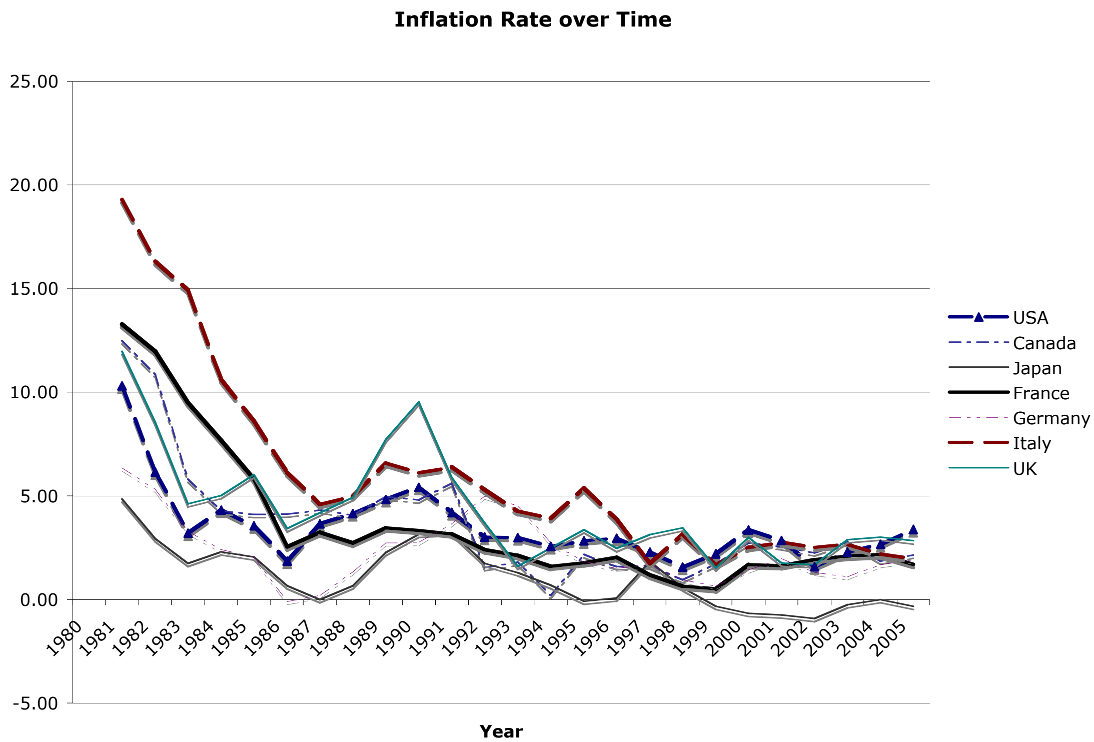
CHAPTER 1:
THE NATURE OF REGRESSION ANALYSIS

1.1 (a) These rates (%) are as follows. They are year-over-year, starting with 1981.

	USA	Canada	Japan	France	Germany	Italy	UK
1980							
1981	10.32	12.48	4.84	13.28	6.34	19.30	11.97
1982	6.16	10.86	2.94	11.97	5.31	16.31	8.53
1983	3.21	5.80	1.73	9.49	3.30	14.94	4.61
1984	4.32	4.28	2.30	7.67	2.39	10.62	5.01
1985	3.56	4.11	2.06	5.83	2.04	8.61	6.01
1986	1.86	4.13	0.67	2.53	-0.10	6.11	3.42
1987	3.65	4.32	0.00	3.24	0.19	4.59	4.18
1988	4.14	4.05	0.67	2.73	1.33	4.99	4.93
1989	4.82	4.95	2.27	3.46	2.73	6.59	7.72
1990	5.40	4.80	3.15	3.34	2.75	6.12	9.53
1991	4.21	5.61	3.23	3.16	3.65	6.39	5.87
1992	3.01	1.54	1.74	2.41	4.99	5.30	3.70
1993	2.99	1.79	1.28	2.14	4.50	4.25	1.60
1994	2.56	0.20	0.68	1.60	2.74	3.92	2.48
1995	2.83	2.16	-0.08	1.78	1.83	5.37	3.36
1996	2.95	1.59	0.08	2.02	1.50	3.87	2.46
1997	2.29	1.63	1.84	1.19	1.70	1.75	3.12
1998	1.56	0.96	0.58	0.65	0.94	3.15	3.46
1999	2.21	1.71	-0.33	0.52	0.65	1.66	1.52
2000	3.36	2.74	-0.66	1.68	1.43	2.52	2.99
2001	2.85	2.55	-0.74	1.65	1.97	2.76	1.75
2002	1.58	2.25	-0.92	1.94	1.31	2.52	1.67
2003	2.28	2.78	-0.25	2.08	1.09	2.66	2.90
2004	2.66	1.86	0.00	2.16	1.69	2.19	3.00
2005	3.39	2.15	-0.34	1.70	1.92	1.95	2.83

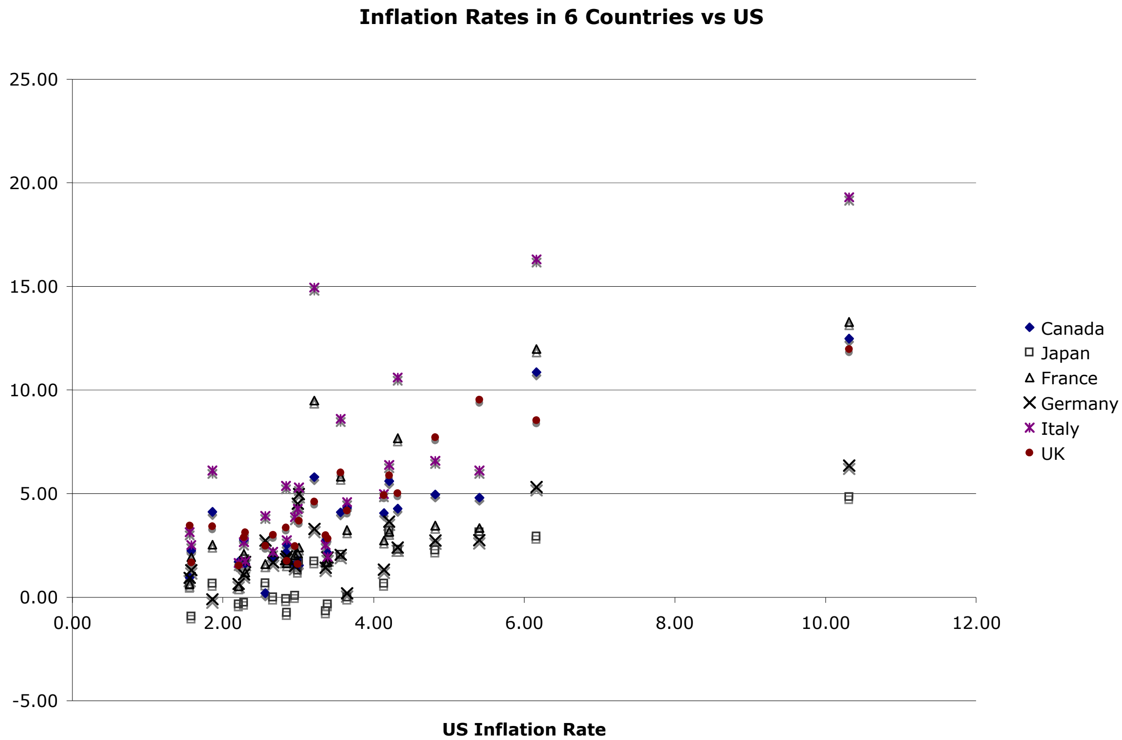
(b)



(c) As you can see from this figure, the inflation rate of each of the countries has *generally* declined over the years.

(d) As a measure of variability, we can use the standard deviation. These standard deviations are 1.81, 2.85, 1.49, 3.40, 1.60, 4.70, and 2.65, respectively, for the US, Canada, Japan, France, Germany, Italy, and the UK. The highest variability is thus found for Italy and the lowest for Japan.

