CHAPTER 5:
TWO-VARIABLE REGRESSION:
INTERVAL ESTIMATION AND HYPOTHESIS TESTING

Questions

5.1 (*a*) *True*. The *t* test is based on variables with a normal distribution. Since the estimators of β_1 and β_2 are linear combinations of the error u_i, which is assumed to be normally distributed under CLRM, these estimators are also normally distributed.

(*b*) *True.* So long as E(u_i) = 0, the OLS estimators are unbiased. No probabilistic assumptions are required to establish unbiasedness.

(*c*) *True.* In this case the Eq. (1) in App. 3A, Sec. 3A.1, will be absent. This topic is discussed more fully in Chap. 6, Sec. 6.1.

(*d*) *True.* The *p value* is the smallest level of significance at which the null hypothesis can be rejected. The terms level of significance and size of the test are synonymous.

(*e*) *True.* This follows from Eq. (1) of App. 3A, Sec. 3A.1.

(*f*) *False.* All we can say is that the data at hand does not permit us to reject the null hypothesis.

(g) *False.* A larger σ^2 may be counterbalanced by a larger $\sum x_i^2$. It is only if the latter is held constant, the statement can be true.

(*h*) *False.* The conditional mean of a random variable depends on the values taken by another (conditioning) variable. Only if the two variables are independent, that the conditional and unconditional means can be the same.

(*i*) *True.* This is obvious from Eq. (3.1.7).

(*j*) *True.* Refer of Eq. (3.5.2). If *X* has no influence on *Y*, $\hat{\beta}_2$ will be zero, in which case $\sum y_i^2 = \sum \hat{u}_i^2$.

5.2 ANOVA table for the Food Expenditure in India

Source of variation	SS	df	MSS
Due to regression (ESS)	139023	1	139023
Due to residual (RSS)	236894	53	4470
TSS	375916		

$$F = \frac{139023}{4470} = 31.1013 \text{ with df} = 1 \text{ and } 53\text{, respectively.}$$

Under the hypothesis that there is no relationship between food expenditure and total expenditure, the *p value* of obtaining such an F value is almost zero, suggesting that one can strongly reject the null hypothesis.

5.3 (*a*) se of the intercept coefficient is 6.1523, so the *t* value under H_0 : $\beta_1 = 0$, is: $\frac{14.4773}{6.1523} = 2.3532$. With 32 degrees of freedom, the cutoff for the 5% level of significance is 2.042 (using 30 d.f. since 32 is not in the table in the textbook's appendix), so the intercept IS statistically significant.

(*b*) se of the slope coefficient is 0.00032, so the *t* value under H_0 : $\beta_2 = 0$, is: $\frac{0.0022}{0.00032} = 6.8750$. As noted in part *a*, the cutoff for the 5% level of significance is 2.042, so the slope IS statistically significant.

(*c*) The 95% confidence interval for the true slope coefficient would be: $0.0022 \pm (2.042)(0.00032) \rightarrow [0.0015, 0.0029]$.

(*d*) If per capita income is \$9000, the mean forecast value of cell phones demanded is 14.4773 + 0.0022 (9000) = 34.2773 per 100 persons. For the prediction confidence interval, we first need to compute $\operatorname{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$.

$$\operatorname{var}(\hat{Y}_0) = 422.526 \left[\frac{1}{34} + \frac{(9000 - 15819.865)^2}{12{,}668{,}291{,}885} \right] = 13.9785.$$ Now the confidence interval is given as

$$\Pr\left[\hat{Y}_0 - t_{\alpha/2}\, se\left(\hat{Y}_0\right) \le Y_0 \le \hat{Y}_0 + t_{\alpha/2}\, se\left(\hat{Y}_0\right)\right] = 1-\alpha$$
$$= \Pr\left[34.2773 - 2.042\left(3.7388\right) \le Y_0 \le 34.2773 + 2.042\left(3.7388\right)\right] = 0.95$$
$$\rightarrow \left[26.6427,\ 41.9119\right]$$

5.4 Verbally, the hypothesis states that there is no correlation between the two variables. Therefore, if we can show that the covariance between the two variables is zero, then the correlation must be zero.

5.5 (*a*) Use the *t* test to test the hypothesis that the true slope coefficient is one. That is obtain: $t = \dfrac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} = \dfrac{1.0598 - 1}{0.0728} = 0.821$

For 238 df this *t* value is not significant even at $\alpha = 10\%$. The conclusion is that over the sample period, IBM was not a volatile security.

