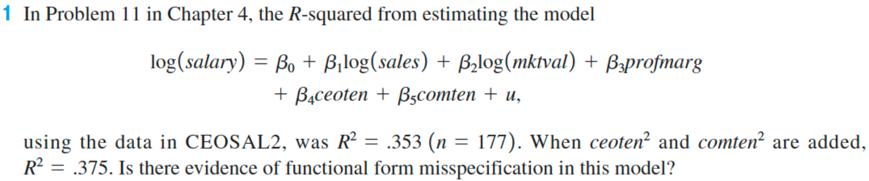
9.7 Problem Set 4

Question 1, page 324, 9th Edition

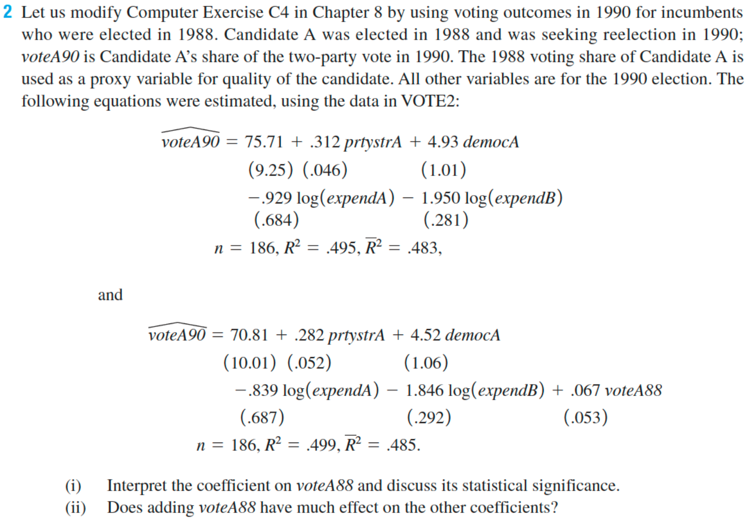


1 In Problem 11 in Chapter 4, the *R*-squared from estimating the model

$$\log(salary) = \beta_0 + \beta_1\log(sales) + \beta_2\log(mktval) + \beta_3 profmarg + \beta_4 ceoten + \beta_5 comten + u,$$

using the data in CEOSAL2, was $R^2 = .353$ ($n = 177$). When $ceoten^2$ and $comten^2$ are added, $R^2 = .375$. Is there evidence of functional form misspecification in this model?

Question 2, page 324, 9th Edition



2 Let us modify Computer Exercise C4 in Chapter 8 by using voting outcomes in 1990 for incumbents who were elected in 1988. Candidate A was elected in 1988 and was seeking reelection in 1990; *voteA90* is Candidate A's share of the two-party vote in 1990. The 1988 voting share of Candidate A is used as a proxy variable for quality of the candidate. All other variables are for the 1990 election. The following equations were estimated, using the data in VOTE2:

$$\begin{aligned}\widehat{voteA90} = {} & \underset{(9.25)}{75.71} + \underset{(.046)}{.312}\, prtystrA + \underset{(1.01)}{4.93}\, democA \\ & - \underset{(.684)}{.929} \log(expendA) - \underset{(.281)}{1.950} \log(expendB) \\ & n = 186,\ R^2 = .495,\ \bar{R}^2 = .483,\end{aligned}$$

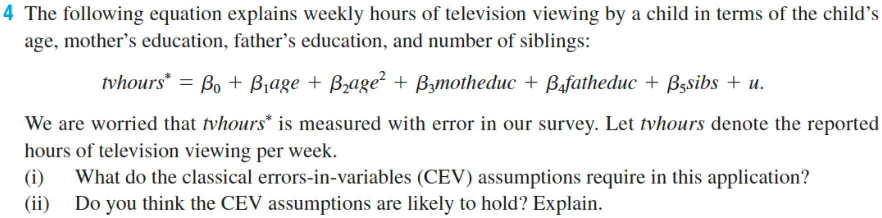
and

$$\begin{aligned}\widehat{voteA90} = {} & \underset{(10.01)}{70.81} + \underset{(.052)}{.282}\, prtystrA + \underset{(1.06)}{4.52}\, democA \\ & - \underset{(.687)}{.839} \log(expendA) - \underset{(.292)}{1.846} \log(expendB) + \underset{(.053)}{.067}\, voteA88 \\ & n = 186,\ R^2 = .499,\ \bar{R}^2 = .485.\end{aligned}$$

(i) Interpret the coefficient on *voteA88* and discuss its statistical significance.

(ii) Does adding *voteA88* have much effect on the other coefficients?