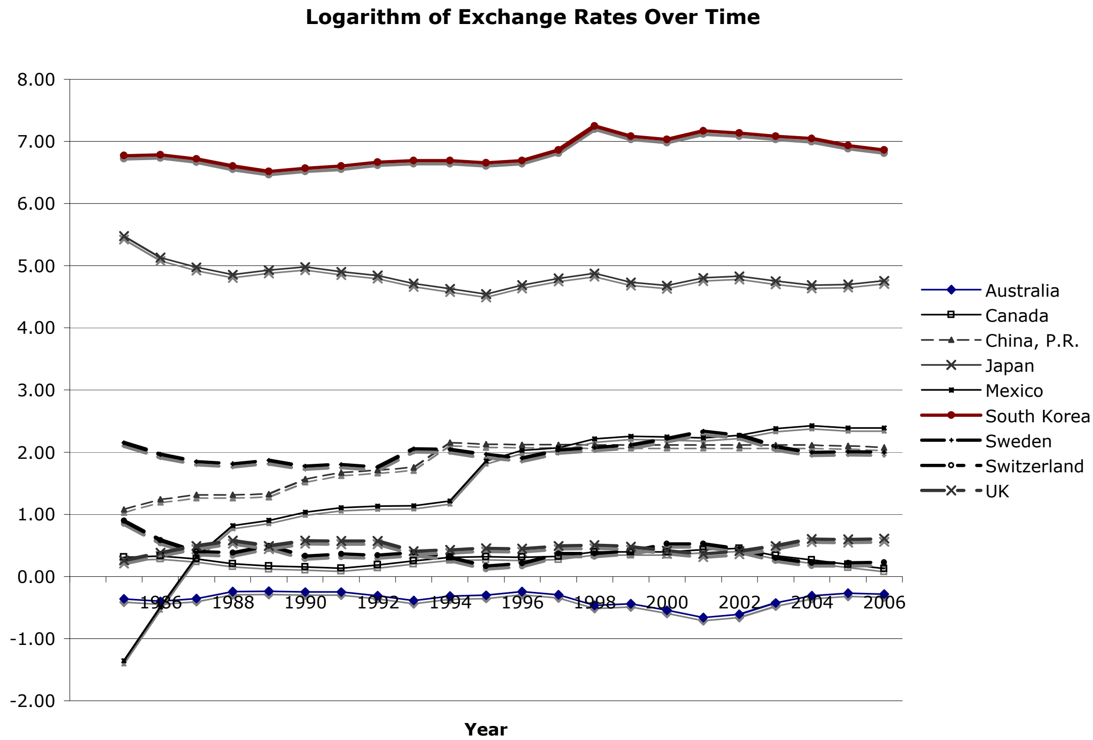
1.2. (a) The graph of the inflation rates of the six countries plotted against the US inflation rate is as follows:



(b) As the figure shows, in general the inflation rates of the six countries are positively correlated with the US inflation rate.

(c) Remember that correlation does not mean causation. One may have to consult a book on international macroeconomics to find out if there is any causal connection between the US and the other countries' inflation rates.

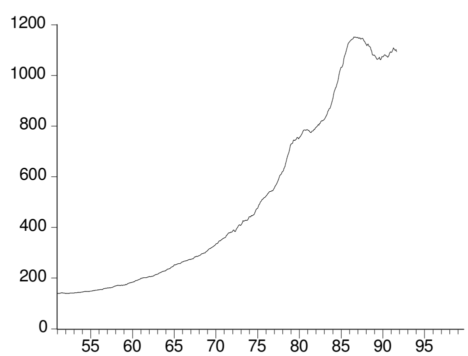
1.3 (a) For better visual impression the logarithm of the exchange rate is plotted on the vertical axis and time on the horizontal axis.



As you can see, the exchange rates show a good deal of variability. For example, in 1985 one US dollar only bought about 0.257 Pesos, but in 2004 it could buy about 11.29 Pesos.

(b) Again, the picture is mixed. For instance, between 1985 and 2006, the U.S. dollar appreciated at a relatively high rate against the Peso, but for most of the other currencies the relationship more slowly and steadily increased.

1.4. The graph of the M1 money supply is as follows:



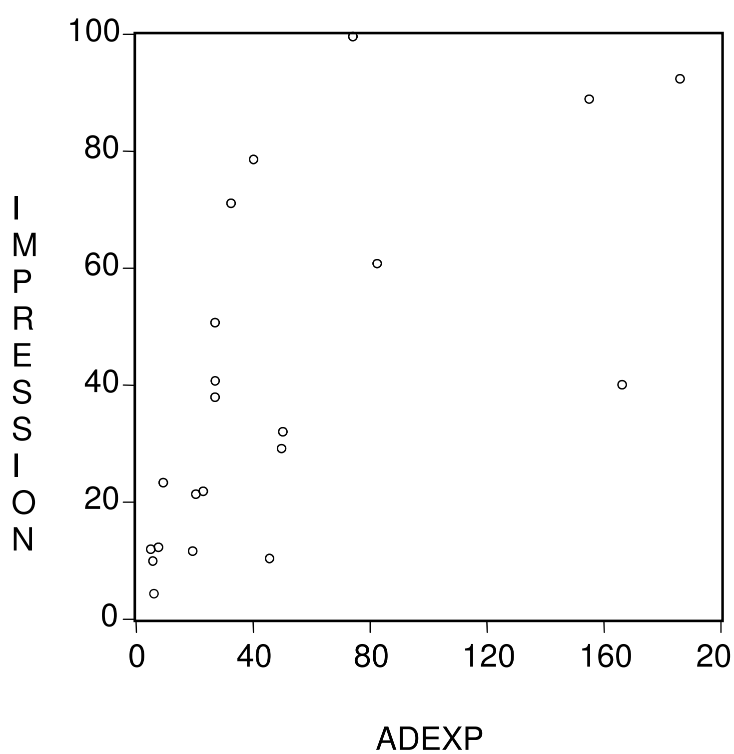
M1 Money Supply: USA, 1951:01-1999:09

As GDP increases over time, naturally a higher amount of the money supply is needed to finance the increased output.

1.5. Some of the relevant variables would include: (1) wages or earnings in criminal activity, (2) hourly wages or earnings in non-criminal activity, (3) probability of getting caught, (4) probability of conviction, (5) expected sentence after conviction. Note that it may not be easy to get data on earnings in the illegal activities. Anyway, refer to the Becker article cited in the text.

1.6. One key factor in the analysis would be the labor force participation rate of people in the 65-69 age category. Data on labor force participation are collected by the Labor Department. If, after the new law went into effect, we find increased participation of these "senior" citizens in the labor force, that would be a strong indication that the earlier law had artificially restricted their labor market participation. It would also be interesting to find out what kinds of of jobs these workers get and what they earn.

1.7 (*a*), (*b*) & (*c*). As the following figure shows, there seems to be a positive relationship between the two variables, although it does not seem to be very strong. This probably suggests that it pays to advertise; otherwise, it is bad news for the advertising industry.



CHAPTER 2:
TWO-VARIABLE REGRESSION ANALYSIS: SOME BASIC IDEAS

2.1 It tells how the mean or average response of the sub-populations of Y varies with the fixed values of the explanatory variable (s).

2.2 The distinction between the sample regression function and the population regression function is important, for the former is is an estimator of the latter; in most situations we have a sample of observations from a given population and we try to learn something about the population from the given sample.

2.3 A regression model can never be a completely accurate description of reality. Therefore, there is bound to be some difference between the actual values of the regressand and its values estimated from the chosen model. This difference is simply the stochastic error term, whose various forms are discussed in the chapter. The residual is the sample counterpart of the stochastic error term.

2.4 Although we can certainly use the mean value, standard deviation and other summary measures to describe the behavior the of the regressand, we are often interested in finding out if there are any causal forces that affect the regressand. If so, we will be able to better predict the mean value of the regressand. Also, remember that econometric models are often developed to test one or more economic theories.

2.5 A model that is linear in the parameters; it may or may not be linear in the variables.

2.6 Models (a), (b), (c) and (e) are linear (in the parameter) regression models. If we let $\alpha = \ln \beta_1$, then model (d) is also linear.

2.7 (a) Taking the natural log, we find that $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$, which becomes a linear regression model.
(b) The following transformation, known as the **logit** transformation, makes this model a linear regression model:

$$\ln [(1- Y_i)/Y_i] = \beta_1 + \beta_2 X_i + u_i$$

(c) A linear regression model
(d) A nonlinear regression model
(e) A nonlinear regression model, as β_2 is raised to the third power.

2.8 A model that can be made linear in the parameters is called an intrinsically linear regression model, as model (a) above. If β_2 is 0.8 in model (d) of Question 2.7, it becomes a linear regression model, as $e^{-0.8(X_i - 2)}$ can be easily computed.

2.9 (a) Transforming the model as $(1/Y_i) = \beta_1 + \beta_2 X_i$ makes it a linear regression model.
(b) Writing the model as $(X_i/Y_i) = \beta_1 + \beta_2 X_i$ makes it a linear regression model.
(c) The transformation $\ln[(1 - Y_i)/Y_i] = -\beta_1 - \beta_2 X_i$ makes it a linear regression model.
Note: Thus the original models are intrinsically linear models.