(*b*) Since $t = \dfrac{0.7264}{0.3001} = 2.4205$, which is significant at the two percent level of significance. But it has little economic meaning. Literally interpreted, the intercept value of about 0.73 means that even if the market portfolio has zero return, the security's return is 0.73 percent.

5.6 Under the normality assumption, $\hat{\beta}_2$ is normally distributed. But since a normally distributed variable is continuous, we know from probability theory that the probability that a continuous random variable takes on a specific value is zero. Therefore, it makes no difference if the equality is strong or weak.

5.7 Under the hypothesis that $\beta_2 = 0$, we obtain

$$t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{\hat{\beta}_2\sqrt{\sum x_i^2}}{\hat{\sigma}} = \frac{\hat{\beta}_2\sqrt{\sum x_i^2}}{\sqrt{\dfrac{\sum y_i^2(1-r^2)}{(n-2)}}}$$

because $\hat{\sigma}^2 = \dfrac{\sum \hat{u}_i^2}{(n-2)} = \dfrac{\sum y_i^2(1-r^2)}{(n-2)}$, from Eq.(3.5.10)

$$= \frac{\hat{\beta}_2\sqrt{\sum x_i^2}\sqrt{(n-2)}}{\sqrt{\sum y_i^2}\sqrt{(1-r^2)}}$$

But since $r^2 = \hat{\beta}_2^2 \dfrac{\sum x_i^2}{\sum y_i^2}$, then $r = \hat{\beta}_2 \sqrt{\dfrac{\sum x_i^2}{\sum y_i^2}}$, from Eq.(3.5.6).

Thus, $t = \dfrac{r\sqrt{(n-2)}}{\sqrt{(1-r)^2}} = \dfrac{\hat{\beta}_2 \sqrt{x_i^2}}{\hat{\sigma}}$, and

$$t = F = \frac{r^2(n-2)}{1-r^2} = \hat{\beta}_2^2 \frac{\sum x_i^2}{\hat{\sigma}^2}, \text{ from Eq. (5.9.1)}$$

Empirical Exercises

5.8 (*a*) There is a positive association in the LFPR in 1972 and 1968, which is not surprising in view of the fact since WW II there has been a steady increase in the LFPR of women.

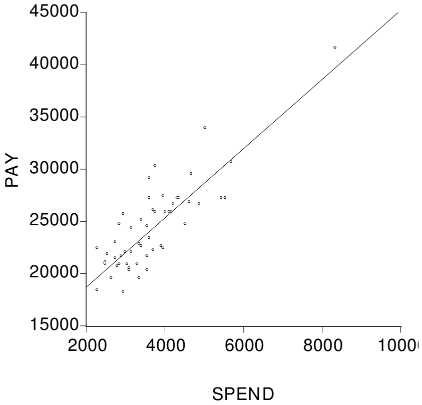
(*b*) Use the one-tail *t test.*

$t = \dfrac{0.6560-1}{0.1961} = -1.7542$. For 17 df, the one-tailed t value at α=5% is 1.740. Since the estimated t value is significant, at this level of significance, we can reject the hypothesis that the true slope coefficient is 1 or greater.

(c) The mean LFPR is : 0.2033 + 0.6560 (0.58) ≈ 0.5838. To establish a 95% confidence interval for this forecast value, use the formula: 0.5838 $\pm$ 2.11(se of the mean forecast value), where 2.11 is the 5% critical t value for 17 df. To get the standard error of the forecast value, use Eq. (5.10.2). But note that since the authors do not give the mean value of the LFPR of women in 1968, we cannot compute this standard error.

(*d*) Without the actual data, we will not be able to answer this question because we need the values of the residuals to plot them and obtain the Normal Probability Plot or to compute the value of the Jarque-Bera test.

5.9 (a)



(b) $\text{Pay}_i = 12129.37 + 3.3076\ \text{Spend}$

$\text{se} = (1197.351)\ (0.3117) \qquad r^2 = 0.6968;\ \text{RSS} = 2.65\text{E}+08$

(c) If the spending per pupil increases by a dollar, the average pay increases by about \$3.31. The intercept term has no viable economic meaning.

