CHAPTER 8

MULTICOLLINEARITY: WHAT HAPPENS IF EXPLANATORY VARIABLES ARE CORRELATED?

QUESTIONS

8.1. An exact linear relationship between two or more (explanatory) variables; more than one exact linear relationship between two or more explanatory variables.

8.2. In perfect collinearity there is an *exact linear* relationship between two or more variables, whereas in imperfect collinearity this relationship is not exact but an approximate one.

8.3. Since 1 foot = 12 inches, there is an exact linear relationship between the variables "height in inches" and "height in feet", if both variables are included in the same regression. In this case, we have only one independent explanatory variable and not two.

8.4. Disagree. The variables X^2 and X^3 are *nonlinear* functions of X. Hence, their inclusion in the regression model does not violate the assumption of the classical linear regression model (CLRM) of "no exact *linear* relationship among explanatory variables."

8.5. Consider, for instance, Eq. (4.21). Let $x_{3i} = 2x_{2i}$ .Substituting this into Equation (4.21), we obtain:

$$b_2 = \frac{(\sum yx_2)(4\sum x_2^2) - (2\sum yx_2)(2\sum x_2^2)}{(\sum x_2^2)(4\sum x_2^2) - (2\sum x_2^2)^2} = \frac{0}{0}$$

which is an indeterminate expression. The same is true of Equations (4.22), (4.25), and (4.27).

8.6. OLS estimators are still BLUE.

8.7. (1) Large variances and covariances of OLS estimators

(2) Wider confidence intervals

(3) Insignificant t ratios

(4) A high R^2 but few significant t ratios

(5) Sensitivity of OLS estimators and their standard errors to changes in the data.

(6) Wrong signs for regression coefficients.

(7) Difficulty in assessing the individual contributions of the explanatory variables to the ESS or R^2.

8.8. The VIF measures the increase in the variances OLS estimators as the degree of collinearity, as measured by R^2, increases. If the (explanatory) variables are uncorrelated, the least value of VIF is 1, but if the variables are perfectly correlated, VIF is infinite.

8.9. (***a***) large; small

(***b***) undefined; undefined

(***c***) variances

8.10. (***a***) *False*. In cases of perfect multicollinearity, OLS estimators are not even defined.

(***b***) *True*.

(***c***) *Uncertain*. A high R^2 can be offset by a low σ^2 or a high variance of the relevant explanatory variable included in the model, or both.

(***d***) *True*. A simple correlation between two explanatory variables may be high, but when account is taken of other explanatory variables in the model, the partial correlation between those two variables may be low.

(***e***) *Uncertain*. It is true only if the collinearity observed in the given sample continues to hold in the post sample period. If that is not the case, then this statement is false.

8.11. This is because business cycles and or trends (reflecting growth rates) dominate most economic time series. Therefore, in a regression of the consumer price index (CPI) on the money supply and the unemployment rate, the latter two variables are likely to exhibit collinearity.

8.12. (***a***) Yes. In the course of a business cycle, variables such as income at times t and (t -1) usually tend to move in the same direction. Thus, in the upswing

phase of a business cycle, income this period is generally greater than the income in the previous time period.

(*b*) There are various methods of resolving the problem, such as the Koyck transformation, or the first difference transformation. Some of these methods are discussed in Chapter 16.

PROBLEMS

8.13. **(*a*)** No, because $X_3 = 2X_2 - 1$, which is perfect collinearity.

(*b*) $Y = B_1 + B_2X_2 + B_3(2X_2 - 1) + u$

$= A_1 + A_2X_2 + u$

where $A_1 = (B_1 - B_3)$ and $A_2 = (B_2 + 2B_3)$.

If we regress Y on X_2, we can obtain estimates of the combinations of the Bs as shown above, but not individual estimates of the Bs. Incidentally, the regression of Y on X_2 is:

$$\hat{Y} = -12.0 + 2.0\ X_2$$

Therefore, $(B_1 - B_3) = -12.0$ and $(B_2 + 2B_3) = 2.0$.

8.14. **(*a*)** X_2 is a product specific price index, whereas X_3 is the general price index. It is possible that the two indexes may not move together if there is a lead-lag relationship between the two.