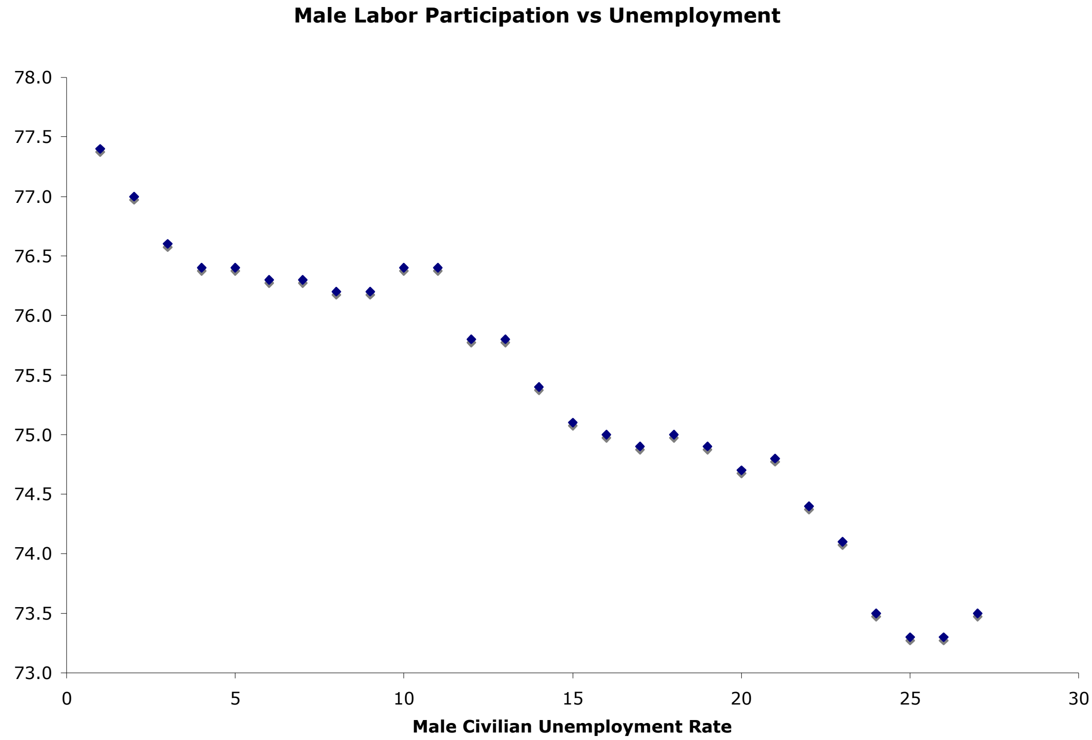
2.10 This scattergram shows that more export-oriented countries on average have more growth in real wages than less export oriented countries. That is why many developing countries have followed an export-led growth policy. The regression line sketched in the diagram is a sample regression line, as it is based on a sample of 50 developing countries.

2.11 According to the well-known Heckscher-Ohlin model of trade, countries tend to export goods whose production makes intensive use of their more abundant factors of production. In other words, this model emphasizes the relation between factor endowments and comparative advantage.

2.12 This figure shows that the higher is the minimum wage, the lower is per head GNP, thus suggesting that minimum wage laws may not be good for developing countries. But this topic is controversial. The effect of minimum wages may depend on their effect on employment, the nature of the industry where it is imposed, and how strongly the government enforces it.

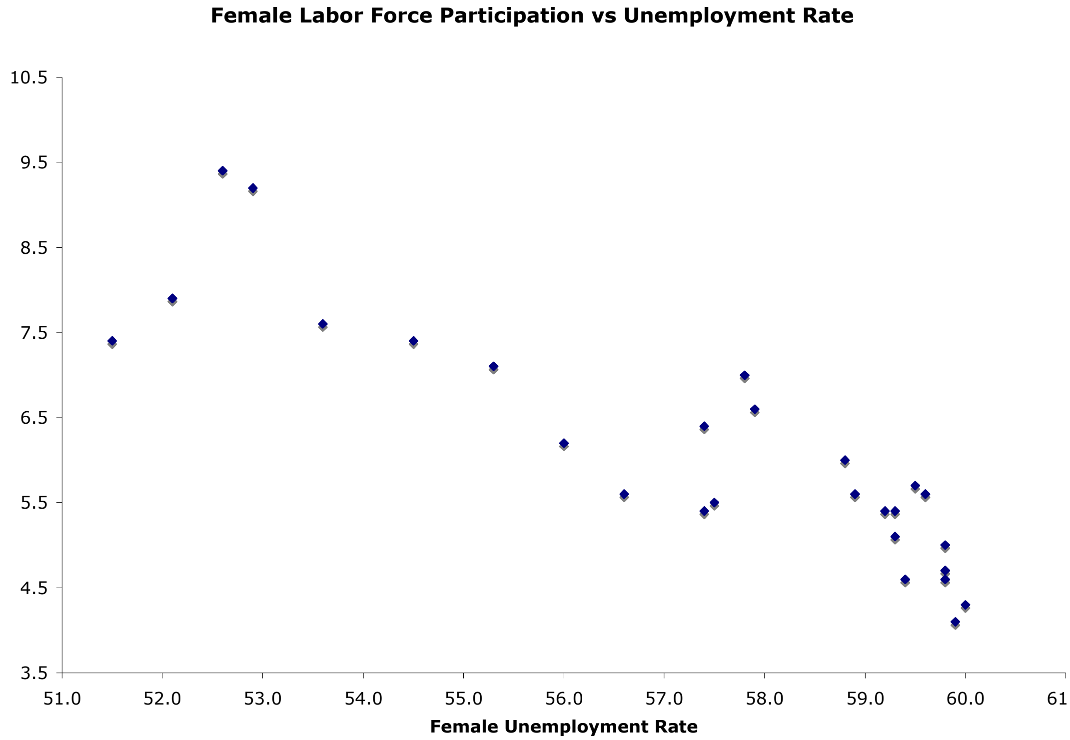
2.13 It is a sample regression line because it is based on a sample of 15 years of observations. The scatter points around the regression line are the actual data points. The difference between the actual consumption expenditure and that estimated from the regression line represents the (sample) residual. Besides GDP, factors such as wealth, interest rate, etc. might also affect consumption expenditure.

2.14 (a) The scattergram is as follows:



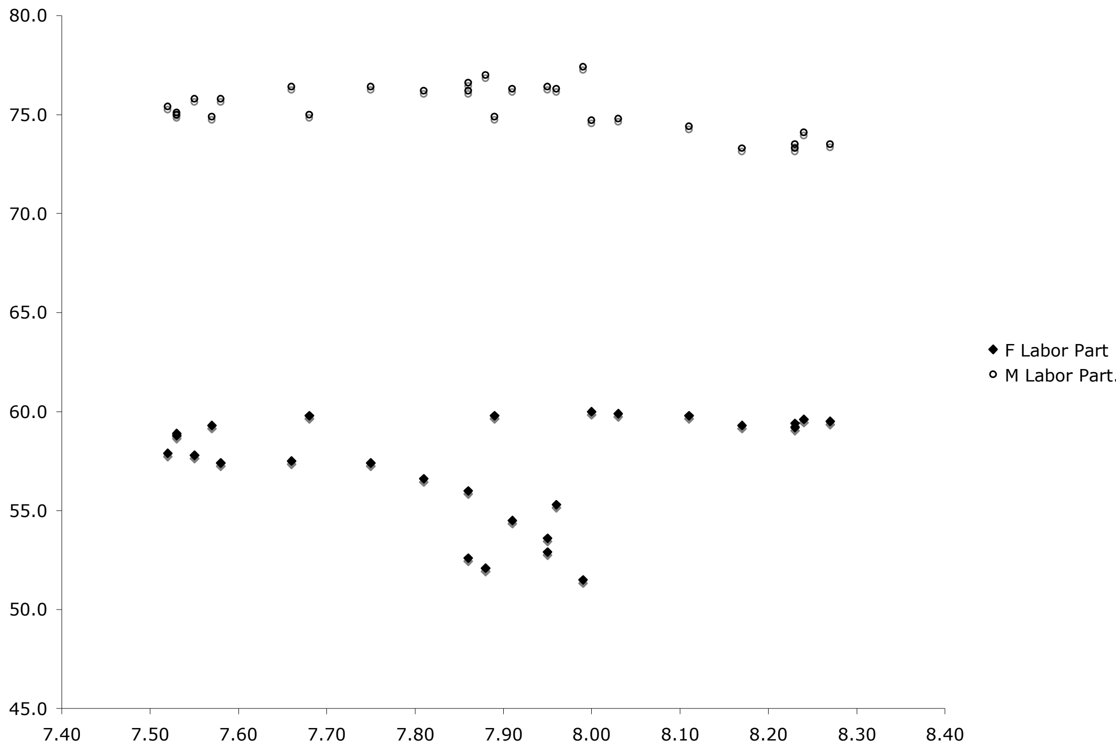
The negative relationship between the two variables seems seems relatively reasonable. As the unemployment rate increases, the labor force participation rate decreases, although there are several minor peaks and valleys in the graph.

(b) The scattergram is as follows:



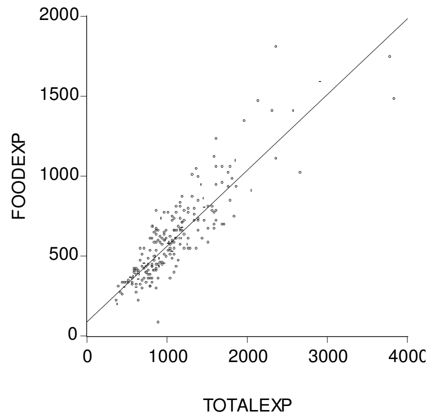
Here the **discouraged worker hypothesis** of labor economics seems to be at work: unemployment discourages female workers from participating in the labor force because they fear that there are no job opportunities.

(c) The plot of Male and Female Labor Force Participation against AH82 shows the following:



There is a similar relationship between the two variables for males and females, although the Male Labor Participation Rate is always significantly higher than that of the Females. Also, there is quite a bit more variability among the Female Rates. To validate these statements, the average Male Rate is 75.4 %, whereas the Female average is only 57.3 %. With respect to variability, the Male Rate standard deviation is only 1.17 %, but the Female standard deviation is 2.73 %, more than double that of the Male Rate. Keep in mind that we are doing simple bivariate regressions here. When we study multiple regression analysis, we may have some different conclusions.