(d) The 95% CI for β_2 is: $3.3076 \pm 2(0.3117) = (2.6842, 3.931)$ Based on this CI you will not reject the null hypothesis that the true slope coefficient is 3.

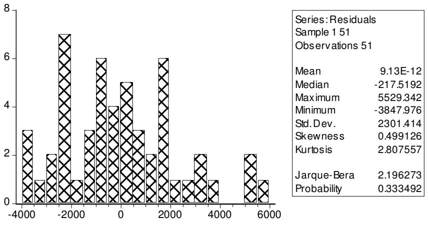
(*e*)The mean and individual forecast values are the same, namely, $12129.37 + 3.3076(5000) \approx 28{,}667$. The standard error of the mean forecast value, using eq.(5.10.2), is 520.5117 (dollars) and the standard error of the individual forecast, using Eq.(5.10.6), is 2382.337. The confidence intervals are:

Mean Prediction: $28{,}667 \pm 2(520.5117)$, that is,
(\$27,626, \$29,708)

Individual Prediction: $28667 \pm 2(2382.337)$, that is,
(\$ 23,902, \$33,432)

As expected, the latter interval is wider than the former.

(*f*)



Series: Residuals
Sample 1 51
Observations 51

Mean	9.13E-12
Median	-217.5192
Maximum	5529.342
Minimum	-3847.976
Std. Dev.	2301.414
Skewness	0.499126
Kurtosis	2.807557
Jarque-Bera	2.196273
Probability	0.333492

The histogram of the residuals can be approximated by a normal curve. The Jarque-Bera statistic is 2.1927 and its *p value* is about 0.33. So, we do not reject the normality assumption on the basis of this test, assuming the sample size of 51 observations is reasonably large.

5.10 The ANOVA table for the *business sector* is as follows:

Source of Variation	SS	df	MSS
Due to Regression(ESS)	91915.2537	1	91915.2537
Due to residual (RSS)	2610.9211	44	59.3391
Total(TSS)	94525.1748		

The F value is $\dfrac{91914.2537}{59.3391} = 1548.9657$

Under the null hypothesis that there is no relationship between wages and productivity in the business sector, this F value follows the F distribution with 1 and 44 df in the numerator and denominator, respectively. The probability of obtaining such an F value is 0.0000, that is, practically zero. Thus, we can reject the null hypothesis, which should come as no surprise.

(*b*) For the *non-farm business sector*, the ANOVA table is as follows:

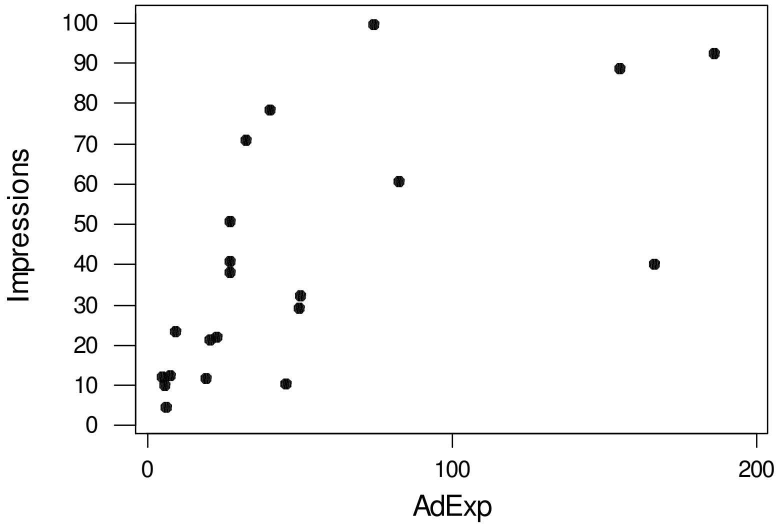
Source of Variation	SS	df	MSS
Due to regression (ESS)	90303.3157	1	90303.3157
Due to residual (RSS)	2714.7626	44	61.6991
Total	93018.0783		

Under the null hypothesis that the true slope coefficient is is zero, the computed F value is:

$$F = \frac{90303.3157}{61.6991} \approx 1463.6071$$

If the null hypothesis were true, the probability of obtaining such an F value is practically zero, thus leading to the rejection of the the null hypothesis.

5.11 (*a*) The plot shown below indicates that the relationship between



the two variables is nonlinear. Initially, as advertising expenditure increases, the number of impressions retained increases, but gradually they taper off.

