CHAPTER 6

DUMMY VARIABLE REGRESSION MODELS

QUESTIONS

6.1. (***a***) and (***b***) These are variables that cannot be quantified on a cardinal scale. They usually denote the possession or nonpossession of an attribute, such as nationality, religion, sex, color, etc.

(***c***) Regression models in which explanatory variables are qualitative are known as ANOVA models.

(***d***) Regression models in which one or more explanatory variables are quantitative, although others may be qualitative, are known as ANCOVA models.

(***e***) In a regression model with an intercept, if a qualitative variable has m categories, one must introduce only $(m - 1)$ dummy variables. If we introduce m dummies in such a model, we fall into the dummy variable trap, that is, we cannot estimate the parameters of such models because of perfect (multi)collinearity.

(***f***) They tell whether the average value of the dependent variable varies from group to group.

(***g***) If the rate of change of the mean value of the dependent variable varies between categories, the differential slope dummies will point that out.

6.2. (***a***) Quantitative (***b***) qualitative (***c***) quantitative

(***d***) qualitative (***e***) quantitative (***f***) qualitative, if expressed in broad categories, but quantitative if expressed as years of schooling

(***g***) qualitative (***h***) qualitative (***i***) qualitative

(***j***) qualitative.

6.3. (***a***) If there is an intercept term in the model, 11 dummies.

(***b***) If there is an intercept term in the model, 5 dummies.

6.4. (***a***) Here we will fall into the *dummy variable trap*, because the four columns of the dummy variables will add up to the first column (representing the intercept).

(***b***) This equation can be written as:

$$GNP_t = B_1 + (B_2 + B_4)M_t + (B_3 - B_4)M_{t-1} + u_t$$
$$= B_1 + A_2 M_t + A_3 M_{t-1} + u_t$$

where $A_2 = (B_2 + B_4)$ and $A_3 = (B_3 - B_4)$.

Although we can estimate B_1, A_2, and A_3, we cannot estimate B_2, B_3, and B_4 uniquely. The problem here is that the third explanatory variable in the original model, $(M_t - M_{t-1})$, is a linear combination of M_t and M_{t-1}, thereby leading to perfect collinearity.

6.5. (***a***) *False*. Letting D take the values of (0, 2) will halve both the estimated B_2 and its standard error, leaving the t ratio unchanged.

(***b***) *False*. Since the dummy variables do not violate any of the assumptions of OLS, the estimators obtained by OLS are unbiased in small as well as large samples.

6.6. (***a***) Each regression coefficient is expected to be positive.

(***b***) B_2 tells us by how much the average salary of a Harvard MBA differs from the base category, which is non-Harvard and non-Wharton MBAs.

(***c***) It probably suggests that the Harvard MBA has a premium over the Wharton MBA.

6.7. (***a***) The model given in the previous question assumes that the average starting salaries of Harvard and Wharton MBAs are different from that of the other MBAs, but the rate of change of salary with respect to years of service is the same for all graduates. On the other hand, the model given in this question assumes that the average starting salary as well as the progression of salary (i.e., the rate of change) over years of service is different among Harvard, Wharton, and other MBAs.

(***b***) B_4 and B_5 are *differential slopes.*

(***c***) Yes, otherwise, we will be committing the "omission of relevant variable" bias.

(***d***) This can be tested by the *F* test.

PROBLEMS

6.8. (***a***) The coefficient -0.1647 is the own-price elasticity, 0.5115 is the income elasticity, and 0.1483 is the cross-price elasticity.

(***b***) It is inelastic because, in absolute value, the coefficient is less than one.

(***c***) Since the cross-price elasticity is positive, coffee and tea are substitute products.

(***d***) and (***e***) The trend coefficient of -0.0089 suggests that over the sample period coffee consumption had been declining at the quarterly rate of 0.89 percent. Among other things, the side effects of caffeine may have something to do with the decline.

(***f***) 0.5115.

(***g***) The estimated *t* value of the income elasticity coefficient is 1.23, which is not statistically significant. Therefore, it does not make much sense to test the hypothesis that it is not different from one.