(*b*) It is an indicator of employment conditions in the labor markets. *Ceteris paribus*, the higher the level of employment, the higher the demand for automobiles will be.

(*c*) Since we are dealing with a log-linear model, the partial slope coefficients are partial elasticities of the dependent variable with respect to the given variables.

(*d*) Running the logarithmic regression with $\ln Y_t$ as the dependent variable, and including all the variables, we obtain the following results for the aggregate demand function for passenger cars:

Variable name	Coefficient	*t* value
Constant	11.0582	0.5086 **
$\ln X_{2t}$	1.9409	2.1099 **
$\ln X_{3t}$	-4.6815	-2.5475 *
$\ln X_{4t}$	2.7164	1.8438 **
$\ln X_{5t}$	-0.0259	-0.2106 **
$\ln X_{6t}$	-0.5821	-0.2496 **
$R^2 = 0.8551$		

* Significant at the 5% level (two-tailed);
** Not significant at the 5% level (two-tailed).

8.15. From the results given in problem 8.14, multicollinearity may be present in the data. First, the R^2 value is reasonably high, but only one *t* value is statistically significant. Second, the general price index (X_3) has a negative sign, but the new car price index (X_2) has a positive sign. The latter may not make economic sense. Thirdly, neither the income variable (X_4) nor the employment variable (X_6) has any impact on the demand for autos, a rather surprising result. The interest rate (X_5) is also insignificant.

8.16 If you regress the natural log of each explanatory variable on the natural logs of the remaining explanatory variables, you will find that the R^2s of all these auxiliary regressions are very high, as the following table shows:

Dependent variable	Independent variables	R^2
$\ln X_{2t}$	$\ln X_{3t}$, $\ln X_{4t}$, $\ln X_{5t}$, $\ln X_{6t}$	0.9963
$\ln X_{3t}$	$\ln X_{2t}$, $\ln X_{4t}$, $\ln X_{5t}$, $\ln X_{6t}$	0.9995
$\ln X_{4t}$	$\ln X_{2t}$, $\ln X_{3t}$, $\ln X_{5t}$, $\ln X_{6t}$	0.9995
$\ln X_{5t}$	$\ln X_{2t}$, $\ln X_{3t}$, $\ln X_{4t}$, $\ln X_{6t}$	0.8734
$\ln X_{6t}$	$\ln X_{2t}$, $\ln X_{3t}$, $\ln X_{4t}$, $\ln X_{5t}$	0.9961

8.17. The simple correlation matrix of the natural logs of the X variables is:

	$\ln X_{2t}$	$\ln X_{3t}$	$\ln X_{4t}$	$\ln X_{5t}$	$\ln X_{6t}$
$\ln X_{2t}$	1.0000				
$\ln X_{3t}$	0.9960	1.0000			
$\ln X_{4t}$	0.9931	0.9964	1.0000		
$\ln X_{5t}$	0.5850	0.6138	0.5850	1.0000	
$\ln X_{6t}$	0.9737	0.9740	0.9868	0.5995	1.0000

Since the civilian employment (X_6) and disposable personal income (X_4) are likely to move together, one of them can be dropped from the model; notice that the correlation between the logs of these two variables is 0.9868. Similarly, since the two price indexes X_2 and X_3 are also likely to move together, one of them can be dropped; the simple correlation between the logs of these variables is 0.9960. But keep in mind the warning given in the text that simple correlations are not infallible indicators of multicollinearity. Also, keep in mind the "omission of relevant variables" bias if we drop one or more of these variables.

8.18. The following models may be acceptable on the basis of the usual economic (i.e., signs of the variables) and statistical criteria:

$$\widehat{\ln Y_t} = -22.104 - 1.038 \ln X_{2t} - 0.295 \ln X_{5t} + 3.244 \ln X_{6t}$$
$$t = (-2.640) \quad (-3.143) \quad (-4.002) \quad (3.719)$$
$$R^2 = 0.6849$$

$$\widehat{\ln Y_t} = -27.755 - 0.904 \ln X_{3t} - 0.251 \ln X_{5t} + 3.692 \ln X_{6t}$$
$$t = (-3.876) \quad (-4.491) \quad (-4.074) \quad (5.165)$$
$$R^2 = 0.7857$$

Compared with the original model, these two models have the correct signs for the various coefficients and all the individual coefficients are statistically significant. It is true that the R^2s of the these two models are not as high as

that of the original model. Therefore, for forecasting purposes the original model might be better, provided the collinearity observed in the sample continues in the future. But that is a big proviso.