Question 4, page 325, 9th Edition



4 The following equation explains weekly hours of television viewing by a child in terms of the child's age, mother's education, father's education, and number of siblings:

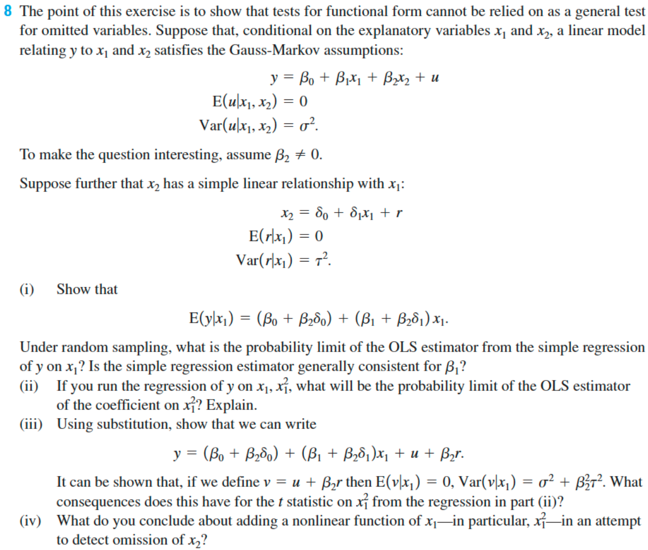
$$tvhours^{*} = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 motheduc + \beta_4 fatheduc + \beta_5 sibs + u.$$

We are worried that *tvhours** is measured with error in our survey. Let *tvhours* denote the reported hours of television viewing per week.

(i) What do the classical errors-in-variables (CEV) assumptions require in this application?

(ii) Do you think the CEV assumptions are likely to hold? Explain.

Question 8, page 326, 9th Edition



8 The point of this exercise is to show that tests for functional form cannot be relied on as a general test for omitted variables. Suppose that, conditional on the explanatory variables x_1 and x_2, a linear model relating y to x_1 and x_2 satisfies the Gauss-Markov assumptions:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ \mathrm{E}(u|x_1, x_2) &= 0 \\ \mathrm{Var}(u|x_1, x_2) &= \sigma^2. \end{aligned}$$

To make the question interesting, assume $\beta_2 \neq 0$.

Suppose further that x_2 has a simple linear relationship with x_1:

$$\begin{aligned} x_2 &= \delta_0 + \delta_1 x_1 + r \\ \mathrm{E}(r|x_1) &= 0 \\ \mathrm{Var}(r|x_1) &= \tau^2. \end{aligned}$$

(i) Show that

$$\mathrm{E}(y|x_1) = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1.$$

Under random sampling, what is the probability limit of the OLS estimator from the simple regression of y on x_1? Is the simple regression estimator generally consistent for β_1?

(ii) If you run the regression of y on x_1, x_1^2, what will be the probability limit of the OLS estimator of the coefficient on x_1^2? Explain.

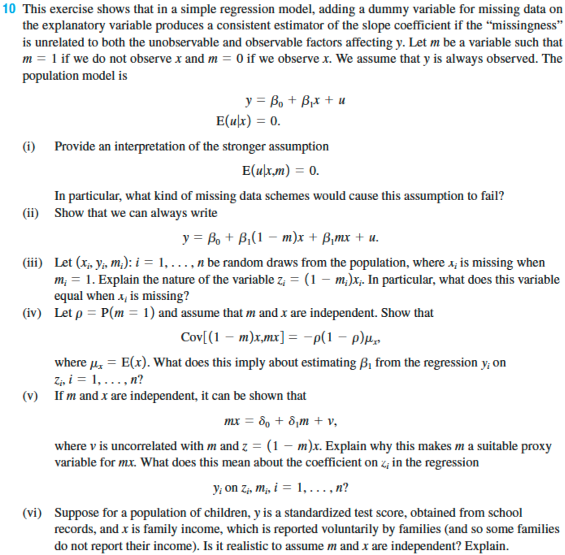
(iii) Using substitution, show that we can write

$$y = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + u + \beta_2 r.$$

It can be shown that, if we define $v = u + \beta_2 r$ then $\mathrm{E}(v|x_1) = 0$, $\mathrm{Var}(v|x_1) = \sigma^2 + \beta_2^2\tau^2$. What consequences does this have for the t statistic on x_1^2 from the regression in part (ii)?

(iv) What do you conclude about adding a nonlinear function of x_1—in particular, x_1^2—in an attempt to detect omission of x_2?

Question 10, page 327, 9th Edition



10 This exercise shows that in a simple regression model, adding a dummy variable for missing data on the explanatory variable produces a consistent estimator of the slope coefficient if the "missingness" is unrelated to both the unobservable and observable factors affecting y. Let m be a variable such that $m = 1$ if we do not observe x and $m = 0$ if we observe x. We assume that y is always observed. The population model is

$$y = \beta_0 + \beta_1 x + u$$
$$\mathrm{E}(u|x) = 0.$$

(i) Provide an interpretation of the stronger assumption

$$\mathrm{E}(u|x,m) = 0.$$

In particular, what kind of missing data schemes would cause this assumption to fail?