(***h***) The dummies here perhaps represent seasonal effects, if any.

(***i***) Each dummy coefficient tells by how much the average value of ln Q is different from that of the base quarter, which is the fourth quarter. The actual values of the intercepts in the various quarters are, respectively, 1.1828, 1.1219, 1.2692, and 1.2789. Taking the antilogs of these values, we obtain: 3.2635, 3.0707, 3.5580, and 3.5927 as the average pounds of coffee consumed per capita in the first, second, third, and the fourth quarter, holding the values of the logs of all explanatory variables zero.

Note: On the general interpretation of the dummy variables in a semi-log model, see Robert Halvorsen and Raymond Palmquist, "The Interpretation of Dummy Variables in Semilogarithmic Equations," *The American Economic Review*, vol. 70 (June 1980), no.3, pp. 474-475.

(*j*) The dummy coefficients D_1 and D_2 are individually statistically significant.

(*k*) That seems to be the case in quarters one and two. Among other things, coffee prices and weather may have something to do with the observed seasonal pattern in these two quarters.

(*l*) The benchmark is the fourth quarter. If we choose another quarter for the base, the numerical values of the dummy coefficients will change.

(*m*) The implicit assumption that is made is that the partial slope coefficients do not change among quarters.

(*n*) We can incorporate *differential slope* dummies as follows:

$$\begin{aligned}\ln Q = {} & B_1 + B_2 \ln P + B_3 \ln I + B_4 \ln P' + B_5 T \\ & + B_6 D_1 + B_7 D_2 + B_8 D_3 \\ & + B_9(D_1 \ln P) + B_{10}(D_2 \ln P) + B_{11}(D_3 \ln P) \\ & + B_{12}(D_1 \ln I) + B_{13}(D_2 \ln I) + B_{14}(D_3 \ln I) \\ & + B_{15}(D_1 \ln P') + B_{16}(D_2 \ln P') + B_{17}(D_3 \ln P') + u\end{aligned}$$

Note: The subscript "*t*" has been omitted to avoid cluttering the equation. The first two rows of the equation are the same as in the text. The *differential slope* dummies are in the last three rows.

(*o*) One could estimate the model given in (*n*). If there are other substitutes for coffee, they can be brought in the model.

6.9. **(*a*)** It is a way of finding out if there are economies or diseconomies of scale. In general, if at a given point, the first derivative (i.e., the slope) is negative but the second derivative is positive, it means the slope is *negative* and *increasing*, that is, the negative slope tends to be less steep as the value of the variable increases.

(*b*) The same reasoning as in (*a*), except that miles has a positive sign and miles squared has a negative sign. In general, if at a given point, the first derivative is positive but the second derivative is negative, it means that the value of the function is increasing at a decreasing rate. In the present case,

this is an indication of economies of scale, for the longer the distance in miles is, the lesser is the incremental fare.

(***c***) Population may be a proxy for traffic volume. The negative sign here indicates perhaps some type of economies of scale.

(***d***) Although negative, the coefficient is significant only for the "discount" category. This sign is rather puzzling.

(***e***) The negative sign makes economic sense in the sense that the higher the number of stopovers, the greater is the time spent traveling. Hence, the fare is lower to induce passengers to travel with several stopovers.

(***f***) It suggests that the average level of fare for Continental Airlines is lower than its competitors'.

(***g***) The critical Z value is 1.96 (5%, two-tailed) or 1.65 (5%, one tailed). If the computed Z value exceeds these critical values, the coefficient in question is statistically significant.

(***h***) Although this dummy coefficient is expected to be positive for all categories, it is not clear why it is significant only for the "discount" category.

(***i***) Yes, these observations can be pooled. In that case, introduce an additional dummy for the "coach" or "discount" fares.

(***j***) Overall, the results are a mixed bag. Although the R^2s are quite high for this sample size, and although several coefficients are statistically significant, some of the coefficients have dubious signs.