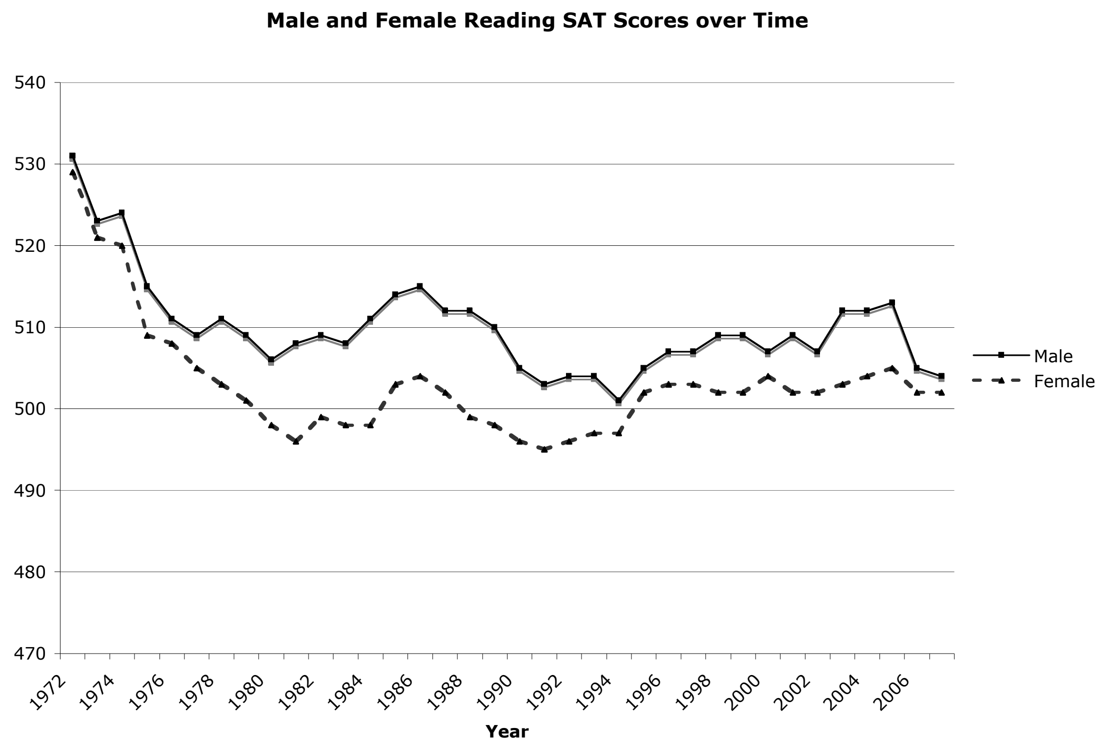
2.15 (a) The scattergram and the regression line look as follows:



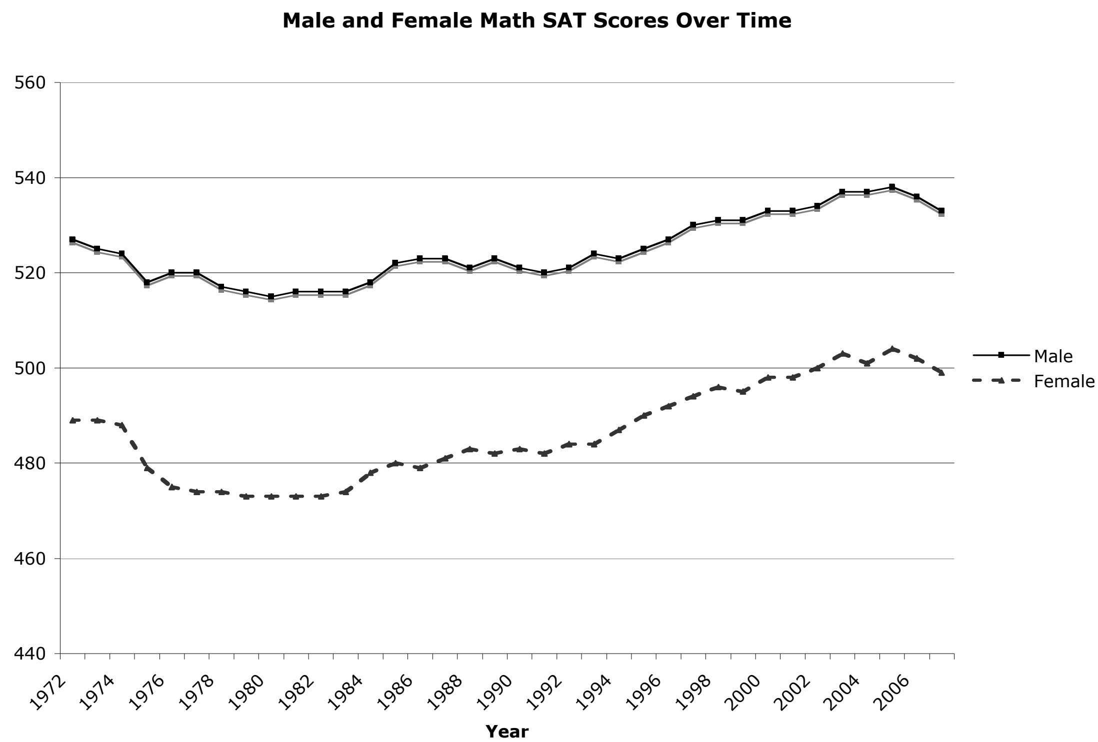
(b) As total expenditure increases, on the average, expenditure on food also increases. But there is greater variability between the two after the total expenditure exceeds the level of Rs. 2000.

(c) We would not expect the expenditure on food to increase linearly (i.e., in a straight line fashion) for ever. Once basic needs are satisfied, people will spend relatively less on food as their income increases. That is, at higher levels of income consumers will have more discretionary income. There is some evidence of this from the scattergram shown in (a): At the income level beyond Rs. 2000, expenditure on food shows much more variability.

2.16 (a) The scatter plot for male and female verbal scores is as follows:



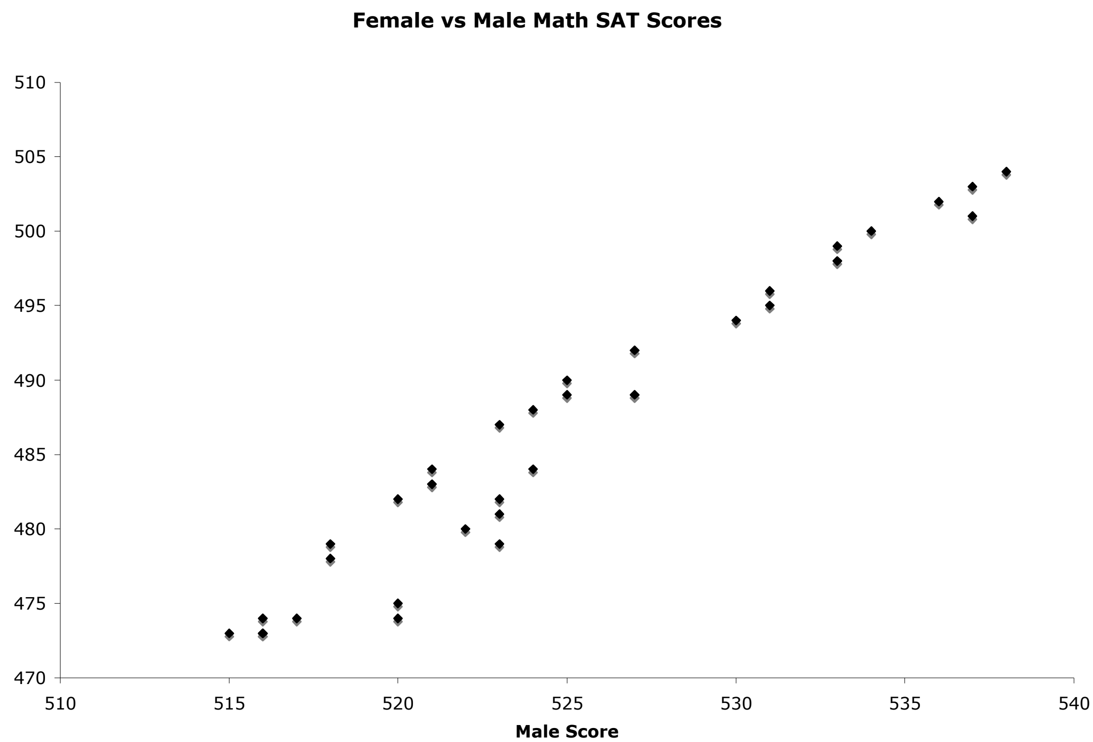
And the corresponding plot for male and female math score is as follows:



(b) Over the years, the male and female reading scores show a slight downward trend, although there seems to be a leveling in the mid-1990's. The math scores, however, show a slight increasing trend, especially starting in the early 1990's. In both graphs it seems the male scores are generally higher than the female scores, of course with year-to-year variation.

