, called an *outlier*, pertaining to Mexico.

(***b***) $\hat{Y}_i = 2.7131 + 1.2320\, X_i$

(***c***) The value of the slope coefficient is expected to be 1, because, according to the Fisher equation, the following relationship holds true approximately: nominal interest rate = expected real interest rate + expected inflation rate. Thus, the intercept in the Fisher equation is the expected real rate of interest. In the present example, we cannot tell whether the Fisher equation holds because the inflation rate used is the actual inflation rate. In terms of the actual inflation rate, the nominal rate, on average, seems to increase more than one percent for a one percent increase in the (actual) inflation rate, for the slope coefficient is 1.2320. Applying the techniques discussed in the next chapter, this slope coefficient is statistically significantly greater than 1.

2.14. (***a***) This is straightforward.

(***b***) $\hat{\text{NE}}_{\text{US}} = 0.0088 + 1.1274\ \text{RE}_{\text{US}}$

(***c***) Positive.

(***d***) Yes.

(***e***) $\ln \hat{\text{NE}}_{\text{US}} = 0.1233 + 1.0034 \ln \text{RE}_{\text{US}}$

Yes, the results are qualitatively the same. But note that the slope coefficient in the double-log model represents the elasticity coefficient, whereas that in the linear model represents the absolute rate of change in the

(mean) value of NE_{US} for a unit change in RE_{US}. See Chapter 5 for the various functional forms.

2.15 (***a***) Repeating the five questions, we have:

- The scattergram is straightforward.
- As before, the relationship between the two is expected to be positive.
- The regression equation for the 1990-2007 period is:

$$\widehat{(S\&P)}_t = -1611.5024 + 15.0550\ CPI_t$$

- The positive slope makes economic sense but the intercept does not.
- The 1988 S&P decline is not applicable here.

(***b***) The results are in accord with prior expectations, although numerical values of the two period regression coefficients are vastly different.

(***c***) Combining the two data sets, we get the following results:

$$\widehat{(S\&P)}_t = -906.8409 + 10.8914\ CPI_t$$

(***d***) Since the regression results of the two sub-periods are different (which can be proved using the dummy variable technique discussed in Chapter 6 or by the Chow test), the preceding regression results that are based on the pooled data are not meaningful.

2.16. (***a***) ASP = - 88,220.4947 + 55,227.4336 GPA

It seems GPA has a positive impact on ASP.

(***b***) ASP = - 241,386.602 + 511.721 GMAT

GMAT also seems to have a positive impact on ASP.

(***c***) ASP = 42,878.332 + 1.635 TUITION

Tuition also seems to have positive impact on ASP.

Top business schools generally have top teachers and researchers. This means that these schools have to pay higher salaries to attract quality faculty. In this sense high tuition may be a proxy for high quality education, which may result in higher ASP for graduates from such schools.

(***d***) ASP = -29,943.604 + 37,300.297 RECRUITER

This positive relationship suggests that recruiter perception has a positive bearing on ASP.

Note: In the next chapter we will see if the regressions presented above are statistically significant.

2.17. **(*a*)** Given the formulation of Okun's law in Equation (6.22), the new variables based on the real GDP (RGDP) and the unemployment rate (UNRATE) data from Table 2-13 can be calculated as follows:

CHUNRATE = Change in UNRATE = UNRATE – UNRATE(-1)

PCTCRGDP = % Change in RGDP = [RGDP / RGDP(-1)]*100-100

Note: UNRATE – UNRATE(-1) means subtracting the previous period's unemployment rate from the current period's unemployment rate. For example, looking at the first two observations, UNRATE – UNRATE(-1) = 6.7 – 5.5, and so on. Similarly for RGDP and RGDP(-1), except in this case we divide by the previous period's observation.

The regression equation is:

$$\widehat{\text{CHUNRATE}} = 1.2334 - 0.3734\ \text{PCTCRGDP}$$

The slope coefficients in the two regressions are about the same. If you simplify (2.22), the result is: CHUNRATE = 1.00 – 0.40 PCTCRGDP.

Therefore, the intercepts in the two regressions are about the same. Perhaps Okun's law may have some universal validity.

(*b*) Reversing the roles of CHUNRATE and PCTCRGDP, we have:

$$\widehat{\text{PCTCRGDP}} = 3.3191 - 1.8630\ \text{CHUNRATE}$$

For a unit change in CHUNRATE, real GDP growth changes by about 1.86 percent in the opposite direction.

(*c*) If CHUNRATE in **(*b*)** is zero, real GDP growth is about 3.3%. We may interpret this as the natural rate of growth in real GDP. In the original Okun model it was assumed to be about 2.5%, the growth rate then prevailing.

2.18. **(*a*)** Straightforward. Any minor differences may be solely due to rounding issues.