8.19. Prices of used cars, expenditure on advertising, a dummy variable to represent regional variation, import restrictions on foreign cars, and special incentives offered by the auto manufacturers (e.g., zero-interest financing or instant cash rebates) are some of the relevant variables that may further explain the demand for automobiles. But keep in mind that we need many more observations to include all these variables, assuming that the data on some of these variables are available.

8.20. **(*a*)** The slope coefficients in the first model are partial elasticities. In the second model, the coefficients of log K and log H are, as well, elasticities. The coefficient of the trend variable, t, suggests that, holding other things constant, (the index of) production has been increasing at the annual rate of about 2.7%.

(*b*) The t values of the regression coefficients are, respectively, -3.600, 10.195, and 6.518. For 18 d.f., the t values are significant at the 5% level, since the critical t value is 2.101.

(*c*) The t values of the trend variable and log K are 1.333 and 1.381, respectively, which are not statistically significant.

(*d*) It may be that the trend variable (perhaps representing technology) and log K are collinear.

(*e*) Even though a high pairwise correlation does not necessarily suggest collinearity, sometimes this may be the case.

(*f*) This hypothesis can be rejected, for the F value (using the R^2 variant) is 45.3844, which is significant beyond the 1% level, for the 1% critical F value for 3 and 17 d.f, $F_{3,17}$, is 5.18.

(*g*) The returns to scale are: 0.887 + 0.893 = 1.780, that is, increasing returns to scale.

8.21. For instance, we have:

$$\text{var}(b_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}\sigma^2$$

Now: $$r_{23}^2 = \frac{(\sum x_{2i}x_{3i})^2}{(\sum x_{2i}^2)(\sum x_{3i}^2)}$$

Substituting the latter into the former, we obtain, after simple algebraic manipulations, Equation (8.12). The same holds true of Equation (8.13).

8.22. (***a***) $\hat{Y} = 24.3370 + 0.8716\,X_2 - 0.0350\,X_3$

$t = (3.8753)\quad (2.7726)\qquad (-1.1604)\qquad\qquad R^2 = 0.9682$

(***b***) Collinearity may be present in the data, because despite the high R^2 value, only the coefficient of the income variable is statistically significant. In addition, the wealth coefficient has the wrong sign.

(***c***) $\hat{Y} = 24.4545 + 0.5091\,X_2$

$t = (3.8128)\quad (14.2432)\qquad\qquad r^2 = 0.9621$

$\hat{Y} = 26.4520 + 0.0480\,X_3$

$t = (3.1318)\quad (10.5752)\qquad\qquad r^2 = 0.9332$

Now individually both slope coefficients are statistically significant and they each have the correct sign.

(***d***) $\hat{X}_3 = -3.3636 + 10.3727\,X_2$

$t = (-0.0456)\quad (25.2530)\qquad\qquad r^2 = 0.9876$

This regression shows that the two variables are highly collinear.

(***e***) We can drop either X_2 or X_3 from the model. But keep in mind that in that case we will be committing a specification error. The problem here is that our sample is too small to isolate the individual impact of income and wealth on consumption expenditure.

8.23. Let $Y' = Y - 0.9$ earnings. Using the data given in Table 8-1, we obtain:

$$\hat{Y}' = -220.2613 - 0.3527\,X_2$$

$t = (-202.1192)\quad (-2.0080)\qquad\qquad r^2 = 0.9483$

These results are vastly different from the ones in Equation (8.8), showing that unless the prior information is reliable, one can obtain dubious results.

Note: The r^2 value given has been corrected so that it can be directly compared with the r^2 value obtained from Equation (8.8).