6.10. (***a***) Since the coefficient of the Dumsex dummy is statistically significant, Model 2 is preferable to Model 1.

(***b***) The error of omitting a relevant variable.

(***c***) *Ceteris paribus*, the average weight of males is greater than that of females.

(***d***) There is an additional variable, Dumht, in Model 3, which is statistically insignificant. As shown in Chapter 11, if an "unnecessary" variable is added to a model, the OLS estimators, while unbiased and consistent, are generally inefficient. This can be seen from Model 3. In Model 2 the Dumsex

variable was statistically significant, but is insignificant in Model 3 because of the apparently superfluous Dumht variable. Also, keep in mind the possibility of multicollinearity.

(*e*) Choose Model 2. Not only is the Dumsex variable statistically significant in this model, but the coefficient of the height variable is about the same in both Models 2 and 3. On the other hand, neither dummy variable is statistically significant in Model 3.

(*f*) We observe from the correlation matrix that the coefficient of correlation between Dumsex and Dumht is very high, almost unity. As we show in the chapter on multicollinearity, in cases of very high collinearity, OLS estimators, although unbiased, have relatively large standard errors. Also, the signs and magnitudes of the coefficients can change with slight alterations in the data or in the specification of the model.

6.11. **(*a*)** $\widehat{\text{Sales}}_t = 930.4118 + 58.6667\, D_{2t} + 57.6091\, D_{3t} + 1338.1091\, D_{4t}$

$t =$ (21.598) (0.963) (0.931) (21.629)

$R^2 = 0.9130$

(*b*) The average sales in the first quarter was about $930 million. In the second quarter it was higher by about $59 million, in the third quarter by about $58 million, and in the fourth quarter by about $1338 million, but only the fourth quarter dummy variable appears to be statistically significant. The second and third quarter dummy variables are not significant, indicating there isn't a significant difference between sales in the second and first (since it was not explicitly incorporated in the model) quarters or between the third and first quarters. The actual values of the intercepts in the various quarters can be obtained by adding the differential intercept dummies to the base quarter value. The individual intercept values are, respectively, (all in millions of dollars):

1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
930.412	989.078	988.021	2268.521

(***c***) It makes sense that hobby, toy, and game sales would be higher during certain parts of the year. Quarter 4 is likely to be the winter season, which contains several gift-giving holidays.

(d) To deseasonalize the data, subtract from each quarter's sales figure the dummy coefficient of that quarter. For instance, if you subtract from the sales figure for the fourth quarter of each year the number 1338.109, the resulting figure for that quarter will indicate the seasonally adjusted sales for that quarter. Thus, the seasonally adjusted figures for the fourth quarter of 1992, 1993, 1994, and 1995 are as follows:

4th Quarter 1992	**4th Quarter 1993**	**4th Quarter 1994**	**4th Quarter 1995**
458.558	527.891	712.891	822.224

6.12. (***a***) $\widehat{\text{Sales}}_t = 930.412\, D_{1t} + 989.078\, D_{2t} + 988.021\, D_{3t} + 2268.521\, D_{4t}$

$t =$ (21.5983) (22.9602) (22.251) (51.0885)

This model gives directly the intercept values for all the four quarters, whereas, as shown in problem 6.11, the intercept values for the second, third, and fourth quarters were obtained by adding the differential intercept dummy values to the intercept value of the base quarter. Of course, both procedures give identical results, as they should.

Note: The R^2 value of this model is not presented for the reasons explained in the text (Ch. 5).

(b) In this model, we have assigned a dummy coefficient for each quarter. But notice that, to avoid the dummy variable trap, we have omitted the intercept term from this model and have run a regression through the origin.

(***c***) The results are virtually identical between these two models. The first approach may be a nicer approach simply because the R-squared value is given explicitly and a few slope comparisons are made within the output.