(***b***) For model (2.24), the output is as follows:

obs	Actual	Fitted	Residual	Residual Plot
1980	118.780	491.045	-372.265	| .* | . |
1981	128.050	475.464	-347.414	| .* | . |
1982	119.710	497.694	-377.984	| .* | . |
1983	160.410	519.918	-359.508	| .* | . |
1984	160.460	508.464	-348.004	| .* | . |
1985	186.840	537.677	-350.837	| .* | . |
1986	236.340	571.107	-334.767	| .* | . |
1987	286.830	575.689	-288.859	| . * | . |
1988	265.790	553.415	-287.625	| . * | . |
1989	322.840	527.173	-204.333	| . * | . |
1990	334.590	537.145	-202.555	| . * | . |
1991	376.180	588.330	-212.150	| . * | . |
1992	415.740	693.353	-277.613	| . * | . |
1993	451.410	734.495	-283.085	| . * | . |
1994	460.420	636.776	-176.356	| . * | . |
1995	541.720	585.326	-43.6060	| . *| . |
1996	670.500	602.985	67.5145	| . |* . |
1997	873.430	601.027	272.403	| . | * . |
1998	1085.50	611.655	473.845	| . | .* |
1999	1327.33	618.326	709.004	| . | . * |
2000	1427.22	574.811	852.409	| . | . *|
2001	1194.18	693.353	500.827	| . | .* |
2002	993.940	1019.76	-25.8158	| . * . |
2003	965.230	1381.73	-416.496	| * | . |
2004	1130.65	1126.77	3.87696	| . * . |
2005	1207.23	719.870	487.360	| . | .* |
2006	1310.46	615.160	695.300	| . | . * |
2007	1477.19	630.453	846.737	| . | . *|

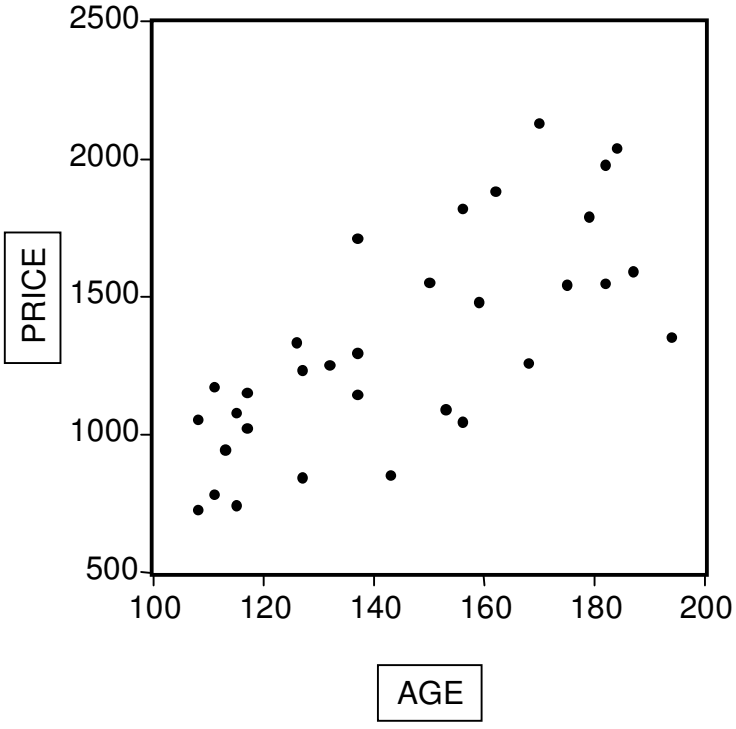
For model (2.25) the output is:

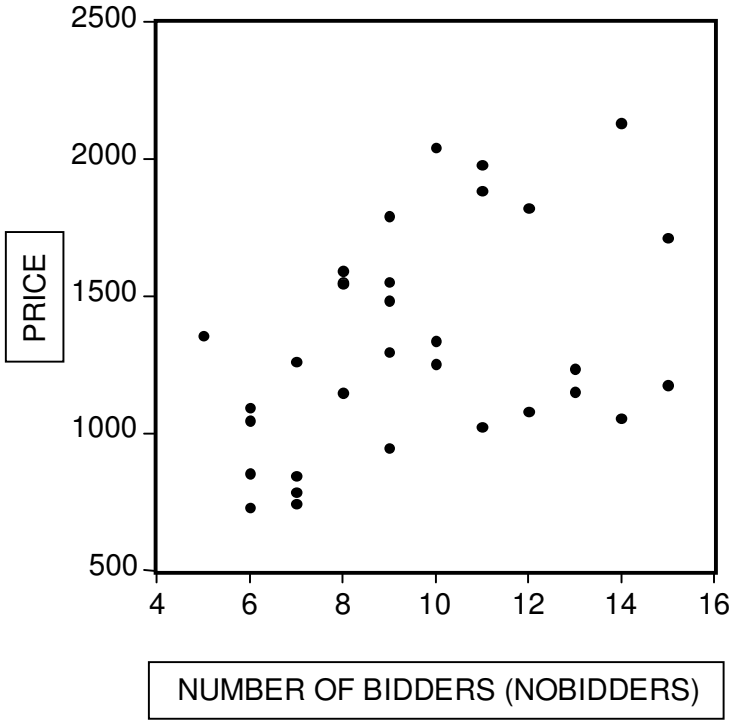
obs	Actual	Fitted	Residual	Residual Plot
1980	118.780	85.6288	33.1512	| . |* . |
1981	128.050	-165.161	293.211	| . | *. |
1982	119.710	167.138	-47.4280	| . *| . |
1983	160.410	371.507	-211.097	| . * | . |
1984	160.460	277.076	-116.616	| . * | . |
1985	186.840	485.819	-298.979	| .* | . |
1986	236.340	634.921	-398.581	| * | . |
1987	286.830	650.825	-363.995	| * | . |
1988	265.790	564.346	-298.556	| .* | . |
1989	322.840	422.202	-99.3620	| . * | . |
1990	334.590	482.837	-148.247	| . * | . |

1991	376.180	690.586	-314.406	| .* | . |
1992	415.740	886.407	-470.667	| *. | . |
1993	451.410	929.149	-477.739	| *. | . |
1994	460.420	802.909	-342.489	| .* | . |
1995	541.720	681.640	-139.920	| . * | . |
1996	670.500	730.346	-59.8463	| . *| . |
1997	873.430	725.376	148.054	| . |* . |
1998	1085.50	751.221	334.279	| . | *. |
1999	1327.33	766.131	561.199	| . | . * |
2000	1427.22	647.843	779.377	| . | . *|
2001	1194.18	886.407	307.773	| . | *. |
2002	993.940	1068.31	-74.3711	| . *| . |
2003	965.230	1127.95	-162.722	| . * | . |
2004	1130.65	1092.17	38.4826	| . |* . |
2005	1207.23	915.233	291.997	| . | *. |
2006	1310.46	759.173	551.287	| . | . * |
2007	1477.19	790.981	686.209	| . | . * |

The residual plots of the two models seem similar. To choose between the two models, we need model selection criteria, discussed in Chapter 7.

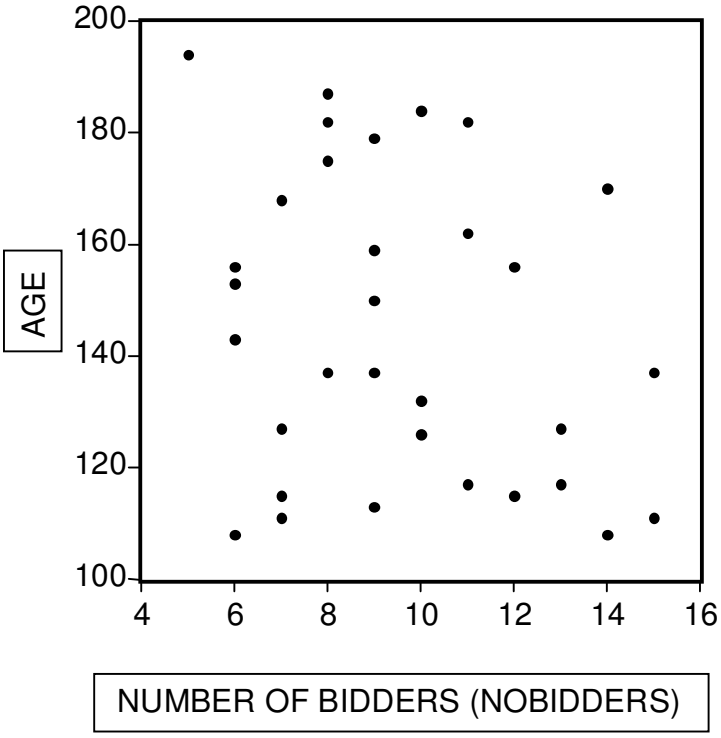
2.19 ***(a)*** The graphs are as follows:





This graph shows that the higher the number of bidders, the higher the price is. This probably is true of the antique clock auction market. As a first approximation, the linear model may be appropriate for the price/ age relationship, but may not be quite appropriate for the price/number of bidders relationship.