8.24. Use the formula: $F = \dfrac{R^2/3}{(1-R^2)/19}$, which follows the usual F distribution with 3 and 19 d.f., respectively, in the numerator and denominator. The results of the F test will show that all the R^2s shown in Table 8-4 are highly statistically significant (at the 1% level, $F_{3,19} = 5.01$).

8.25. In Problem 7.19 we showed that when all the explanatory variables are included in the model, there is collinearity among these explanatory variables. There we also gave another version of the model. Variables such as education and median household earnings are likely to be correlated, as per human capital theory of labor economics. Likewise, spending per pupil is likely to be correlated with median household income. Wealthier school districts generally spend more on schooling. It is left for the reader to develop suitable models taking into account these factors.

8.26. A priori, some multicollinearity might be expected in this regression since the variables are quite related to each other. Including the required variables in the model, and using *EViews*, we obtained the following regression results:

Dependent Variable: ASP
Sample: 1 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	73669.310	54193.296	1.359	0.181
GPA	-733.771	12621.572	-0.058	0.954
GMAT	-101.673	81.618	-1.246	0.220
PCTACCEPT	-4689.750	14042.724	-3.182	0.003
TUITION	0.4944	0.1859	2.659	0.011
RATING	27256.581	4695.142	5.805	0.000
R-squared	0.870004			

In addition, we can generate the Analysis of Variance output from *Excel*, which is as follows:

Source of variation	SS	df	MS	*F*	*p*-value
Regression	10358466314	5	2071693263	51.579	0.000
Residual	1727118549	43	40165547.65		
Total	12085584863	48			

Note: In the source of variation, Regression is ESS, Residual is RSS, and Total is TSS.

As these results suggest, the coefficients of GPA and GMAT are not statistically significant, perhaps due to collinearity. There seems to be collinearity among other variables. If we include only the percentage accepted, tuition, and recruiter rating as variables, we get the following results:

Dependent Variable: ASP
Sample: 1 47

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12908.672	13896.631	0.929	0.358
PCTACCEPT	-34871.744	11280.574	-3.091	0.003
TUITION	0.498	0.165	3.011	0.004
RATING	23491.321	3657.113	6.423	0.000
R-squared	0.851			

You are invited to use the data to develop other models.

8.27. ***(a)*** Model 1: A one percent increase in the GDP in the U.K. should correspond to about a 193% increase in imports, whereas a one percent increase in the CPI should result in about 27% increase in imports. Note that only the log GDP variable seems to be significant.

Model 2: A one percent increase in the GDP in the U.K. should correspond to about a 197% increase in imports, a one percent increase in the CPI should result in about 103% increase in imports, and a one percent increase in PPI would decrease imports by about 77%. Note that all three variables seem to be statistically significant.

Model 3: A one percent increase in the GDP in the U.K. should correspond to about a 209% increase in imports, whereas a one percent increase in PPI

would increase imports by about 12%. Note that only the log GDP variable seems to be statistically significant.
(b) This actually does not make economic sense.

(c) This also does not make economic sense.

(d) This is most likely due to inherent multicollinearity between variables in the model.

(e) Yes, as discussed above.

(f) Yes it would, especially since that level of correlation is greater than the overall correlation between all the independent variables and log Imports.

(g) It is unnecessary to include both CPI and PPI variables as the level of correlation between them indicates potential multicollinearity and a redundancy of information. Therefore, the choice is between models 1 and 3. Since the R-squared value is higher in model 1 and both independent variables are statistically significant, it is probably the better choice.

8.28. ***(a)***

Dependent Variable: LIMPORTS
Sample: 1970 1998
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.975260	0.782070	2.525683	0.0180
LGDP	1.043167	0.405783	2.570749	0.0162
LCPI	0.446142	0.569840	0.782925	0.4407

R-squared	0.982318	Mean dependent var	12.49048
Adjusted R-squared	0.980958	S.D. dependent var	0.904848
S.E. of regression	0.124862	Akaike info criterion	-1.225512
Sum squared resid	0.405356	Schwarz criterion	-1.084068
Log likelihood	20.76993	F-statistic	722.2174
Durbin-Watson stat	0.461405	Prob(F-statistic)	0.000000