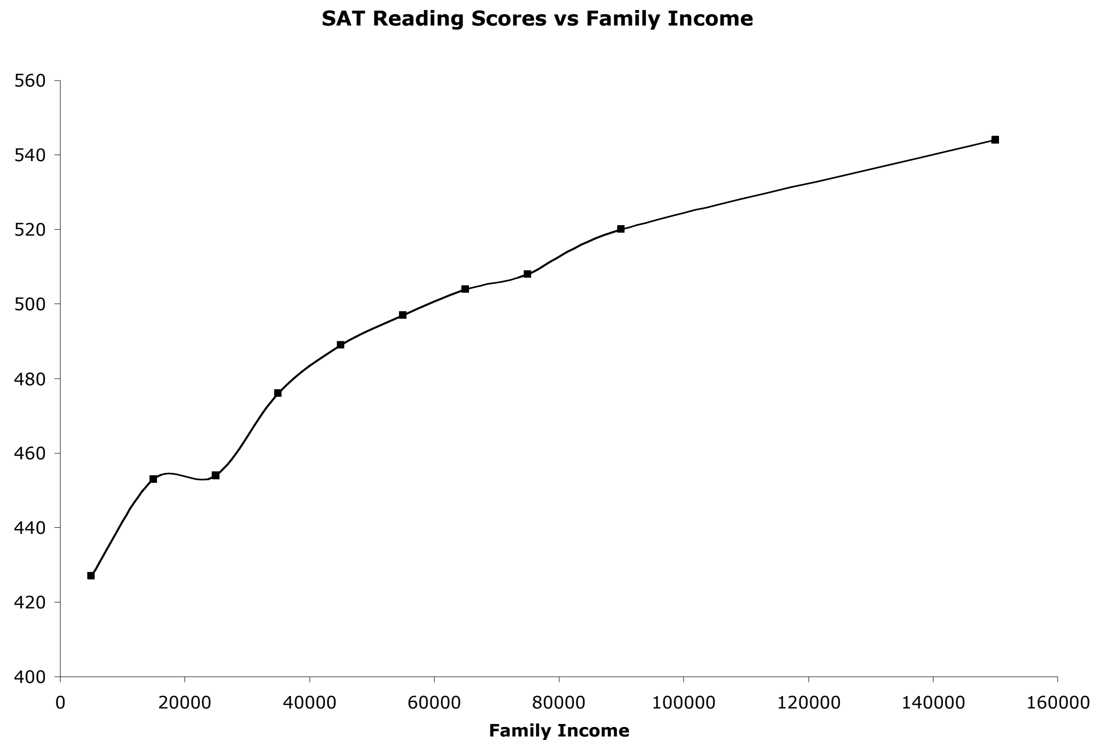
(c) We can develop a simple regression model regressing the math score on the verbal score for both sexes.

(d) The plot is as follows:



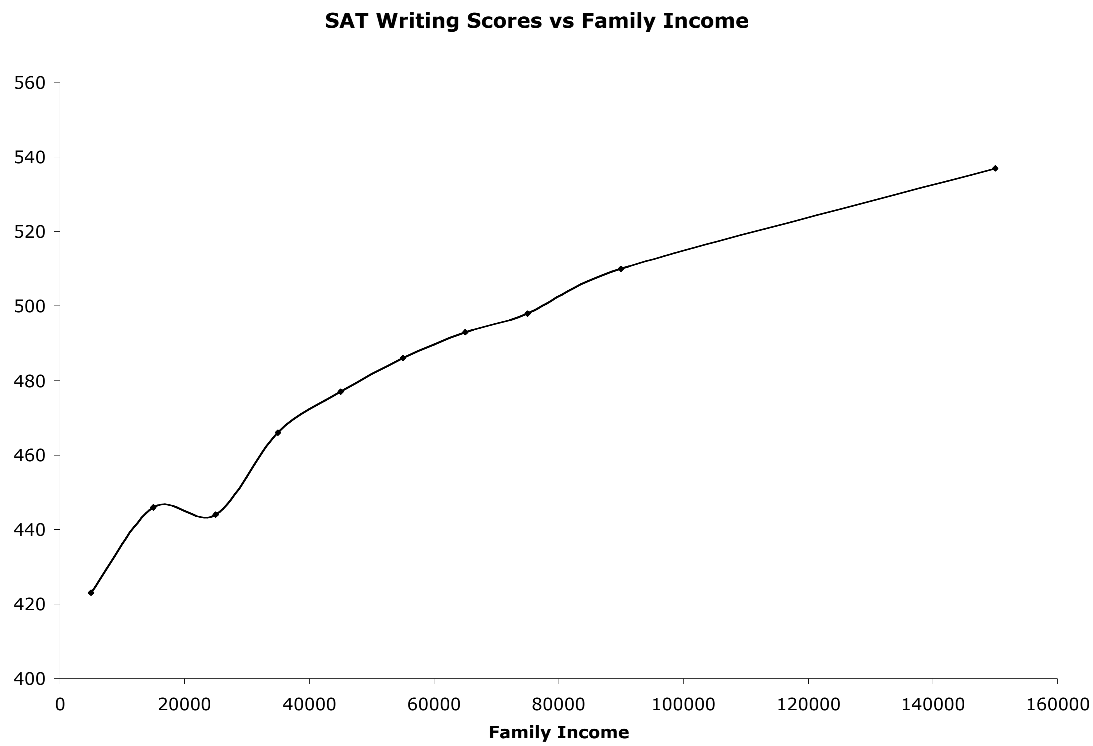
As the graph shows, the two genders seem to move together, although the male scores are always higher than the female scores.

2.17 (a) The scatter plot for male and female verbal scores is as follows:



The results are somewhat similar to those from Figure 2.7. It seems that the average Reading Score increases as the average Family Income increases.

b)



This graph looks almost identical to the previous ones, especially the Reading Score graph.

c) Apparently, there seems to be a positive relationship between average Family Income and SAT scores.

CHAPTER 3: TWO-VARIABLE REGRESSION MODEL: THE PROBLEM OF ESTIMATION

3.1 (1) $Y_i = \beta_1 + \beta_2 X_i + u_i$. Therefore,

$E(Y_i|X_i) = E[(\beta_1 + \beta_2 X_i + u_i)|X_i]$

$= \beta_1 + \beta_2 X_i + E(u_i|X_i)$, since the β's are constants and X is nonstochastic.

$= \beta_1 + \beta_2 X_i$, since $E(u_i|X_i)$ is zero by assumption.

(2) Given $cov(u_i u_j) = 0$ for $\forall$ for all i,j ($i \neq j$), then

$cov(Y_i Y_j) = E\{[Y_i - E(Y_i)][Y_j - E(Y_j)]\}$

$= E(u_i u_j)$, from the results in (1)

$= E(u_i)E(u_j)$, because the error terms are not correlated by assumption,

$= 0$, since each u_i has zero mean by assumption.

(3) Given $var(u_i \backslash X_i) = \sigma^2$, $var(Y_i \backslash X_i) = E[Y_i - E(Y_i)]^2 = E(u_i^2) = var(u_i \backslash X_i) = \sigma^2$, by assumption.

3.2

	Y_i	X_i	y_i	x_i	$x_i y_i$	x_i^2
	4	1	-3	-3	9	9
	5	4	-2	0	0	0
	7	5	0	1	0	1
	12	6	5	2	10	4
sum	28	16	0	0	19	14

Note: $\bar{Y} = 7$ and $\bar{X} = 4$

Therefore, $\hat{\beta}_2 = \dfrac{\sum x_i y_i}{\sum x_i^2} = \dfrac{19}{14} = 1.357$; $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 1.572$

3.3 The PRF is: $Y_i = \beta_1 + \beta_2 X_i + u_i$

Situation I: $\beta_1 = 0, \beta_2 = 1,$ and $E(u_i) = 0$, which gives $E(Y_i|X_i) = X_i$

Situation 2: $\beta_1 = 1, \beta_2 = 0$, and $E(u_i) = (X_i - 1)$, which gives

$\overline{E(Y_i|X_i) = X_i}$

which is the same as Situation 1. Therefore, without the assumption $E(u_i) = 0$, one cannot estimate the parameters, because, as just shown, one obtains the same conditional distribution of Y although the assumed parameter values in the two situations are quit different.

3.4 Imposing the first restriction, we obtain:

$\sum \hat{u}_i = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$

Simplifying this yields the first normal equation.

Imposing the second restriction, we obtain:

$\sum \hat{u}_i X_i = \sum [(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i] = 0$

Simplifying this yields the second normal equation.

The first restriction corresponds to the assumption that $E(u_i \backslash X_i) = 0$. The second restriction corresponds to the assumption that the population error term is uncorrelated with the explanatory variable X_i, i.e., $\text{cov}(u_i X_i) = 0$.

3.5 From the Cauchy-Schwarz inequality it follows that:

$$\frac{E(XY)^2}{E(X^2)E(Y^2)} \leq 1$$

Now $r^2 = \dfrac{\sum (x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \leq 1$, by analogy with the Cauchy-Schwarz inequality. This also holds true of ρ^2, the squared population correlation coefficient.

3.6 Note that:

$$\beta_{yx} = \frac{\sum x_i y_i}{\sum x_i^2} \text{ and } \beta_{xy} = \frac{\sum x_i y_i}{\sum y_i^2}$$

Multiplying the two, we obtain the expression for r^2, the squared sample correlation coefficient.

