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

$E(Y_i | X_i) = X_i$

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., $cov(u_i X_i) = 0$.

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

$= \bar{Y}$, 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 $x_i = (X_i - \bar{X})$

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:

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

$$= \frac{\sum x_2^2}{\sqrt{(\sum x_1^2 + \sum x_2^2)(\sum x_2^2 + \sum x_3^2)}}, \text{ because the X's are}$$

uncorrelated. *Note:* We have omitted the observation subscript for convenience.

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

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

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

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

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

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

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	
$\mathbf{d^2}$	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.

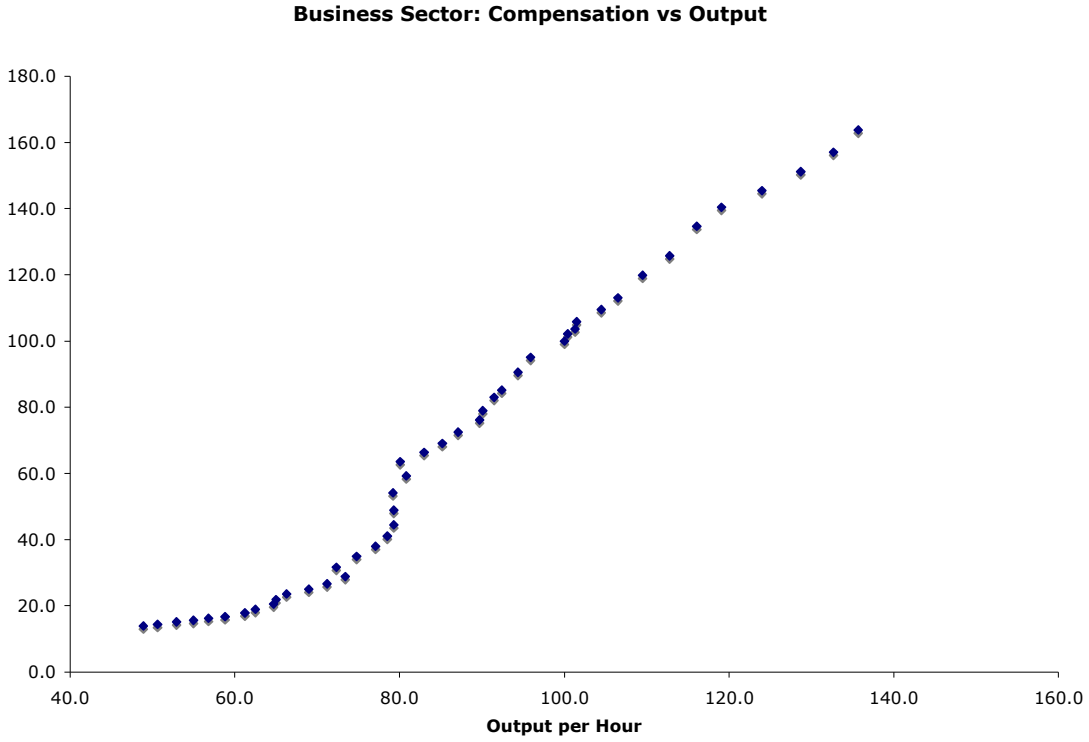
3.19 (a) The slope value of 2.250 suggests that over the period 1985-2005, for every unit increase in the ratio of the US to Canadian CPI, on average, the Canadian to US dollar exchange rate ratio increased by about 2.250 units. That is, as the US dollar strengthened against the