Dependent Variable: LN_IMPORTS
Sample: 1975 2005
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.409416	0.270074	5.218629	0.0000
LN_GDP	1.850099	0.182912	10.11471	0.0000
LN_CPI	-0.873369	0.284805	-3.066548	0.0048

R-squared	0.992005	Mean dependent var	13.08472
Adjusted R-squared	0.991434	S.D. dependent var	0.762092
S.E. of regression	0.070532	Akaike info criterion	-2.373736
Sum squared resid	0.139293	Schwarz criterion	-2.234963
Log likelihood	39.79291	F-statistic	1737.193
Durbin-Watson stat	0.650260	Prob(F-statistic)	0.000000

(*b*) Judged by the high R^2 value and the negative coefficient on the log CPI variable, there *might* some multicollinearity in the data.

(*c*)

Dependent Variable: LN_IMPORTS
Sample: 1975 2005
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.001910	0.214308	9.341269	0.0000
LN_GDP	1.293252	0.024951	51.83130	0.0000

R-squared	0.989321	Mean dependent var	13.08472
Adjusted R-squared	0.988952	S.D. dependent var	0.762092
S.E. of regression	0.080102	Akaike info criterion	-2.148687
Sum squared resid	0.186074	Schwarz criterion	-2.056171
Log likelihood	35.30465	F-statistic	2686.484
Durbin-Watson stat	0.517695	Prob(F-statistic)	0.000000

Dependent Variable: LN_IMPORTS
Sample: 1975 2005
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.578293	0.348057	10.28076	0.0000
LN_CPI	1.986499	0.072515	27.39449	0.0000

R-squared	0.962795	Mean dependent var	13.08472
Adjusted R-squared	0.961512	S.D. dependent var	0.762092
S.E. of regression	0.149510	Akaike info criterion	-0.900561
Sum squared resid	0.648248	Schwarz criterion	-0.808045
Log likelihood	15.95869	F-statistic	750.4582
Durbin-Watson stat	0.279792	Prob(F-statistic)	0.000000

Dependent Variable: LN_GDP
Sample: 1975 2005
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.172304	0.166696	7.032591	0.0000
LN_CPI	1.545793	0.034730	44.50942	0.0000

R-squared	0.985573	Mean dependent var	8.569723
Adjusted R-squared	0.985075	S.D. dependent var	0.586128
S.E. of regression	0.071605	Akaike info criterion	-2.372953
Sum squared resid	0.148693	Schwarz criterion	-2.280437
Log likelihood	38.78076	F-statistic	1981.089
Durbin-Watson stat	0.199406	Prob(F-statistic)	0.000000

The auxiliary regression of LN_GDP on LN_CPI shows that the two variables are highly correlated, perhaps suggesting that the data suffer from the collinearity problem.

(***d***) The best solutions here would be to express imports and GDP in real terms by dividing each by CPI (recall the ratio method discussed in the chapter). The results are as follows:

Dependent Variable: LN(IMP/CPI)
Sample: 1975 2005
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.442445	0.221017	6.526390	0.0000
LN(GDP/CPI)	1.811942	0.058312	31.07304	0.0000

R-squared	0.970841	Mean dependent var	8.299204
Adjusted R-squared	0.969835	S.D. dependent var	0.399385
S.E. of regression	0.069365	Akaike info criterion	-2.436517
Sum squared resid	0.139535	Schwarz criterion	-2.344002
Log likelihood	39.76602	F-statistic	965.5338
Durbin-Watson stat	0.647203	Prob(F-statistic)	0.000000

8.29. (***a***) and (***c***)Examining the correlation coefficients between the possible explanatory variables, one observes a very high correlation between the new car CPI and the general CPI (0.997) and between PDI and the new car CPI (0.991). Others are relatively high, but they should remain in the model for theoretical reasons. PDI is also closely related to the employment level, the

correlation between the two being 0.972. Therefore, one could drop general CPI and PDI and estimate the following model.