3.7 Even though $\hat{\beta}_{yx} . \hat{\beta}_{xy} = 1$, it may still matter (for causality and theory) if Y is regressed on X or X on Y, since it is just the product of the two that equals 1. This does not say that $\hat{\beta}_{yx} = \hat{\beta}_{xy}$.

3.8 The means of the two-variables are: $\bar{Y} = \bar{X} = \dfrac{n+1}{2}$ and the correlation between the two rankings is:

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \qquad (1)$$

where small letters as usual denote deviation from the mean values. Since the rankings are permutations of the first *n* natural numbers,

$$\sum x_i^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}$$

and similarly,

$\sum y_i^2 = \frac{n(n^2-1)}{12}$, Then

$$\sum d^2 = \sum (X_i - Y_i)^2 = \sum (X_i^2 + Y_i^2 - 2X_iY_i)$$

$$= \frac{2n(n+1)(2n+1)}{6} - 2\sum X_iY_i$$

Therefore, $\sum X_iY_i = \frac{n(n+1)(2n+1)}{6} - \frac{\sum d^2}{2}$ (2)

Since $\sum x_iy_i = \sum X_iY_i - \frac{\sum X_i \sum Y_i}{n}$, using (2), we obtain

$$\frac{n(n+1)(2n+1)}{3} - \frac{\sum d^2}{2} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12} - \frac{\sum d^2}{2} \qquad (3)$$

Now substituting the preceding equations in (1), you will get the answer.

3.9 (a) $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 X_i$ and $\hat{\alpha}_1 = \overline{Y} - \hat{\beta}_2 \bar{x}$ [Note: $x_i = (X_i - \overline{X})$]

$$= \overline{Y}, \text{ since } \sum x_i = 0$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 \text{ and } \text{var}(\hat{\alpha}_1) = \frac{\sum x_i^2}{n\sum x_i^2}\sigma^2 = \frac{\sigma^2}{n}$$

Therefore, neither the estimates nor the variances of the two estimators are the same.

(b) $\hat{\beta}_2 = \frac{\sum x_iy_i}{\sum x_i^2}$ and $\hat{\alpha}_1 = \frac{\sum x_iy_i}{\sum x_i^2}$, since $x_i = (X_i - \overline{X})$

It is easy to verify that var($\hat{\beta}_2$) = var($\hat{\alpha}_2$) = $\dfrac{\sigma^2}{\sum x_i^2}$

That is, the estimates and variances of the two slope estimators are the same.

(c) Model II may be easier to use with large X numbers, although with high speed computers this is no longer a problem.

3.10 Since $\sum x_i = \sum y_i = 0$, that is, the sum of the deviations from mean value is always zero, $\bar{x} = \bar{y} = 0$ are also zero. Therefore, $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 0$. The point here is that if both Y and X are expressed as deviations from their mean values, the regression line will pass through the origin.

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$, since means of the two variables are zero. This is equation (3.1.6).

3.11 Let $Z_i = aX_i + b$ and $W_i = cYi + d$. In deviation form, these become: $z_i = ax_i$ and $w_i = cy_i$. By definition,

$$r_2 = \frac{\sum z_i w_i}{\sqrt{\sum z_i^2 \sum w_i^2}} = \frac{ac\sum x_i y_i}{ac\sqrt{\sum x_i^2 \sum y_i^2}} = r_1 \text{ in Eq.(3.5.13)}$$

3.12 (a) True. Let a and c equal -1 and b and d equal 0 in Question 3.11.

(b) False. Again using Question 3.11, it will be negative.

(c) True. Since $r_{xy} = r_{yx} > 0$, S_x and S_y (the standard deviations of X and Y, respectively) are both positive, and $r_{yx} = \beta_{yx}\frac{S_x}{S_y}$ and $r_{xy} = \beta_{xy}\frac{S_y}{S_x}$, then β_{xy} *and* β_{yx} must be positive.

3.13 Let Z = $X_1 + X_2$ and W = X_2 and X_3. In deviation form, we can write these as z = $x_1 + x_2$ and w = $x_2 + x_3$. By definition the correlation between Z and W is:

$$r_{zw} = \frac{\sum z_i w_i}{\sqrt{\sum z_i^2 \sum w_i^2}} = \frac{\sum (x_1 + x_2)(x_2 + x_3)}{\sqrt{\sum (x_1 + x_2)^2 \sum (x_2 + x_3)^2}}$$

$$= \frac{\sum x_2^2}{\sqrt{(\sum x_1^2 + \sum x_2^2)(\sum x_2^2 + \sum x_3^2)}}$$, because the X's are uncorrelated. *Note:* We have omitted the observation subscript for convenience.

$$= \frac{\sigma^2}{\sqrt{(2\sigma^2 + 2\sigma^2)}} = \frac{1}{2}$$, where σ^2 is the common variance.

The coefficient is not zero because, even though the X's are individually uncorrelated, the pairwise combinations are not.

As just shown, $\sum zw = \sigma^2$, meaning that the covariance between z and w is some constant other than zero.

3.14 The residuals and fitted values of Y will not change. Let
$Y_i = \beta_1 + \beta_2 X_i + u_i$ and $Y_i = \alpha_1 + \alpha_2 Z_i + u_i$, where Z = 2X
Using the deviation form, we know that

$$\hat{\beta}_2 = \frac{\sum xy}{\sum x^2}$$, omitting the observation subscript.

$$\hat{\alpha}_2 = \frac{\sum z_i y_i}{\sum z_i^2} = \frac{2\sum x_i y_i}{4\sum x_i^2} = \frac{1}{2}\hat{\beta}_2$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\,\bar{X}\,;\ \hat{\alpha}_1 = \bar{Y} - \hat{\alpha}_2\,\bar{Z} = \hat{\beta}_1\ (\text{Note: } \bar{Z} = 2\bar{X}\,)$$

That is the intercept term remains unaffected. As a result, the fitted Y values and the residuals remain the same even if X_i is multiplied by 2. The analysis is *analogous* if a constant is added to X_i.

3.15 By definition,

$$r_{y\hat{y}}^2 = \frac{(\sum y_i\hat{y}_i)^2}{(\sum y_i^2)(\sum \hat{y}_i^2)} = \frac{\left[\sum(\hat{y}_i + \hat{u}_i)(\hat{y}_i)\right]^2}{(\sum y_i^2)(\sum \hat{y}_i^2)} = \frac{\sum \hat{y}_i^2}{\sum y_i^2},$$

since $\sum \hat{y}_i\hat{u}_i = 0$. $= \dfrac{\sum(\hat{\beta}_2 x_i)^2}{\sum y_i^2} = \dfrac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} = r^2$, using (3.5.6).

3.16 (a) *False*. The covariance can assume any value; its value depends on the units of measurement. The correlation coefficient, on the other hand, is unitless, that is, it is a pure number.

(b) *False*. See Fig.3.11*h*. Remember that correlation coefficient is a measure of *linear* relationship between two variables. Hence, as Fig.3.11*h* shows, there is a perfect relationship between Y and X, but that relationship is nonlinear.

(c) *True*. In deviation form, we have

$$y_i = \hat{y}_i + \hat{u}_i$$

Therefore, it is obvious that if we regress y_i on $\hat{y}_i$, the slope coefficient will be one and the intercept zero. But a formal proof can proceed as follows:

If we regress y_i on $\hat{y}_i$, we obtain the slope coefficient, say, $\hat{\alpha}$ as:

$$\hat{\alpha} = \frac{\sum y_i\hat{y}_i}{\sum \hat{y}^2} = \frac{\hat{\beta}\sum x_i y_i}{\hat{\beta}^2 \sum x_i^2} = \frac{\hat{\beta}^2}{\hat{\beta}^2} = 1, \text{ because}$$

$\hat{y}_i = \hat{\beta} x_i$ and $\sum x_i y_i = \hat{\beta}\sum x_i^2$ for the two-variable model. The intercept in this regression is zero.

3.17 Write the sample regression as: $Y_i = \hat{\beta}_1 + \hat{u}_i$. By LS principle, we want to minimize: $\sum \hat{u}_i^{\,2} = \sum (Y_i - \hat{\beta}_1)^2$. Differentiate this equation with the only unknown parameter and set the resulting expression to zero, to obtain:

$$\frac{d(\hat{u}_{i.}{}^2)}{d\hat{\beta}_1} = 2\sum(Y_i - \hat{\beta}_1)(-1) = 0$$

which on simplification gives $\hat{\beta}_1 = \bar{Y}$,that is, the sample mean. And we know that the variance of the sample mean is $\frac{\sigma_y^2}{n}$, where n is the sample size, and σ^2 is the variance of Y. The RSS is $\sum(Y_i - \bar{Y})^2 = \sum y_i^2$ and $\hat{\sigma}^2 = \frac{RSS}{(n-1)} = \frac{\sum y_i^2}{(n-1)}$. It is worth adding the X variable to the model if it reduces $\hat{\sigma}^2$ significantly, which it will if X has any influence on Y. In short, in regression models we hope that the explanatory variable(s) will better predict Y than simply its mean value. As a matter of fact, this can be looked at formally. Recall that for the two-variable model we obtain from (3.5.2),

$$\begin{aligned} RSS &= TSS - ESS \\ &= \sum y_i^2 - \sum \hat{y}_i^2 \\ &= \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2 \end{aligned}$$

Therefore, if $\hat{\beta}_2$ is different from zero, RSS of the model that contains at least one regressor, will be smaller than the model with no regressor. Of course, if there are more regressors in the model and their slope coefficients are different from zero, the RSS will be much smaller than the no-regressor model.