(b) As a result, it would be inappropriate to fit a bivariate linear regression model to the data. At present we do not have the tools to fit an appropriate model. As we will show later, a model of the type:

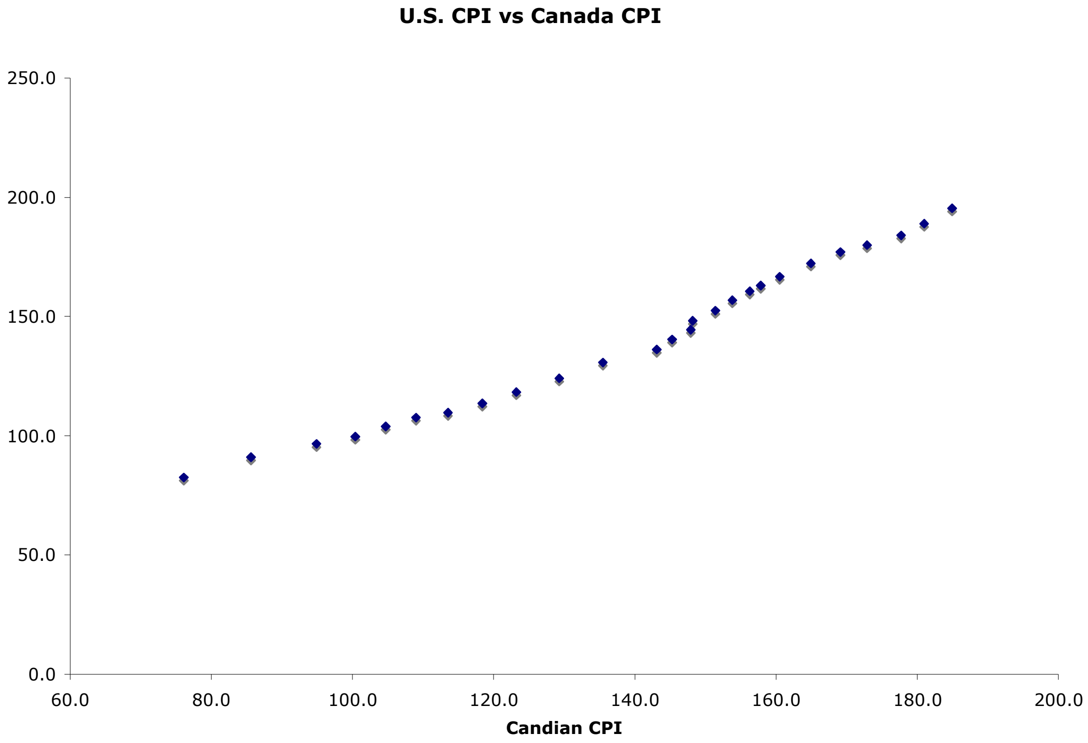
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X^2{}_{2i} + u_i$$

may be appropriate, where Y = impressions retained and X_2 is advertising expenditure. This is an example of a quadratic regression model. But note that this model is still linear in the parameters.

(*c*) The results of blindly using a linear model are as follows:

$$Y_i = 22.163 + 0.3631\ X_i$$
$$\text{se}\quad (7.089)\quad (0.0971) \qquad\qquad r^2 = 0.424$$

5.12 (*a*)



The plot shows that the inflation rates in the two countries generally move together.

(*b*)& (*c*) The following output is obtained from *EViews 3* statistical package.

```
Sample: 1980 2005
Included observations: 26
==========================================================================
        Variable        Coefficient   Std. Error   t-Statistic      Prob.
==========================================================================
           C              -8.5416      4.4795      -1.9068         0.0686
          ICAN             1.0721      0.0316      33.9593         0.0000
==========================================================================
R-squared                  0.9796    F-statistic                1153.2373
Adjusted R-squared         0.9788    Prob(F-statistic)           0.000000
==========================================================================
```

As this output shows, the relationship between the two variables is positive. One can easily reject the null hypothesis that there is no relationship between the two variables, as the *t* value obtained under that hypothesis is 33.9593, and the *p value* of obtaining such a *t* value is practically zero.

Although the two inflation rates are positively related, we cannot infer causality from this finding, for it must be inferred from some underlying economic theory. Remember that regression does not necessarily imply causation.