Dependent Variable: LY
Sample: 1971 1986
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-22.10374	8.373593	-2.639696	0.0216
LX2	-1.037839	0.330227	-3.142805	0.0085
LX5	-0.294929	0.073704	-4.001514	0.0018
LX6	3.243886	0.872231	3.719068	0.0029

R-squared	0.684855	Mean dependent var	9.204273
Adjusted R-squared	0.606069	S.D. dependent var	0.119580
S.E. of regression	0.075053	Akaike info criterion	-2.128930
Sum squared resid	0.067595	Schwarz criterion	-1.935783
Log likelihood	21.03144	F-statistic	8.692569
Durbin-Watson stat	1.309678	Prob(F-statistic)	0.002454

Note: The letter L stands for the "logarithm of."
It seems this model does not suffer from the collinearity problem.

(*b*) If we include all the X variables, we obtain the following results:

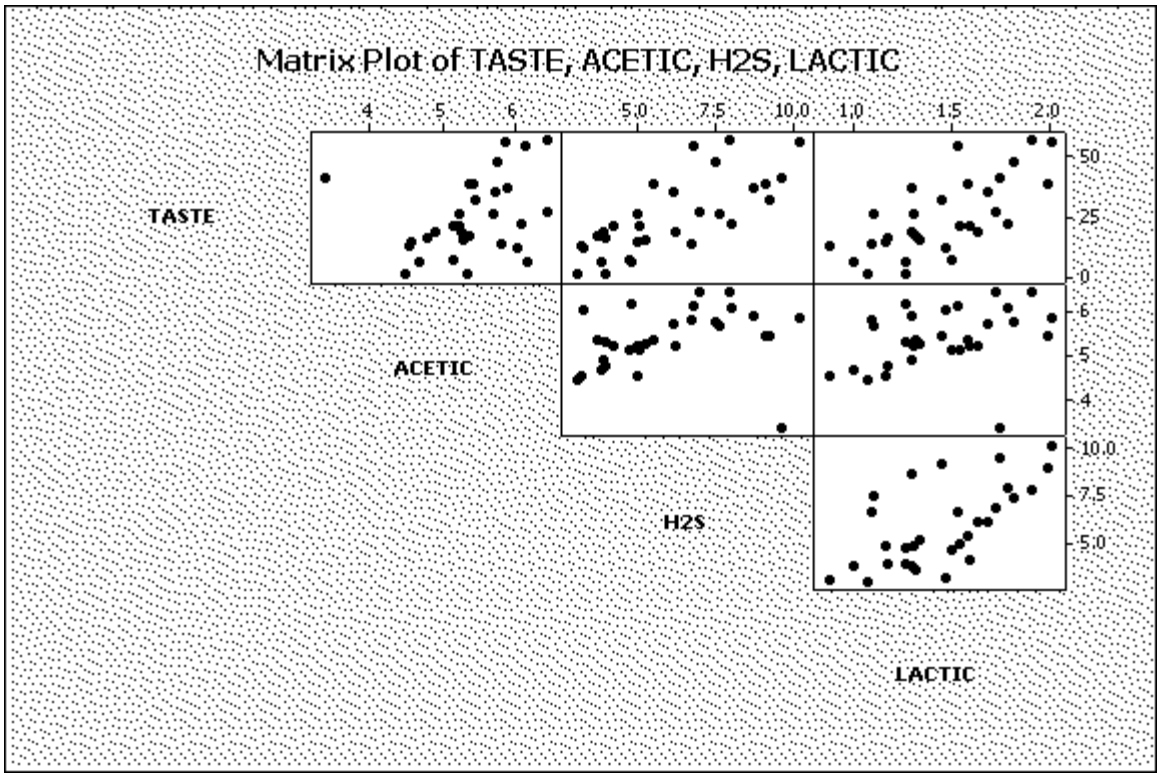
Dependent Variable: LOG(Y)
Sample: 1971 1986
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.254859	19.11656	0.170264	0.8682
LOG(X2)	1.790153	0.873240	2.050012	0.0675
LOG(X3)	-4.108518	1.599678	-2.568341	0.0280
LOG(X4)	2.127199	1.257839	1.691154	0.1217
LOG(X5)	-0.030448	0.121848	-0.249884	0.8077
LOG(X6)	0.277792	2.036975	0.136375	0.8942

R-squared	0.854803	Mean dependent var	9.204273
Adjusted R-squared	0.782205	S.D. dependent var	0.119580
S.E. of regression	0.055806	Akaike info criterion	-2.653874
Sum squared resid	0.031143	Schwarz criterion	-2.364153
Log likelihood	27.23099	F-statistic	11.77442
Durbin-Watson stat	1.793020	Prob(F-statistic)	0.000624

Clearly, this model suffers from collinearity, as suspected.