6.13. In this case the model will be:

$$Accept_i = B_1 + B_2 D_{2i} + B_3 D_{3i} + B_4 Tuition_i + B_5 (D_{2i} Tuition_i) + B_6 (D_{3i} Tuition_i) + u_i$$

The *Minitab* regression results are as follows:

```
Regression Analysis: Acceptance Rate versus N, W, ...

The regression equation is
Acceptance Rate = 66.0 - 8.4 N + 3.7 W - 0.000683 Tuition -
     0.000005 Tuition*N - 0.000470 Tuition*W

Predictor        Coef    SE Coef       T      P
Constant       66.025      9.020    7.32  0.000
N               -8.38      11.99   -0.70  0.487
W                3.66      16.48    0.22  0.825
Tuition    -0.0006825  0.0002737   -2.49  0.015
Tuition*N  -0.0000055  0.0003507   -0.02  0.988
Tuition*W  -0.0004703  0.0004884   -0.96  0.339

S = 12.5435   R-Sq = 37.7%   R-Sq(adj) = 32.5%

Analysis of Variance

Source          DF       SS      MS     F      P
Regression       5   5624.1  1124.8  7.15  0.000
Residual Error  59   9283.1   157.3
Total           64  14907.2
```

Compared with Equation (6.17), these results suggest that there is no regional variation in the coefficient of *Tuition* since the p-values are quite high for the interaction variables. Hence, the results of Equation (6.17) seem acceptable.

6.14. This can be accomplished by adding as variables the product of X_i and D_{2i} and the product of X_i and D_{3i}.

6.15. Using *EViews*, and suppressing the intercept to avoid the problem of perfect multicollinearity , we obtain the following results:

Dependent Variable: FRIG Sample: 1978:1 1985:4				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DUM1MINE	1222.125	59.99041	20.37200	0.0000
DUM2	1467.500	59.99041	24.46224	0.0000
DUM3	1569.750	59.99041	26.16668	0.0000
DUM4	1160.000	59.99041	19.33642	0.0000
R-squared	0.531797			

Here the various dummies represent the average sale of refrigerators in each quarter.

6.16. **(*a*)** By interaction we mean when both effects (sex and race) are present simultaneously.

(*b*) B_2 = differential effect of being a male

B_3 = differential effect of being white

B_4 = differential effect of being a white male

(*c*) $E(Y) = (B_1 + B_2 + B_3 + B_4) + B_5 X_i$

given that $D_{2i} = D_{3i} = 1$. Thus, a white male's mean annual salary is higher by B_4 as compared to the mean salary of a male alone or a white alone.

6.17. We define the new Sex dummy variable as equal to 1 for female and –1 for male and name it SEX1FN1M, to distinguish it from the original dummy variable SEX already in Table 6-2. In SEX1FN1M , "1FN1M" stands for 1 for female and negative 1 for male. The *EViews* output is as follows:

Dependent Variable: FOODEXP Sample: 1 12				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2925.250	164.7874	17.75166	0.0000
SEX1FN1M	-251.5833	164.7874	-1.526714	0.1578
R-squared	0.189026			

With this dummy setup, the constant term represents the "average" intercept of the regression line from which the female and male intercepts differ by 251.5833, in the opposite direction. Thus, the intercept for males is (2,925.250 + 251.5833) = 3,176.8333, and the one for females is calculated as (2,925.250 – 251.5833) = 2,673.6667, which are the values obtained for model (6.1) and shown in Equation (6.4), save any minor rounding errors.

6.18. In this problem, we define the new Sex dummy variable as equal to 2 for female and 1 for male and name it SEX2F1M, to distinguish it from the original dummy variable SEX already in Table 6-2. In SEX2F1M, "2F1M" stands for 2 for female and 1 for male. The regression results are:

Dependent Variable: FOODEXP
Sample: 1 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3680.000	521.1036	7.061935	0.0000
SEX2F1M	-503.1667	329.5749	-1.526714	0.1578
R-squared	0.189026			

These results are precisely the same as in Equation (6.4), by noting that when $D_i = 2$ (female), the female intercept is 3,680.000 – 2(503.1667) = 2,673.6666 and the male intercept is 3,680.000 – 503.1667 = 3,176.8333.