5.13 (a) The two regressions are as follows:

$$\begin{aligned}
\text{Goldprice}_t &= 215.2856 + 1.0384\ \text{CPI}_t \\
se &= (54.4685)\quad (0.4038) \\
t &= (3.9525)\quad (2.5718) \qquad r^2 = 0.1758
\end{aligned}$$

$$\begin{aligned}
\text{NYSEIndex}_t &= -3444.9920 + 50.2972\ \text{CPI}_t \\
se &= (533.9663)\quad (3.9584) \\
t &= (-6.4517)\quad (12.7066) \qquad r^2 = 0.839
\end{aligned}$$

(b) The Jarqu-Bera statistic for the gold price equation is 5.439 with a p value 0.066. The JB statistic for the NYSEIndex equation is 3.084 with a p value 0.214. At the 5% level of significance, in both cases we do not reject the normality assumption.

(c) Using the usual t test procedure, we obtain:

$$t = \frac{1.0384 - 1}{0.4038} = 0.0951$$

Since this t value does not exceed the critical t value of 2.042, we cannot reject the null hypothesis. The true coefficient is not statistically different from 1.

(d) & (e) Using the usual t test procedure, we obtain:

$$t = \frac{50.297 - 1}{3.958} = 12.455$$

Since this t value exceeds the critical t value of 2.042, we reject the null hypothesis. The estimated coefficient is actually greater than 1. For this sample period, investment in the stock market probably was a hedge against inflation. It certainly was a much better hedge against inflation that investment in gold.

5.14 (*a*) None appears to be better than the others. All statistical results are very similar. Each slope coefficient is statistically significant at the 99% level of confidence.

(*b*) The consistently high r^2s cannot be used in deciding which monetary aggregate is best. However, this does not suggest that it makes no difference which equation to use.

(*c*) One cannot tell from the regression results. But lately the Fed seems to be targeting the M2 measure.

5.15 Write the indifference curve model as:

$$Y_i = \beta_1(\frac{1}{X_i}) + \beta_2 + u_i$$

Note that now β_1 becomes the slope parameter and β_2 the intercept. But this is still a linear regression model, as the parameters are linear (more on this in Ch.6). The regression results are as follows:

$$\hat{Y}_i = 3.2827(\frac{1}{X_i}) + 1.1009$$

$$se = (1.2599) \qquad (0.6817) \qquad r^2 = 0.6935$$

The "slope" coefficient is statistically significant at the 92% confidence coefficient. The marginal rate of substitution (MRS) of Y for X is: $\frac{\partial Y}{\partial X} = -0.3287\left(\frac{1}{X_i^2}\right)$.

5.16 (*a*) Let the model be: $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

where *Y* is the actual exchange rate and *X* the implied PPP. If the PPP holds, one would expect the intercept to be zero and the slope to be one.

(*b*) The regression results are as follows:

$$\hat{Y}_i = -33.0917 + 1.8147\ X_i$$

$$se = (26.9878) \quad (0.0274)$$

$$t = (-1.2262) \quad (66.1237) \qquad r^2 = 0.9912$$

To test the hypothesis that $\beta_2 = 1$, we use the t test, which gives

$$t = \frac{1.8147 - 1}{0.0274} = 29.7336$$

This *t* value is highly significant, leading to the rejection of the null hypothesis. Actually, the slope coefficient is is greater than 1. From the given regression, the reader can easily verify that the intercept coefficient is not different from zero, as the *t* value under the hypothesis that the true intercept is zero, is only -1.2262.

Note: Actually, we should be testing the (joint) hypothesis that the intercept is zero and the slope is 1 simultaneously. In Ch. 8, we will show how this is done.

(*c*) Since the Big Max Index is "crude and hilarious" to begin with, it probably doesn't matter. However, for the sample data, the results do not support the theory.

5.17 (*a*) Letting Y represent the male math score and X the female math score, we obtain the following regression:

$$\hat{Y}_i = 198.737 + 0.6704 X_i$$
$$se = (12.875) \quad (0.0265)$$
$$t = (15.435) \quad (25.332) \qquad\qquad r^2 = 0.9497$$

(*b*) The Jarque-Bera statistic is 1.1641 with a *p value* of 0.5588. Therefore, asymptotically we cannot reject the normality assumption.