Empirical Exercises

3.18 Taking the difference between the two ranks, we obtain:

d	-2	1	-1	3	0	-1	-1	-2	1	2
d^2	4	1	1	9	0	1	1	4	1	4

$;\sum d^2 = 26$

Therefore, Spearman's rank correlation coefficient is

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(26)}{10(10^2-1)} = 0.842$$

Thus there is a high degree of correlation between the student's midterm and final ranks. The higher is the rank on the midterm, the higher is the rank on the final.

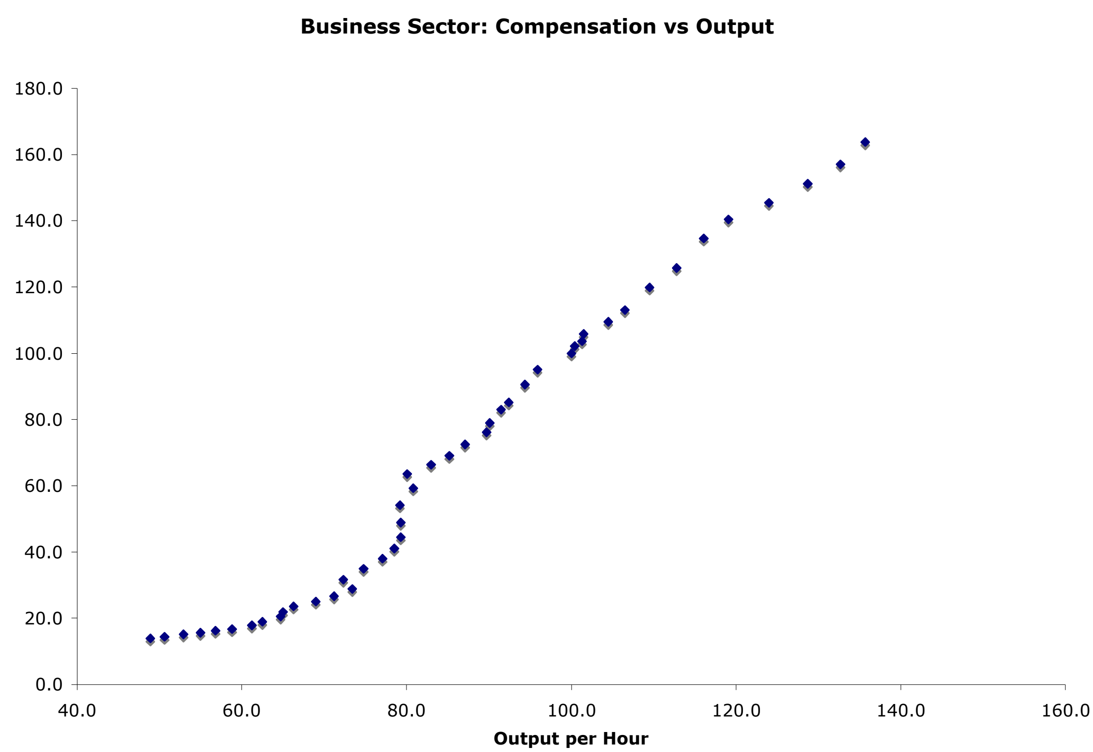
3.19 (a) The slope value of 2.250 suggests that over the period 1985-2005, for every unit increase in the ratio of the US to Canadian CPI, on average, the Canadian to US dollar exchange rate ratio increased by about 2.250 units. That is, as the US dollar strengthened against the

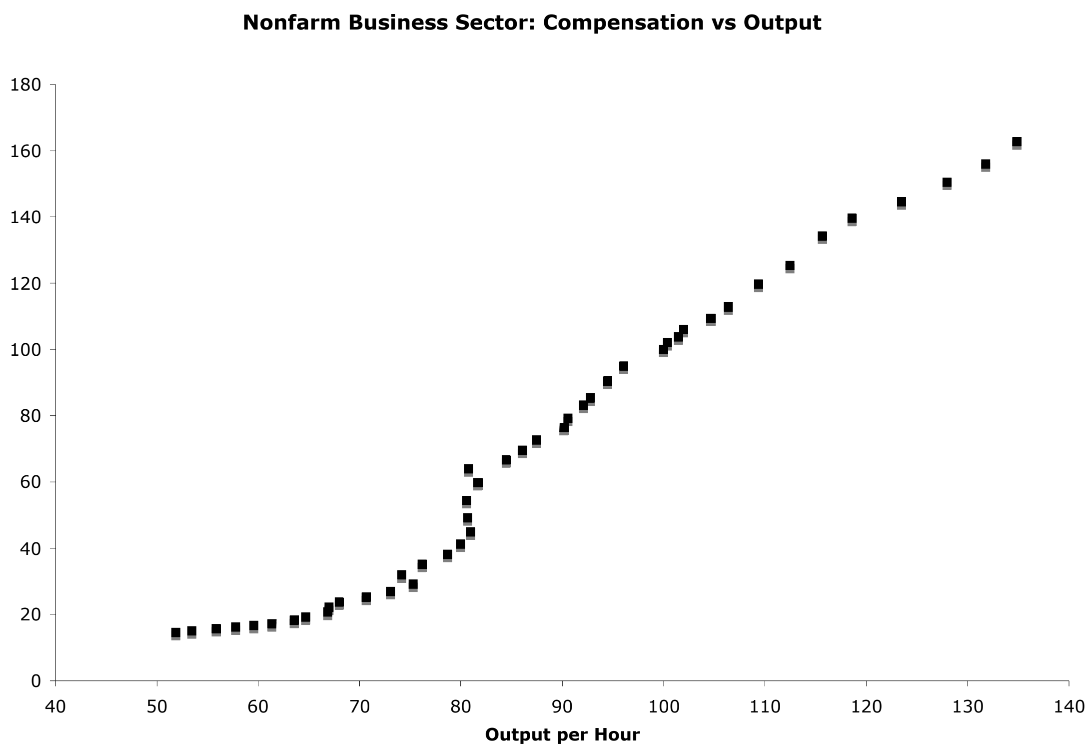
Canadian dollar, one could get more Canadian dollars for each US dollar. Literally interpreted, the intercept value of -0.912 means that if the relative price ratio were zero, a US dollar would exchange for -0.912 Canadian dollars (would lose money). Of course, this interpretation is not economically meaningful. With a fairly low to moderate r^2 of 0.440, we should realize that there is a lot of variability in this result.

(b) The positive value of the slope coefficient makes economic sense because if U.S. prices go up faster than Canadian prices, domestic consumers will switch to Canadian goods because they can buy more, thus increasing the demand for GM, which will lead to appreciation of the German mark. This is the essence of the theory of *purchasing power parity* (PPP), or the law of one price.

(c) In this case the slope coefficient is expected to be negative, for the higher the Canadian CPI relative to the U.S. CPI, the lower the relative inflation rate in Canada which will lead to depreciation of the U.S. dollar. Again, this is in the spirit of the PPP.

3.20 (a) The scattergrams are as follows:





(b) As both the diagrams show, there is a positive relationship between wages and productivity, which is not surprising in view of the *marginal productivity theory* of labor economics.

(c) As the preceding figures show, the relationship between wages is relatively linear, except for a slight upward curve at the lower end of the Output range. Therefore, if we try to fit a straight line regression model to the data we may not get a perfect fit. In a later chapter we will see what types of models are appropriate in this situation. But if we routinely fit the linear model to the data, we obtain the following results.

Business: Compensation = -102.3662 + 1.9924 Output

se = (4.5035) (0.0506) $r^2 = 0.9724$

Nonfarm Business: Compensation = -111.6407 + 2.0757 Output

se = (4.8662) (0.0543) $r^2 = 0.9708$

As expected, the relationship between the two is positive. Surprisingly, the r^2 value is quite high.

3.21 $\sum Y_i$ $\sum X_i$ $\sum X_i Y_i$ $\sum X_i^2$ $\sum Y_i^2$

	$\sum Y_i$	$\sum X_i$	$\sum X_i Y_i$	$\sum X_i^2$	$\sum Y_i^2$
Original data:	1110	1700	205500	322000	132100
Revised data	1110	1680	204200	315400	133300

Therefore, the corrected coefficient of correlation is 0.9688

3.22 (a)

If you plot these variables against time, you will see that there is considerable price volatility for gold, but the NYSE and CPI seem relatively stable.