6.19 **(*a*)** Based on the 19 observations, the *EViews* regression results are:

Dependent Variable: NDIV
Sample: 1997:1 2008:2

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-19.0132	26.2112	-0.7254	0.4721
ATPROFITS	0.6311	0.0311	20.3003	0.0000
R-squared	0.9035			

As these results show, there is a statistically significant positive relationship between the two variables, an unsurprising finding.

(***b***), (***c***),and (***d***) We can introduce three dummies to distinguish four quarters and can also interact them with the profits variable. This exercise yielded no satisfactory results, since both the dummies and interaction terms were completely insignificant, suggesting that perhaps there is no seasonality involved. This makes sense, for most corporations do not change their dividends from quarter to quarter. It seems that there is no reason to consider explicitly seasonality in the present case.

6.20. The reference category (i.e., the category with 0 value for all dummies) is unmarried white male. Therefore, the intercept for this category is 0.501. All other variables remain the same. The intercept for white unmarried female is (0.501 + 0.140) = 0.641. Since the coefficient of DF is not statistically significant at the 5% level, it seems that there is no difference between the two categories in their intercept values. Other variables remain the same.

6.21. You will have to expand the model by including the product of each dummy variable with the other explanatory variables (6 in all). Thus you will have to add (6 × 3) = 18 additional variables to the model. But do not forget the principle of parsimony.

6.22. (***a***) Since the *p* value of the dummy coefficient is about 14%, it seems that product-differentiation does not lead to a higher rate of return.

(***b***) From (*a*) it is obvious that there is no statistical difference in the rate of return for firms that product-differentiate and the firms that do not.

(***c***) Perhaps. If we had the original data, we could verify this. Product differentiation is the result of advertising and marketing strategies. For details, see any industrial organization textbook.

(***d***) To the equation given, add the product of *D* with each of the explanatory variables. Thus, there will be three additional variables in the model.

6.23. (***a***) Since both the differential intercept and slope coefficients are statistically significant, the Phillips curve has changed between the two time periods. The regression models for the two periods derived from this regression are:

$$\underline{1958–1969}: \quad \hat{Y}_t = (10.078 - 10.337) + (-17.549 + 38.137)\left(\frac{1}{X_t}\right)$$

$$= -0.259 + 20.588\left(\frac{1}{X_t}\right)$$

$$\underline{1970–1977}: \quad \hat{Y}_t = \quad 10.078 - 17.549\left(\frac{1}{X_t}\right)$$

What is striking about the latter period is that the slope coefficient is negative! This would imply a "positively" sloped Phillips Curve.

(*b*) The original Phillips curve may be dead but several attempts have been made to revive it. See any modern textbook on macroeconomics.

6.24. From Table 6.10 we observe that of the 40 observations, 6 observations have negative predicted values and 6 have predicted values in excess of 1. Hence, there are 12 incorrect predictions. Therefore,

$$Count\ R^2 = 28 / 40 = 0.7000.$$

The conventional R^2 value is 0.8047.

6.25. **(*a*)** Scatter plots will show that the three expenditure categories are linearly related to PCE.

(*b*) Since the data are seasonally adjusted, if you regress each expenditure category on PCE and include the dummy variables, the dummy coefficients are likely to be insignificant. This, in fact, turns out to be the case. But keep in mind that the method of seasonal adjustment used by the U.S. government is different from the dummy variable method.

Note: *EViews* provides seasonal adjustment options.

(*c*) By including the dummy variables unnecessarily, you will be committing the bias of including superfluous variables. As a result, the standard error of the PCE coefficient is likely to be overestimated, which will lower the *t* values.