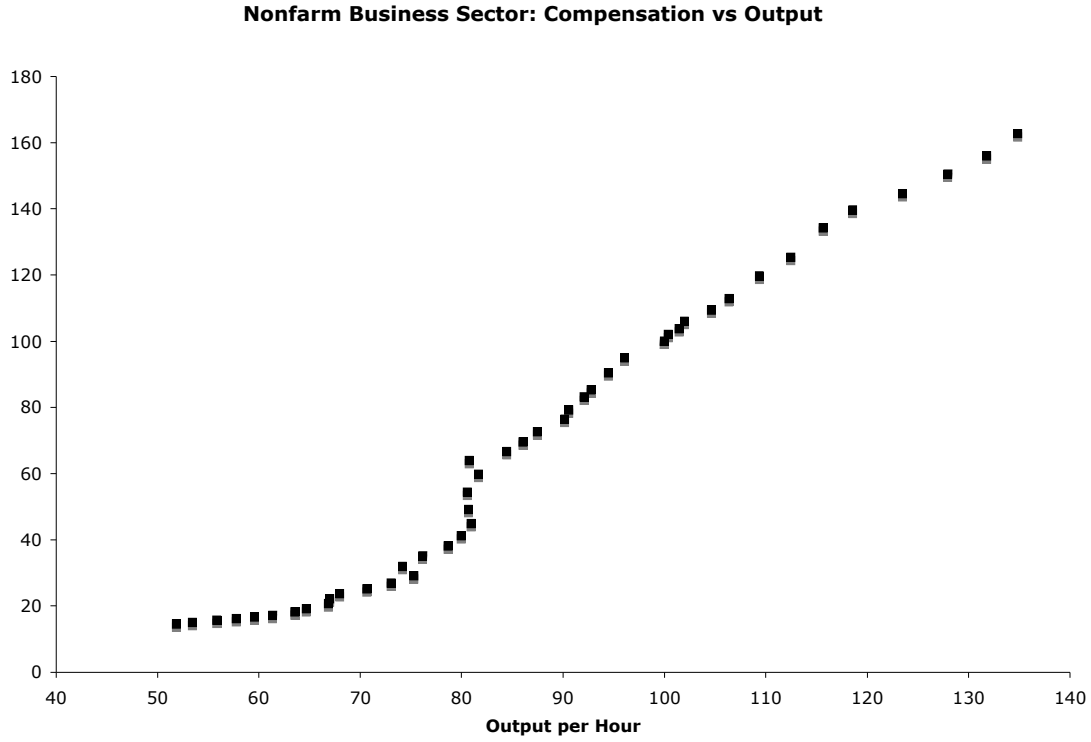
Canadian dollar, one could get more Canadian dollars for each US dollar. Literally interpreted, the intercept value of -0.912 means that if the relative price ratio were zero, a US dollar would exchange for -0.912 Canadian dollars (would lose money). Of course, this interpretation is not economically 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) $r^2 = 0.9708$