(*c*) $t = \dfrac{0.6704 - 1}{0.0265} = -12.4377$. Therefore, with 99% confidence we can reject the hypothesis that $\beta_2 = 1$.

(*d*) The ANOVA table is:

Source of Variation	SS	df	MSS
ESS	1605.916	1	1605.916
RSS	85.084	34	2.502
TSS	1691	35	

Under the null hypothesis that $\beta_2 = 0$, the F value is 641.734, The *p value* of obtaining such an F value is almost zero, leading to the rejection of the null hypothesis.

5.18 (*a*) The regression results are as follows:

$$\hat{Y}_i = 132.778 + 0.750\ X_i$$
$$se = (33.724) \quad (0.067)$$
$$t = (3.937) \quad (11.187) \qquad r^2 = 0.786$$

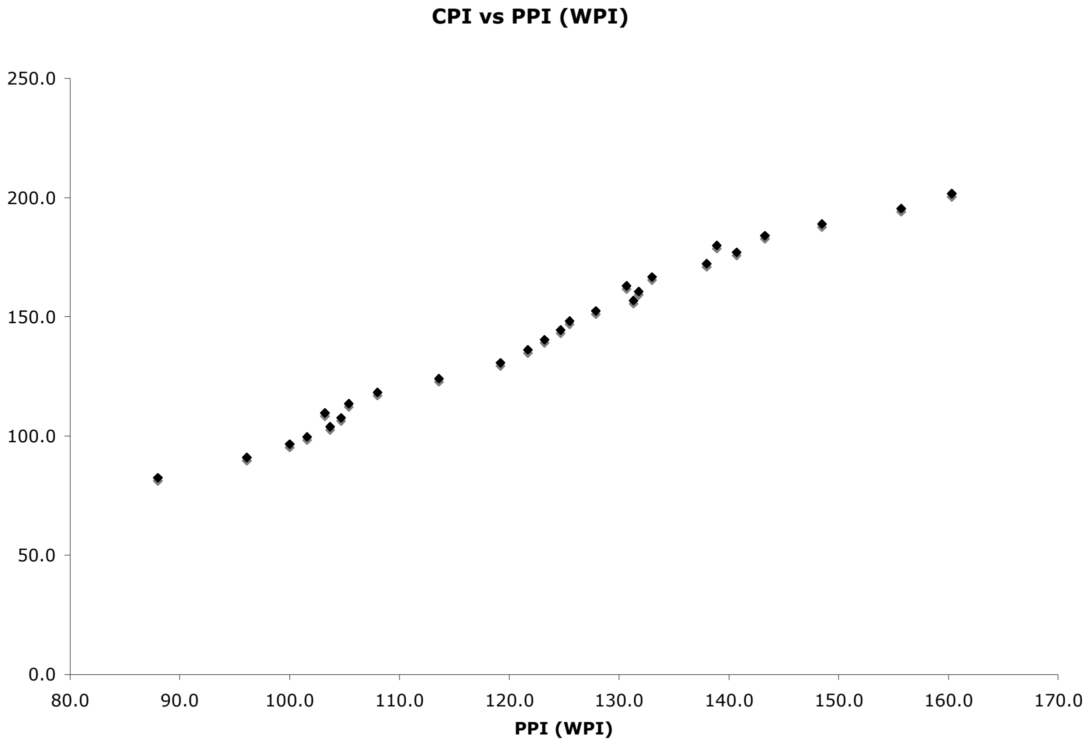
(*b*) The Jarque-Bera statistics is 1.122 with a *p value* of 0.571. Therefore we can reject the null hypothesis of non-normality.

(*c*) Under the null hypothesis, we obtain: $t = \dfrac{0.750 - 1}{0.067} = -3.7313$.

The critical *t* value at the 5% level is 2.042 (or -2.042). Therefore, we can reject the null hypothesis that the true slope coefficient is 1.

(*d*) The ESS, RSS, and TSS values are, respectively, 1005.75 (l df), 273.222 (34 df), and 1278.972 (35 df). Under the usual null hypothesis the F value is 125.156. The *p value* of such an F value is almost zero. Therefore, we can reject the null hypothesis that there is no relationship between the two variables.

5.19 (*a*)



The scattergram as well is shown in the above figure.