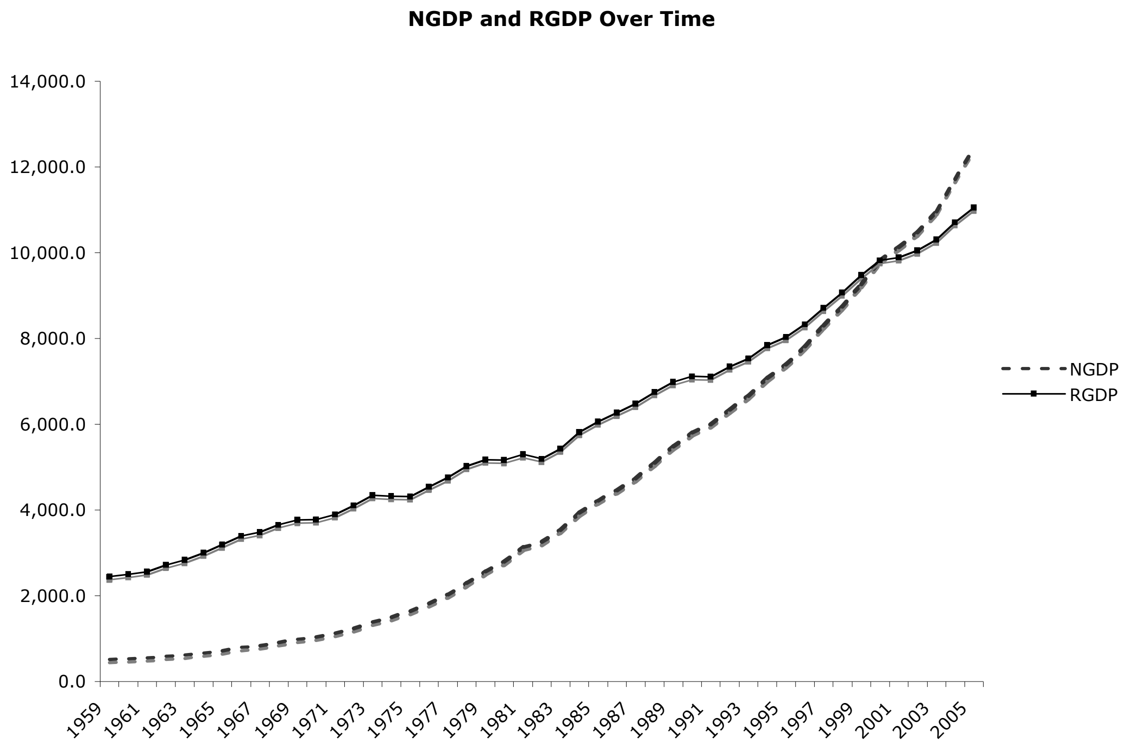
(b) If the hypothesis were true, we would expect $\beta_2 \geq 1$.

Gold Price$_t$ = 215.286 + 1.038 CPI$_t$
se = (54.469) (0.404) $r^2 = 0.1758$

NYSE$_t$ = -3444.992 + 50.297 CPI$_t$
se (533.966) (3.958) $r^2 = 0.8389$

It seems the stock market is a better hedge against inflation than gold.

3.23 (a) The plot is as follows, where NGDP and RGDP are nominal and real GDP.



(b) $$\text{NGDP}_t = -\ 496268 + 252.58 \text{ Year}$$
$$\text{se} = (21089) \quad (10.64) \qquad r^2 = 0.926$$

$$\text{RGDP}_t = -351335 + 180.263 \text{ Year}$$
$$\text{se} = (9070) \quad (4.576) \qquad r^2 = 0.972$$

(c) The slope here gives the rate of change of GDP per year.

(d) The difference between the two represents inflation over time. As the figure and regression results indicate, nominal GDP has been growing at a faster rate than real GDP suggesting that inflation has been rising over time.

3.24 This is straightforward.

3.25 (*a*) See figure in Exercise 2.16 (*d*) (*b*) The regression results are:

$$\hat{Y}_t = -31.76 + 1.0485 X_t$$
$$se = (47.80) \quad (0.0937)$$
$$r^2 = 0.786$$

where Y = female reading score and X = male reading score.

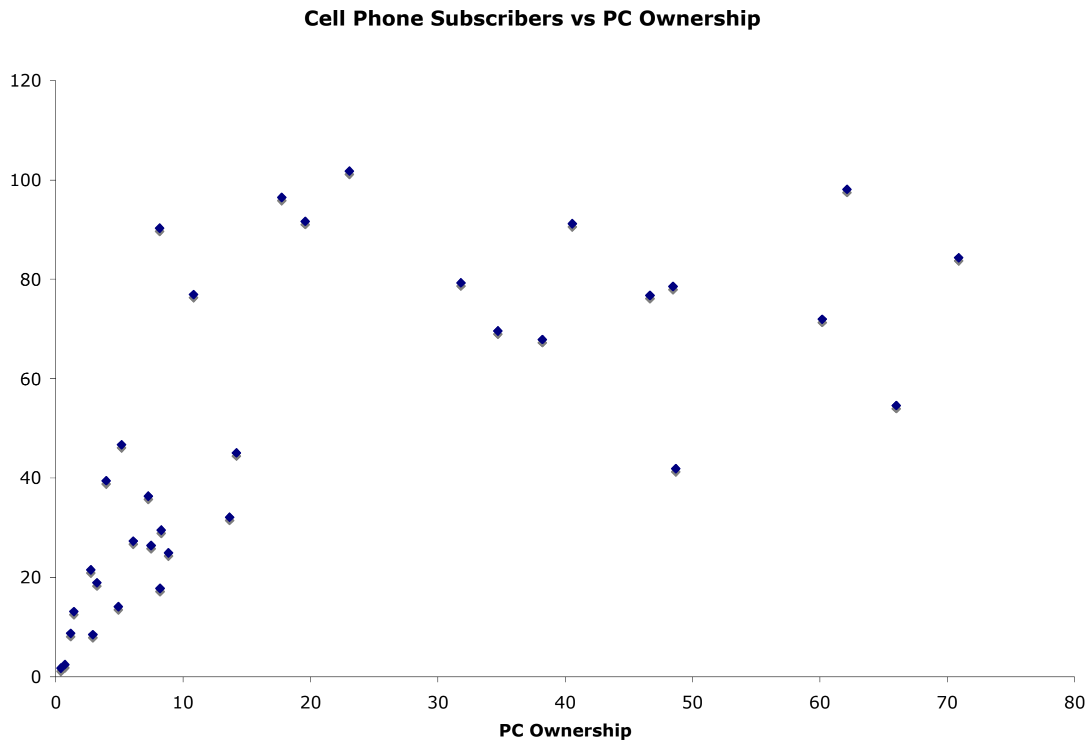
(*c*) As pointed out in the text, a statistical relationship, however strong, does not establish causality, which must be established a priori. In this case, there is no reason to suspect causal relationship between the two variables.

3.26 The regression results are:

$$\hat{Y}_t = -257.02 + 1.416 X_t$$
$$se = (29.35) \quad (0.0559)$$
$$r^2 = 0.950$$

3.27 This is a class project.

3.28



There does seem to be a somewhat positive relationship between these variables, but it is probably better characterized as more logarithmic than linear.

CHAPTER 4:
CLASSICAL NORMAL LINEAR REGRESSION MODEL (CNLRM)

Appendix 4A Exercises

4.1 Given that the coefficient of correlation between Y_1 and Y_2, ρ, is zero, the bivariate normal PDF reduces to:

$$f(Y_1,Y_2) = \frac{1}{2\pi\sigma_1\sigma_2}\exp[-\frac{1}{2}(\frac{Y_1-\mu_1}{\sigma_1})^2 - \frac{1}{2}(\frac{Y_2-\mu_2}{\sigma_2})^2]$$
$$= \{\frac{1}{\sigma_1\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{Y_1-\mu_1}{\sigma_1})^2\}\{\frac{1}{\sigma_2\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{Y_2-\mu_2}{\sigma_2})^2\}$$
$$= f(Y_1)\,f(Y_2)$$

where $f(Y_1)$ and $f(Y_2)$ are the univariate normal PDFs. Thus, when ρ is zero, $f(Y_1,Y_2) = f(Y_1)f(Y_2)$, which is the condition for statistical independence. Therefore, in the bivariate normal case, zero correlation implies statistical independence.

4.2 To ensure that the maximum likelihood estimators maximize the likelihood function, the second derivatives from Eq. (5) in App. 4A must be less than zero, which will ensure that RSS is minimized.

$$\frac{\partial^2 \ln LF}{\partial \beta_1^2} = -\frac{n}{\sigma^2} < 0$$

$$\frac{\partial^2 \ln LF}{\partial \beta_2^2} = -\frac{\sum X^2{}_i}{\sigma^2} < 0$$

$$\frac{\partial^2 \ln LF}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum (Y_i - \beta_1 - \beta_2 X_i)^2$$

$$= \frac{1}{\sigma^2}\left(\frac{n}{2\sigma^2} - \frac{1}{(\sigma^2)^2}\sum \hat{u}^2{}_i\right)$$

since $\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$

$$= \frac{1}{\sigma^2}(\frac{1}{2(\sigma^2)^2}\sum \hat{u}^2{}_i - \frac{1}{(\sigma^2)^2}\sum \hat{u}^2{}_i)\text{, from Eq.(11)}$$

$$= \frac{1}{(\sigma^2)^3}\sum \hat{u}_i^2(\frac{1}{2} - 1) < 0$$