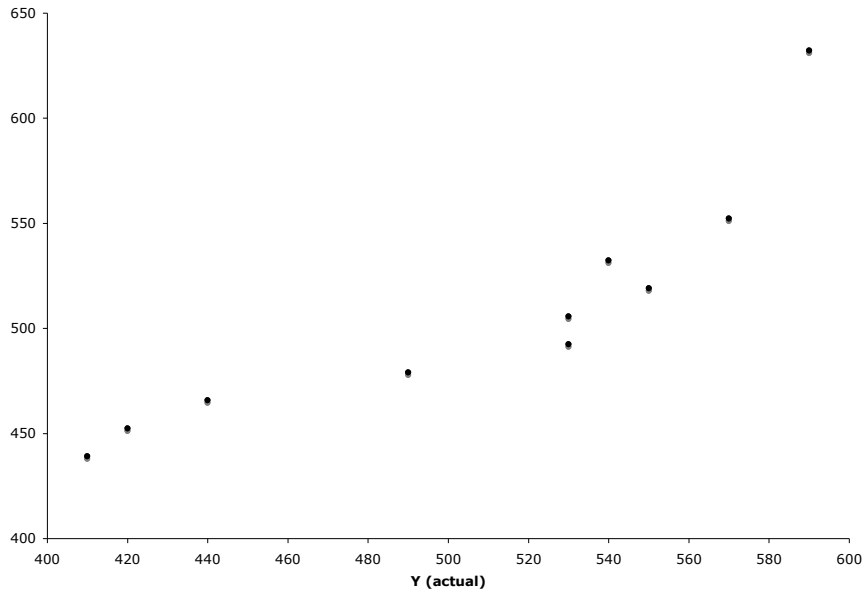
(*b*) The plot of the number of bidders versus age is as follows:



This scatter plot shows a very weak negative relationship between clock age and the number of bidders. This is most likely because, the higher the clock age, the higher the price. There will be fewer people able to bid for the older, more expensive clocks.

2.20. The scatter plot between actual Y (data from Table 2.4) and estimated $\hat{Y}$ values is as follows:

(Graph appears on the following page)



If the fitted model is a good one, the actual and estimated Y values should be very close to each other. In the case where the model is a perfect fit, the scatter points will lie on a straight line.

2.21. **(*a*)** MaleMath = 511.607 + 0.0259 MaleCR

(*b*) This regression suggests that as the male critical reading score goes up by a unit, on average, the male math score goes up by about 0.0259 units.

(*c*) MaleCR = 499.734 + 0.0196 MaleMath

As per this regression if the male math score goes up by a unit, the average male verbal score goes up by about 0.0196 units.

(*d*) If you multiply the slope coefficients in the two preceding equations, you will obtain: (0.0259)(0.0196) = 0.0005

As we show in the next chapter, the r^2 value, which is a measure of how good a chosen regression line fits the actual data, for either of the preceding regressions is 0.0005, which is precisely equal to the product of the slope coefficients in the two preceding regressions. The point to note here is that in a bivariate regression, if we regress Y on X or vice versa, the r^2 value remains the same.

OPTIONAL QUESTIONS

2.22. $\sum e_i = \sum (Y_i - b_1 - b_2 X_i)$

$= n\bar{Y} - \sum(\bar{Y} - b_2\bar{X}) - b_2 \sum X_i \quad [\textit{Note}: b_1 = \bar{Y} - b_2\bar{X}\,]$

$= n\bar{Y} - n\bar{Y} + b_2\, n\,\bar{X} - b_2\, n\,\bar{X} = 0$

2.23. $\sum e_i X_i = \sum (Y_i - b_1 - b_2 X_i) X_i$

$= \sum Y_i X_i - b_1 \sum X_i - b_2 \sum X_i^2$

$= 0$, because of Equation (6.15).

2.24. $\sum e_i \hat{Y}_i = \sum e_i (b_1 + b_2 X_i)$

$= b_1 \sum e_i + b_2 \sum e_i X_i = 0$, using problems (2.22) and (2.23) above.

2.25. Since $Y_i = \hat{Y}_i + e_i$, summing over both sides over the sample, we obtain:

$$\sum Y_i = \sum \hat{Y}_i + \sum e_i$$

Dividing both sides by *n*, we obtain:

$$\sum Y_i / n = \sum \hat{Y}_i / n + \sum e_i / n$$

Since the last term in this equation is zero (why?), the result follows.

2.26. $\sum x_i y_i = \sum x_i (Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y} \sum x_i = \sum x_i Y_i$, since $\bar{Y}$ is a constant and since $\sum x_i = \sum (X_i - \bar{X}) = 0$, as shown in Equation (2.17). The other expressions in this problem can be derived similarly.

2.27. $\sum x_i = \sum (X_i - \bar{X}) = \sum X_i - n\,\bar{X}$, since $\bar{X}$ is a constant

$= n\,\bar{X} - n\,\bar{X} = 0$ since $\bar{X} = \sum X_i / n$

A similar result hold for $\sum y_i$.

It is worth remembering that the sum of deviations of a random variable from its mean value is always zero.

2.28. It is a simple matter of verification, save the rounding errors