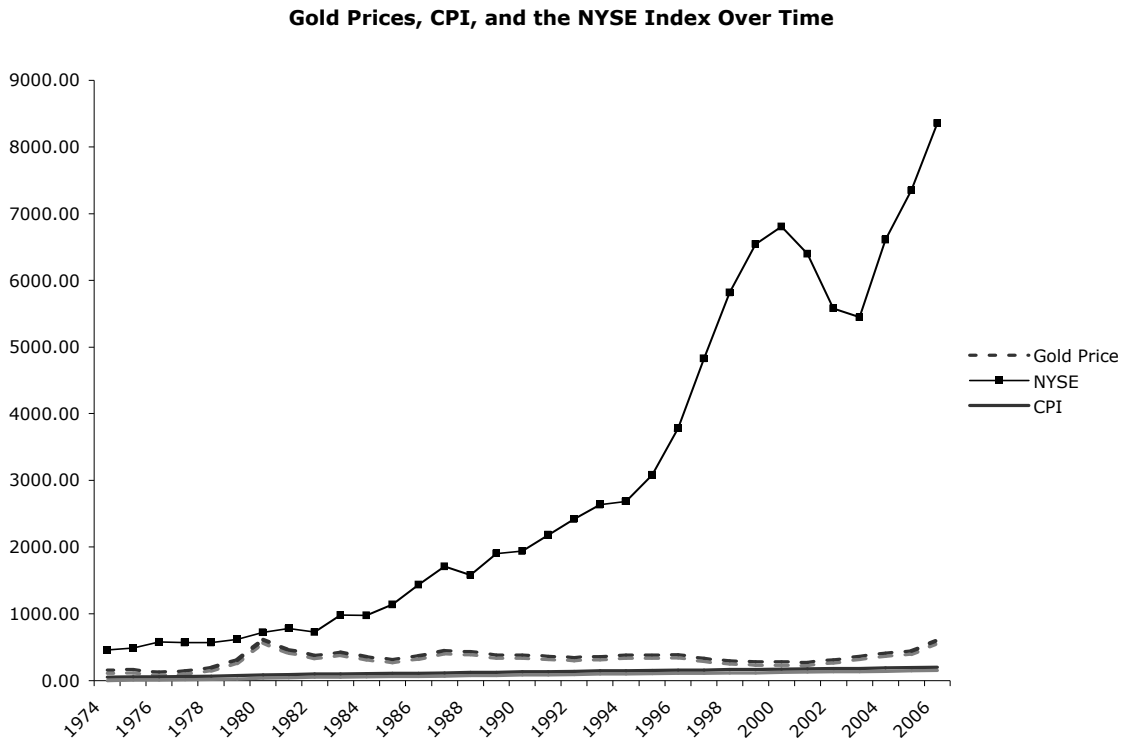
As expected, the relationship between the two is positive. Surprisingly, the r^2 value is quite high.

3.21

$$\sum Y_i \quad \sum X_i \quad \sum X_i Y_i \quad \sum X_i^2 \quad \sum Y_i^2$$

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.

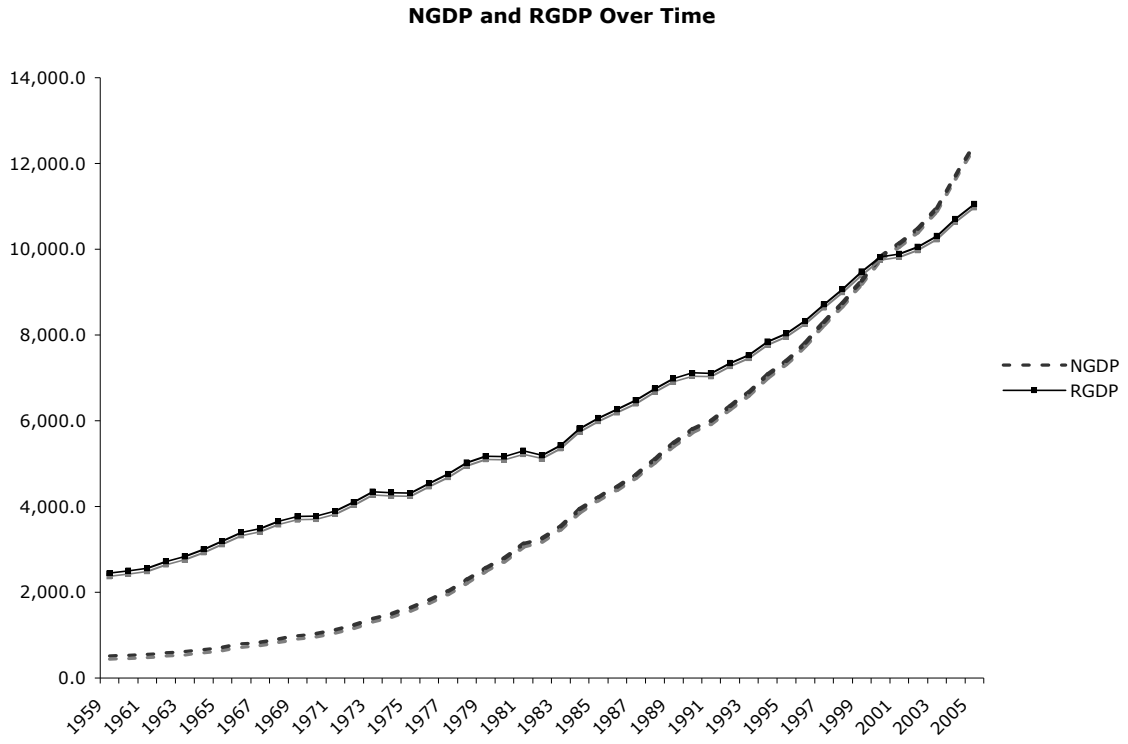
(b) If the hypothesis were true, we would expect $\beta_2 \geq 1$.

