HW 5 Answers

9.28

(a) Model (9.5.4) is a linear model, whereas the present one is a log-linear model. Therefore, the slope coefficients of the regressor in this model are to be interpreted as semi-elasticities. Qualitatively, both models give similar results. Since the regressand in the two models are different, we cannot compare the two R^2's directly

(b) If we take the antilog of the dummy coefficient of 3.6772, what we obtain is the median savings in the period 1970-1981, holding all other factors constant. Now antilog(3.6772)=39.5355. Thus, if income were zero, the median savings in 1970-1981 would be about 40 billion dollars. Now if we take the antilog of (3.6772+1.3971) =150.8602, this would be median savings in the period 1982-1995, holding income constant.

(c) Regressing log of Y on X, the estimated error variances in the two periods are $\hat{\sigma}^2 = 0.0122$ (df=10) and $\hat{\sigma}^2 = 0.0182$ (df=12) Under the null hypothesis that the variances of the two populations are the same, we form

$$F = \frac{0.0182}{0.0122} = 1.4918$$

For 12 and 10 df in the numerator and denominator, respectively, this value is not significant. Hence, we can conclude that the two error variances are the same.

9.29

(a)

Source	SS	df	MS
Model	39.3767288	8	4.9220911
Residual	39.3911662	105	.375153964
Total	78.767895	113	.697061018

Number of obs	= 114
F(8, 105)	= 13.12
Prob > F	= 0.0000
R-squared	= 0.4999
Adj R-squared	= 0.4618
Root MSE	= .6125

lnwi	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0300112	.0047349	6.34	0.000	.0206228	.0393996
dsex	-.5878966	.1452119	-4.05	0.000	-.8758249	-.2999682
de2	.0381515	.1923389	0.20	0.843	-.3432211	.4195241
de3	.3593484	.2114815	1.70	0.092	-.0599803	.7786772
de4	1.757512	.621087	2.83	0.006	.5260109	2.989012
de2dpt	.399618	.3276193	1.22	0.225	-.2499905	1.049227
de3dpt	.438133	.3108017	1.41	0.162	-.1781292	1.054395
de4dpt	-1.199202	.6778215	-1.77	0.080	-2.543197	.1447928
_cons	3.688426	.1757925	20.98	0.000	3.339862	4.03699

Based on the p-values of the new terms, there doesn't really appear to be a significant interaction between the education level and job type (permanent or temporary). The last variable, combining the highest education level with job type, however, is marginally significant with a *p value* of 0.080.

(b) To assess the difference between workers with an education level up to primary and those without a primary education, we will look at both the dummy variable DE2 and the interaction term DE2_DPT. Neither are significant (based on the high *p values*). For workers with a secondary education, look at DE3 and DE3_DPT *p values*, neither of which are significant at the standard 5% level. For the difference between those with an education level beyond secondary and those without a primary level of education, however, there does seem to be a significant difference in the intercept term (the *p value* for DE4 is 0.0056). This is not surprising considering it represents the most extreme disparity in education levels in this dataset.

(c) Having deleted the dummy variables, but retaining the interaction terms, the regression results are now the following:

Source	SS	df	MS
Model	35.3635071	5	7.07270141
Residual	43.4043879	108	.401892481
Total	78.767895	113	.697061018

Number of obs =	114
F(5, 108) =	17.60
Prob > F =	0.0000
R-squared =	0.4490
Adj R-squared =	0.4234
Root MSE =	.63395

lnwi	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0299581	.0047524	6.30	0.000	.0205381	.0393781
dsex	-.632527	.1487523	-4.25	0.000	-.92738	-.3376741
de2dpt	.3683329	.2925221	1.26	0.211	-.2114966	.9481623
de3dpt	.7347649	.2514172	2.92	0.004	.2364123	1.233117
de4dpt	.515768	.2975022	1.73	0.086	-.073933	1.105469
_cons	3.759603	.1666903	22.55	0.000	3.429194	4.090012

Now we see a slightly different result: the interaction term between the secondary education level and job type (DE3_DPT) is very statistically significant. That between education level past secondary and job type (DE4_DPT) is marginally significant, but that between primary level and job type is not. This is not surprising after having deleted the dummy variables; the differences between the education levels is now being picked up in the interaction terms instead.

10.2

(a) No. Variable X_{3i} is an exact linear combination of X_{2i}, because $X_{3i} = 2X_{2i} - 1$

(b) Rewriting the equation yields,

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3(2X_{2i} - 1) + u_i \\ &= (\beta_1 - \beta_3) + (\beta_2 + 2\beta_3)X_{2i} + u_i \\ &= \alpha_1 + \alpha_2 X_{2i} + u_i \\ &\textit{where}\, \alpha_1 = (\beta_1 - \beta_3)\, \textit{and}\, \alpha_2 = (\beta_2 + 2\beta_2) \end{aligned}$$

Therefore, we can estimate $\alpha_1$and $\alpha_2$uniquely, but not the original betas because we have two equations to solve the three unknowns.

10.3

(a) Although the numerical values of the intercept and the slope coefficients of PGNP and FLR have changed, their signs have not. Also, these variables are still statistically significant. These changes are due to the addition of the TFR variable, suggesting that there may be some collinearity among the regressors.

(b) Since the t value of the TFR coefficient is very significant, it seems TFR belongs in the model. The positive sign of this coefficient also makes sense in that the larger the number of children born to a woman, the greater the chances of increased child mortality

(c) Despite possible collinearity, the individual coefficients are still statistically significant.

10.12State with reason whether the following statements are true, false, or uncertain:

a. Despite perfect multicollinearity, OLS estimators are BLUE.
   False. If exact linear relationship(s) exist among variables, we cannot even estimate the coefficients or their standard errors.

b. In cases of high multicollinearity, it is not possible to assess the individual significance of one or more partial regression coefficients.
   False. One may be able to obtain one or more significant *t* values.

c. High pair-wise correlations do not suggest that there is high multicollinearity.
   Uncertain. If a model has only two regressors, high pairwise correlation coefficients may suggest multicollinearity. If one or more regressors enter non-linearly, the pairwise correlations may give misleading answers.

d. You will not obtain a high R-squared value in a multiple regression if all the partial slope coefficients are individually statistically insignificant on the basis of the usual t-test.
   False. One usually obtains high *R*2's in models with highly correlated regressors.

e. Ceteris Paribus, the higher the VIF is, the larger the variances of OLS estimators.
   False. The variance of an OLS estimator is given by the following formula:

$$\mathrm{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right)$$

A high R_j^2 can be counterbalanced by a low σ^2 or high $\sum x_j^2$