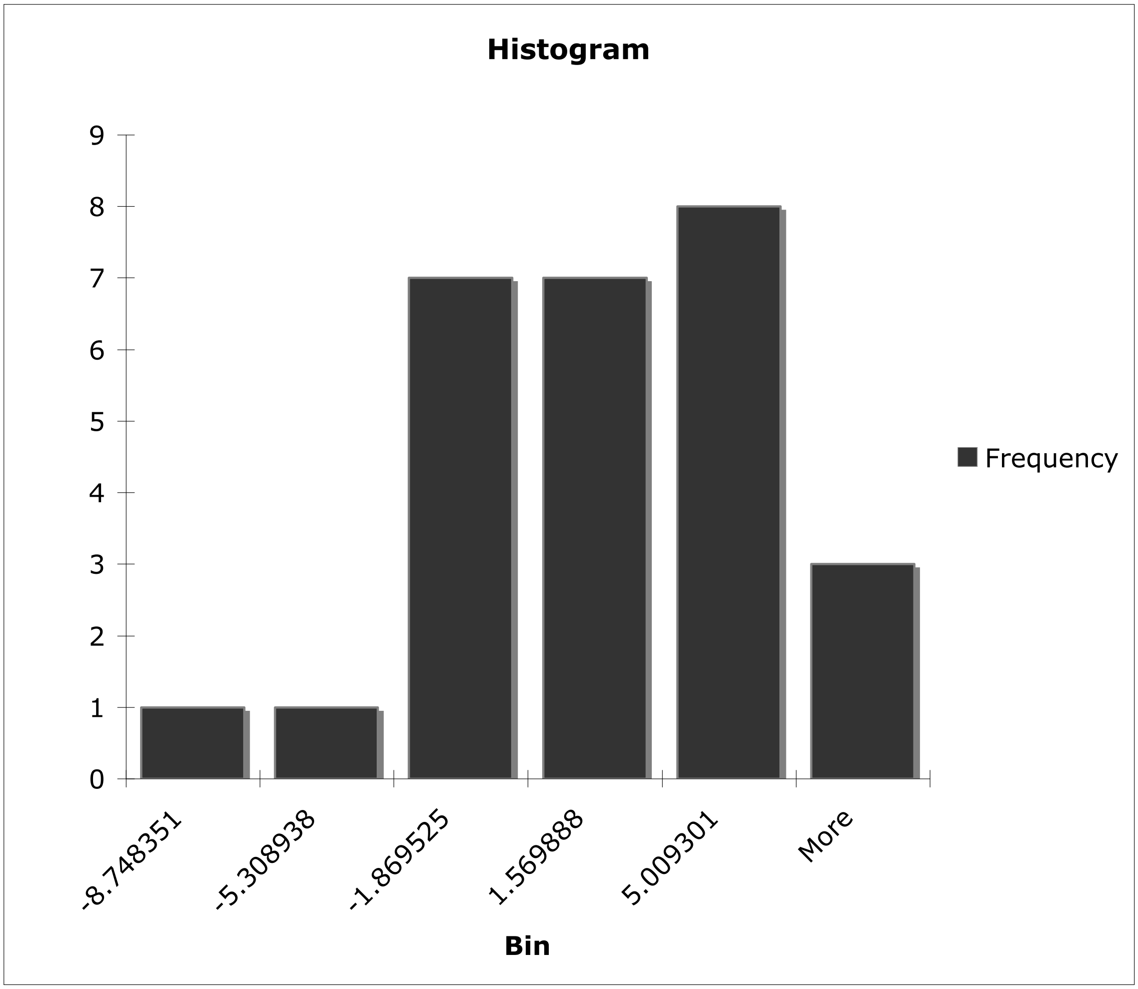
(*b*) Treat CPI as the regressand and WPI as the regressor. The CPI represents the prices paid by the consumers, whereas the WPI represents the prices paid by the producers. The former are usually a markup on the latter.