6.26. ***(a)***, ***(b)***, ***(c)*** *Minitab* results:

```
Regression Analysis: % change versus shift, 1/X, shift*(1/X)

The regression equation is
% change = 8.52 - 6.38 shift - 14.0 1/X + 19.6 shift*(1/X)

Predictor       Coef  SE Coef      T      P
Constant      8.5165   0.9201   9.26  0.000
shift         -6.379    1.471  -4.34  0.000
1/X          -14.011    4.743  -2.95  0.005
shift*(1/X)   19.606    7.894   2.48  0.017

S = 1.11262   R-Sq = 65.7%   R-Sq(adj) = 63.0%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       3   92.371  30.790  24.87  0.000
Residual Error  39   48.279   1.238
Total           42  140.650
```

(d) It appears that all three independent variables are statistically significant because of their low p-values.

(e) The significant results indicate that there was an economic shift around 1982-1983 that resulted in an overall decrease in the percent change in the index of hourly earnings, but also an *increase* in the rate at which the hourly earnings changed as a function of the inverse of unemployment rate.

6.27. ***(a)*** No, it does not make sense to use the education variable as it is in the dataset. If it is left as it is (with 1 referring to a high school graduate, 2 referring to a college graduate, and 3 referring to someone with a graduate degree), the result would be forced to give the same increase in salary to a college graduate over a high school graduate as it would for the increase in salary to a graduate school graduate over someone with a college degree.

(b) The best way to address the issue above is to create dummy variables for the education levels, leaving one out to avoid the dummy variable trap. Results using *Minitab* are as follows:

```
Regression Analysis: Salary versus Experience, Management,
College, Grad

The regression equation is
Salary = 8036 + 546 Experience + 6884 Management + 3144
College + 2996 Grad

Predictor      Coef  SE Coef      T      P
```

```
Constant    8035.6    386.7  20.78  0.000
Experience  546.18    30.52  17.90  0.000
Management  6883.5    313.9  21.93  0.000
College     3144.0    362.0   8.69  0.000
Grad        2996.2    411.8   7.28  0.000

S = 1027.44   R-Sq = 95.7%   R-Sq(adj) = 95.3%


Analysis of Variance
Source           DF          SS         MS       F      P
Regression        4   957816858  239454214  226.84  0.000
Residual Error   41    43280719    1055627
Total            45  1001097577
```

Since the p-values are all around 0, it seems that all the variables are statistically significant.

(c) This would imply using an interaction effect between the Management and the Experience variables. The model results are:

```
Regression Analysis: Salary vs Experience, Management, ...

The regression equation is
Salary = 8256 + 525 Experience + 6461 Management + 3065
College + 2883 Grad  + 59.2 Mgt*Exp

Predictor     Coef  SE Coef      T      P
Constant    8256.3    459.6  17.96  0.000
Experience  525.16    38.59  13.61  0.000
Management  6460.8    568.1  11.37  0.000
College     3065.0    373.5   8.21  0.000
Grad        2883.5    431.6   6.68  0.000
Mgt*Exp      59.19    66.22   0.89  0.377

S = 1029.96   R-Sq = 95.8%   R-Sq(adj) = 95.2%

Analysis of Variance

Source           DF          SS         MS       F      P
Regression        5   958664481  191732896  180.74  0.000
Residual Error   40    42433096    1060827
Total            45  1001097577
```

These results indicate that there doesn't seem to be a significant difference in the *rate* of change in salary between managers and non-managers.

(d) The full model results, with the inclusion of interaction effects between the education levels and years of experience, is:

```
Regression Analysis: Salary vs Experience, Management, ...

The regression equation is
Salary = 7466 + 614 Experience + 6354 Management + 4190
College + 4132 Grad + 118 Mgt*Exp - 147 Col*Exp - 209
Grad*Exp

Predictor       Coef  SE Coef      T      P
Constant      7466.4    550.5  13.56  0.000
Experience    614.40    52.98  11.60  0.000
Management    6354.0    546.4  11.63  0.000
College       4190.4    659.1   6.36  0.000
Grad          4132.0    679.3   6.08  0.000
Mgt*Exp       118.30    69.35   1.71  0.096
Col*Exp      -147.31    69.40  -2.12  0.040
Grad*Exp     -208.63    95.70  -2.18  0.036

S = 981.510   R-Sq = 96.3%   R-Sq(adj) = 95.7%

Analysis of Variance

Source           DF          SS          MS       F      P
Regression        7   964489826   137784261  143.02  0.000
Residual Error   38    36607751      963362
Total            45  1001097577
```

Although the interaction between Management and Experience has become a bit more significant, it still would not be so at the 0.05 level. The interactions between the education levels and Experience are useful at that level, though, indicating there are different *rates* of salary increase between employees with different education levels.