(ii) Show that we can always write

$$y = \beta_0 + \beta_1(1 - m)x + \beta_1 mx + u.$$

(iii) Let (x_i, y_i, m_i): $i = 1, \ldots, n$ be random draws from the population, where x_i is missing when $m_i = 1$. Explain the nature of the variable $z_i = (1 - m_i)x_i$. In particular, what does this variable equal when x_i is missing?

(iv) Let $\rho = \mathrm{P}(m = 1)$ and assume that m and x are independent. Show that

$$\mathrm{Cov}[(1 - m)x, mx] = -\rho(1 - \rho)\mu_x,$$

where $\mu_x = \mathrm{E}(x)$. What does this imply about estimating β_1 from the regression y_i on z_i, $i = 1, \ldots, n$?

(v) If m and x are independent, it can be shown that

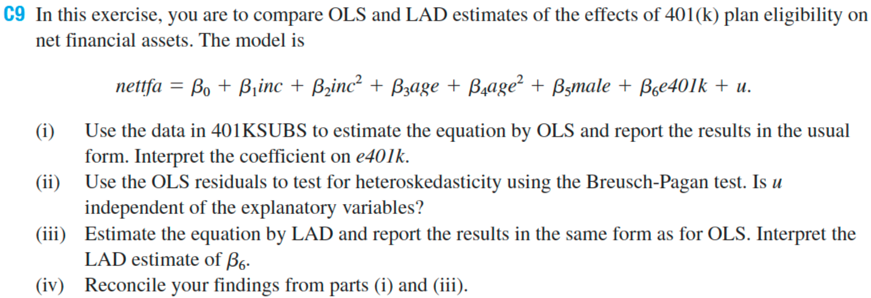
$$mx = \delta_0 + \delta_1 m + v,$$

where v is uncorrelated with m and $z = (1 - m)x$. Explain why this makes m a suitable proxy variable for mx. What does this mean about the coefficient on z_i in the regression

$$y_i \text{ on } z_i, m_i, i = 1, \ldots, n?$$

(vi) Suppose for a population of children, y is a standardized test score, obtained from school records, and x is family income, which is reported voluntarily by families (and so some families do not report their income). Is it realistic to assume m and x are independent? Explain.

Question C9, page 329, 9th Edition



C9 In this exercise, you are to compare OLS and LAD estimates of the effects of 401(k) plan eligibility on net financial assets. The model is

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 male + \beta_6 e401k + u.$$

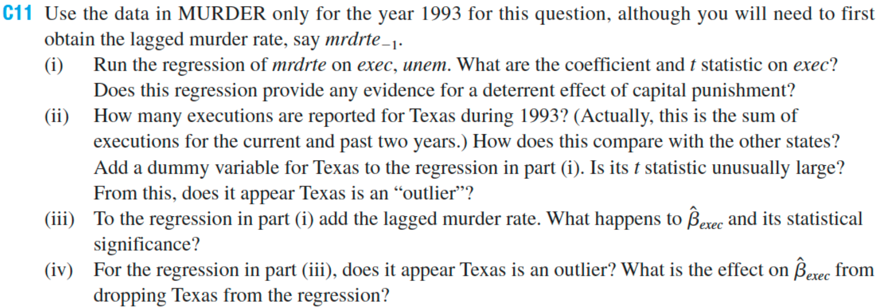
(i) Use the data in 401KSUBS to estimate the equation by OLS and report the results in the usual form. Interpret the coefficient on *e401k*.

(ii) Use the OLS residuals to test for heteroskedasticity using the Breusch-Pagan test. Is *u* independent of the explanatory variables?

(iii) Estimate the equation by LAD and report the results in the same form as for OLS. Interpret the LAD estimate of β_6.

(iv) Reconcile your findings from parts (i) and (iii).

Question C11, page 330, 9th Edition



C11 Use the data in MURDER only for the year 1993 for this question, although you will need to first obtain the lagged murder rate, say $mrdrte_{-1}$.

(i) Run the regression of *mrdrte* on *exec*, *unem*. What are the coefficient and *t* statistic on *exec*? Does this regression provide any evidence for a deterrent effect of capital punishment?

(ii) How many executions are reported for Texas during 1993? (Actually, this is the sum of executions for the current and past two years.) How does this compare with the other states? Add a dummy variable for Texas to the regression in part (i). Is its *t* statistic unusually large? From this, does it appear Texas is an "outlier"?

(iii) To the regression in part (i) add the lagged murder rate. What happens to $\hat{\beta}_{exec}$ and its statistical significance?

(iv) For the regression in part (iii), does it appear Texas is an outlier? What is the effect on $\hat{\beta}_{exec}$ from dropping Texas from the regression?