(c) & (d) The following output obtained from *Eviews6* gives the necessary data.

```
Dependent Variable: CPI
Method: Least Squares
Sample: 1980 2006
Included observations: 27
CPI=C(1)+C(2)*PPI

                    Coefficient  Std. Error    t-Statistic    Prob.

C(1)                -81.01611    5.492246      -14.75100      0.0000
C(2)                1.817620     0.044181      41.14020       0.0000

R-squared               0.985444          Mean dependent var         142.3963
Adjusted R-squared      0.984862          S.D. dependent var         34.67915
S.E. of regression      4.266824          Akaike info criterion      5.810804
Sum squared resid       455.1447          Schwarz criterion          5.906792
Log likelihood          -76.44585         Durbin-Watson stat         0.601660
```

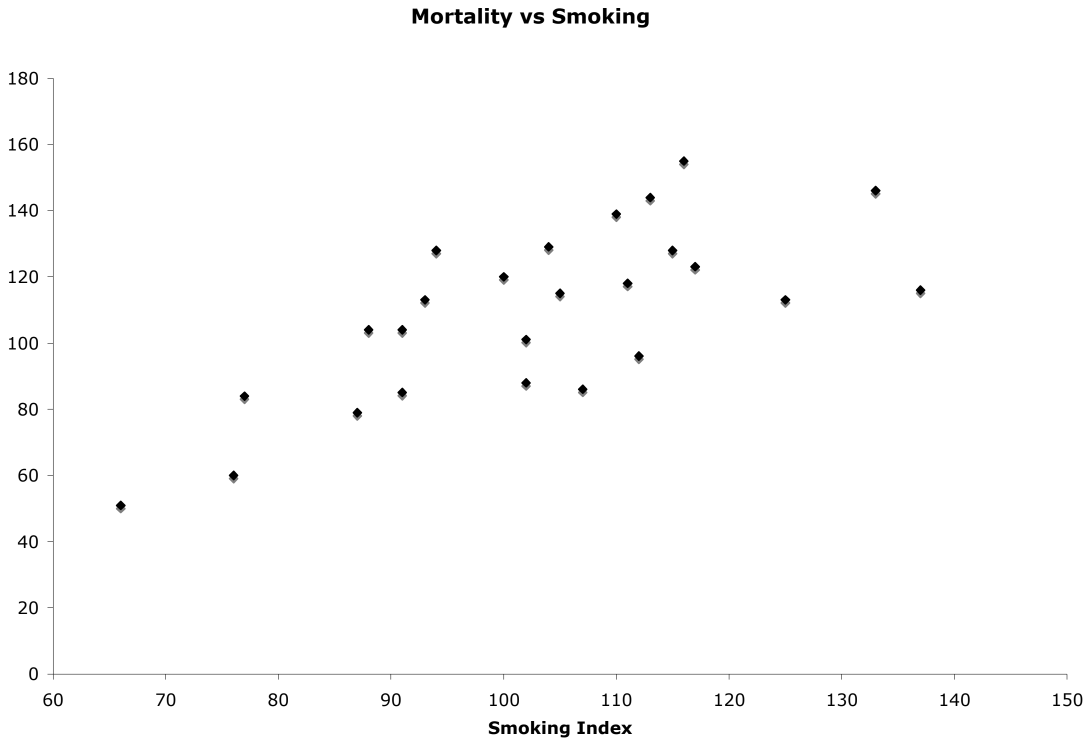
The estimated *t* value of the slope coefficient is 1.8176 under the null hypothesis that there is no relationship between the two indexes. The *p value* of obtaining such a *t* value is almost zero, suggesting the rejection of the null hypothesis.

The histogram and Jarque-Bera test based on the residuals from the preceding regression are given in the following diagram.



The Jarqe-Bera statistic is 0.3927 with a *p value* 0.8217. Therefore, we cannot reject the normality assumption. The histogram also shows that the residuals are slightly left-skewed, but not too far from symmetric.

5.20 (a) There seems to be a general positive relationship between Smoking and Mortality.



(b)

$$\hat{Y}_i = -2.8853 + 1.0875\ X_i$$
$$se = \ (23.0337)\quad(0.2209)$$
$$t = \ (-0.1253)\quad(4.9222)\qquad r^2 = 0.5130$$

(c) The slope coefficient has a *t* statistic of 4.9222, which indicates a *p value* of almost 0. Therefore, we can reject the null hypothesis and conclude that Smoking is related to Mortality at the 5% level of significance.

(d) The riskiest occupations seem to be Furnace forge foundry workers, Construction workers, and Painters and decorators. One reason for why these occupations are more risky could be that they all work around toxic fumes and/or chemicals and therefore breathe in dangerous toxins frequently.

(e) Unless there is a way to categorize the occupations into fewer groups, we cannot include them in the regression analysis (this will be addressed later in the discussion of dummy, or indicator, variables in chapter 9).