6.28. ***(a)*** Regression results are following:

```
Regression Analysis: ln Wage versus AGE, FEMALE, ...

* EXPER is highly correlated with other X variables
* EXPER has been removed from the equation.

The regression equation is
ln Wage = 0.829 + 0.0128 AGE - 0.249 FEMALE - 0.134 NONWHITE
      + 0.180 UNION + 0.0871 EDUCATION

Predictor       Coef   SE Coef      T      P
Constant     0.82894   0.07761  10.68  0.000
AGE         0.012760  0.001172  10.89  0.000
```

```
FEMALE      -0.24915    0.02663  -9.36  0.000
NONWHITE    -0.13354    0.03718  -3.59  0.000
UNION        0.18020    0.03695   4.88  0.000
EDUCATION   0.087110   0.004733  18.40  0.000

S = 0.475237   R-Sq = 34.6%   R-Sq(adj) = 34.3%

Analysis of Variance
Source            DF       SS      MS       F      P
Regression         5  153.065  30.613  135.55  0.000
Residual Error  1283  289.766   0.226
Total           1288  442.831
```

Note: Minitab removed the variable *Experience* from the model as it was highly correlated to another one or more of the independent variables. This issue is known as multicollinearity and will be discussed further in Chapter 8. Looking at the correlation between *Experience* and *Age*, it is apparent that this is the source of the collinearity.

```
Correlations: EXPER, AGE

Pearson correlation of EXPER and AGE = 0.971
```

(b) An employee who is one year older than another, with the values of the other independent variables the same, the wage will be about 1.27% higher. Females, on average, make approximately 24.9% less than male counterparts with the same other values. Nonwhite employees tend to make about 13.3% less than their white counterparts, all else held equal, and union members make approximately 18% more than non-union members. Also, an extra year of education, all else being held the same, is worth approximately 8.7% more in wages.

(c) Based on the p-values, which are all approximately 0, all variables appear to be statistically significant.

(d) Yes, union workers tend to earn about 18% more than non-union workers.

(e) Yes, female workers tend to earn almost 25% less than equivalent male counterparts.

(f) Regression results including an interaction term between Female and Nonwhite are:

```
The regression equation is
ln Wage = 0.838 + 0.0127 AGE - 0.265 FEMALE - 0.190 NONWHITE +
0.182 UNION + 0.0872 EDUCATION + 0.105 Fem*Nonwhite

Predictor         Coef    SE Coef      T      P
```

```
Constant       0.83840   0.07787  10.77  0.000
AGE           0.012688  0.001172  10.82  0.000
FEMALE        -0.26493   0.02886  -9.18  0.000
NONWHITE      -0.19014   0.05463  -3.48  0.001
UNION          0.18202   0.03696   4.92  0.000
EDUCATION     0.087164  0.004732  18.42  0.000
Fem*Nonwhite   0.10452   0.07394   1.41  0.158

S = 0.475053   R-Sq = 34.7%   R-Sq(adj) = 34.4%

Analysis of Variance
Source            DF       SS      MS       F      P
Regression         6  153.516  25.586  113.38  0.000
Residual Error  1282  289.315   0.226
Total           1288  442.831
```

Since the interaction term has a p-value of 0.158, there does not appear to be a statistically significant difference in the *rate* of wage increase between white and nonwhite females. There still is a constant shift for the nonwhite group, however.

(f) There is not a statistically significant effect of an interaction term between Female and Union, so (similar to the question above), the *rate* of wage increase is not significantly different between union and non-union female employees. There is still, however, the dummy variable shift for union workers.

(h) This exercise is left to the reader; there are several possible models.