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

$$\text{NYSE}_t = -3444.992 + 50.297\ \text{CPI}_t$$
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) $NGDP_t = - 496268 + 252.58$ Year
se = (21089) (10.64) $r^2 = 0.926$

$RGDP_t = -351335 + 180.263$ Year
se = (9070) (4.576) $r^2 = 0.972$

(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.

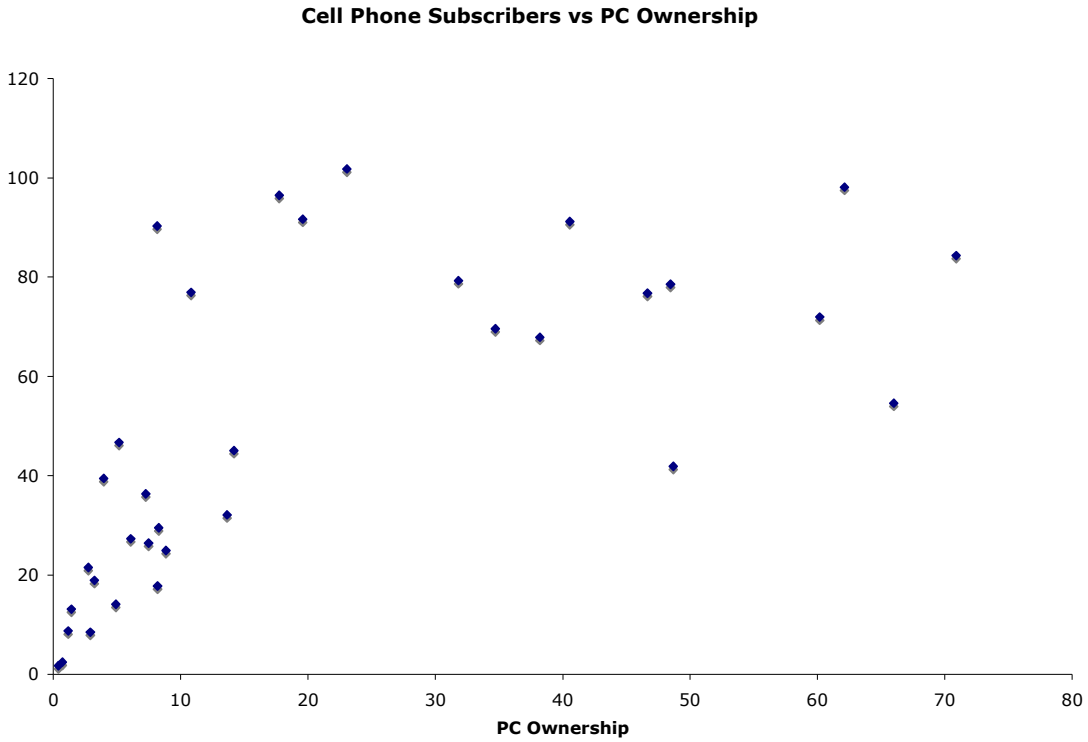
(*c*) As pointed out in the text, a statistical relationship, however strong, does not establish causality, which must be established a priori. In this case, there is no reason to suspect causal relationship between the two variables.

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.