8.30 ***(a)***



(b)

Dependent Variable: TASTE
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-25.67618	16.32978	-1.572353	0.1275
ACETIC	3.253508	3.113979	1.044807	0.3054
H2S	5.499416	0.980733	5.607455	0.0000

R-squared	0.587826	Mean dependent var	24.53333
Adjusted R-squared	0.557294	S.D. dependent var	16.25538
S.E. of regression	10.81570	Akaike info criterion	7.694514
Sum squared resid	3158.444	Schwarz criterion	7.834634
Log likelihood	-112.4177	F-statistic	19.25315
Durbin-Watson stat	1.111606	Prob(F-statistic)	0.000006

Here it seems that only the H2S variable is significant.

(c)

Dependent Variable: TASTE
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-27.59182	8.981825	-3.071961	0.0048
LACTIC	19.88720	7.959009	2.498704	0.0188
H2S	3.946267	1.135692	3.474768	0.0017

R-squared	0.651702	Mean dependent var	24.53333
Adjusted R-squared	0.625903	S.D. dependent var	16.25538
S.E. of regression	9.942362	Akaike info criterion	7.526126
Sum squared resid	2668.965	Schwarz criterion	7.666246
Log likelihood	-109.8919	F-statistic	25.25995
Durbin-Watson stat	1.581086	Prob(F-statistic)	0.000001

This time both variables appear to be significant.

(d)

Dependent Variable: TASTE
Sample: 1 30
Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-34.13491	15.67628	-2.177488	0.0387
ACETIC	1.538645	3.000501	0.512796	0.6124
H2S	3.915241	1.153106	3.395386	0.0022
LACTIC	18.80235	8.342614	2.253772	0.0329

R-squared	0.655190	Mean dependent var	24.53333
Adjusted R-squared	0.615404	S.D. dependent var	16.25538
S.E. of regression	10.08091	Akaike info criterion	7.582729
Sum squared resid	2642.242	Schwarz criterion	7.769556
Log likelihood	-109.7409	F-statistic	16.46793
Durbin-Watson stat	1.441891	Prob(F-statistic)	0.000003

Again it seems that Acetic is not statistically significant within this mutiple regression.

(e) and ***(f)*** It very well may be that the Acetic variable is highly linearly related to H2S. It could be useful to check the significance of Acetic on its own for predicting taste. (this is left to the reader) Due to the lack of coefficient significance, it would be

wiser to choose the model with only Lactic and H2S. On a side note, the R-squared value here is extremely close to that of the last model, also validating this model choice.

8.31 ***(a)*** *Minitab* results are:

```
Regression Analysis: ln Salary versus ln Profit, ln Turnover

The regression equation is
ln Salary = 6.25 + 0.179 ln Profit + 0.0022 ln Turnover

Predictor       Coef  SE Coef      T      P
Constant      6.2516   0.2059  30.36  0.000
ln Profit    0.17873  0.04100   4.36  0.000
ln Turnover  0.00216  0.03978   0.05  0.957

S = 0.320025   R-Sq = 38.6%   R-Sq(adj) = 37.1%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2   5.2156  2.6078  25.46  0.000
Residual Error  81   8.2957  0.1024
Total           83  13.5113
```

(b) Only the ln Profit variable is statistically significant in this model.

(c) Together, the variables' coefficients are statistically significant. This can be seen through the F-test results, indicating a p-value of 0.000. This test assesses whether the *group* of variables contains any significant relationship to ln Salary.

(d) Since the answer to (c) was yes but the answer to (b) was no, there may be some multicollinearity between the variables in the model.

(e) Since there are only two variables in the model, assessment of the correlation coefficient could help to identify this issue. In fact, the correlation between *ln Profit* and *ln Turnover* is 0.787, which is certainly much higher than the correlation between the set of independent variables and the dependent variable. Therefore, multicollinearity may indeed be an issue.