CHAPTER 9

SOLUTIONS TO PROBLEMS

9.1 There is functional form misspecification if $\beta_6 \neq 0$ or $\beta_7 \neq 0$, where these are the population parameters on $ceoten^2$ and $comten^2$, respectively. Therefore, we test the joint significance of these variables using the *R*-squared form of the *F* test: $F = [(.375 - .353)/(1 - .375)][(177 - 8)/2] \approx 2.97$. With 2 and ∞ *df*, the 10% critical value is 2.30, while the 5% critical value is 3.00. Thus, the *p*-value is slightly above .05, which is reasonable evidence of functional form misspecification. (Of course, whether this has a practical impact on the estimated partial effects for various levels of the explanatory variables is a different matter.)

9.3 (i) Eligibility for the federally funded school lunch program is very tightly linked to being economically disadvantaged. Therefore, the percentage of students eligible for the lunch program is very similar to the percentage of students living in poverty.

(ii) We can use our usual reasoning on omitting important variables from a regression equation. The variables log(*expend*) and *lnchprg* are negatively correlated: school districts with poorer children spend, on average, less on schools. Further, $\beta_3 < 0$. From Table 3.2, omitting *lnchprg* (the proxy for *poverty*) from the regression produces an upward biased estimator of β_1 [ignoring the presence of log(*enroll*) in the model]. So when we control for the poverty rate, the effect of spending falls.

(iii) Once we control for *lnchprg*, the coefficient on log(*enroll*) becomes negative and has a *t* of about –2.17, which is significant at the 5% level against a two-sided alternative. The coefficient implies that $\Delta\widehat{math10} \approx -(1.26/100)(\%\Delta enroll) = -.0126(\%\Delta enroll)$. Therefore, a 10% increase in enrollment leads to a drop in *math10* of .126 percentage points.

(iv) Both *math10* and *lnchprg* are percentages. Therefore, a ten percentage point increase in *lnchprg* leads to about a 3.23 percentage point fall in *math10*, a sizeable effect.

(v) In column (1), we explain very little of the variation in pass rates on the MEAP math test: less than 3%. In column (2), we explain almost 19% (which still leaves much variation unexplained). Clearly most of the variation in *math10* is explained by variation in *lnchprg*. This is a common finding in studies of school performance; family income (or related factors, such as living in poverty) are much more important in explaining student performance than are spending per student or other school characteristics.

9.5 The sample selection in this case is arguably endogenous. Because prospective students may look at campus crime as one factor in deciding where to attend college, colleges with high crime rates have an incentive not to report crime statistics. If this is the case, then the chance of appearing in the sample is negatively related to *u* in the crime equation. (For a given school size, higher *u* means more crime, and therefore a smaller probability that the school reports its crime figures.)

9.7 (i) Following the hint, we compute $\text{Cov}(w, y)$ and $\text{Var}(w)$, where $y = \beta_0 + \beta_1 x^* + u$ and $w = (z_1 + ... + z_m)/m$. First, because $z_h = x^* + e_h$, it follows that $w = x^* + \bar{e}$, where $\bar{e}$ is the average of the *m* measures (in the population). Now, by assumption, x^* is uncorrelated with each e_h, and the e_h are pairwise uncorrelated. Therefore,

$$\text{Var}(w) = \text{Var}(x^*) + \text{Var}(\bar{e}) = \sigma_{x^*}^2 + \sigma_e^2/m,$$

where we use $\text{Var}(\bar{e}) = \sigma_e^2/m$. Next,

$$\text{Cov}(w, y) = \text{Cov}(x^* + \bar{e}, \beta_0 + \beta_1 x^* + u) = \beta_1 \text{Cov}(x^*, x^*) = \beta_1 \text{Var}(x^*),$$

where we use the assumption that e_h is uncorrelated with *u* for all *h* and x^* is uncorrelated with *u*. Combining the two pieces gives

$$\frac{\text{Cov}(w, y)}{\text{Var}(w)} = \beta_1 \left\{ \frac{\sigma_{x^*}^2}{[\sigma_{x^*}^2 + (\sigma_e^2/m)]} \right\},$$

which is what we wanted to show.

(ii) Because $\sigma_e^2/m < \sigma_e^2$ for all $m > 1$, $\sigma_{x^*}^2 + (\sigma_e^2/m) < \sigma_{x^*}^2 + \sigma_e^2$ for all $m > 1$. Therefore,

$$1 > \frac{\sigma_{x^*}^2}{[\sigma_{x^*}^2 + (\sigma_e^2/m)]} > \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2},$$

which means the term multiplying β_1 is closer to one when *m* is larger. We have shown that the bias in $\bar{\beta}_1$ is smaller as *m* increases. As *m* grows, the bias disappears completely. Intuitively, this makes sense: the average of several mismeasured variables has less measurement error than a single mismeasured variable. As we average more and more such variables, the attenuation bias can become very small.

9.9 (i) We can use calculus to find the partial derivative of $\text{E}(y \mid \mathbf{x})$ with respect to any element x_j. Using the chain rule and the properties of the exponential function gives

$$\frac{\partial \text{E}(y \mid \mathbf{x})}{\partial x_j} = \left[\beta_j + \frac{1}{2} \cdot \frac{\partial h(\mathbf{x})}{\partial x_j} \right] \exp[\beta_0 + \mathbf{x}\beta + h(\mathbf{x})/2].$$

As the exponential function is strictly positive, $\exp[\beta_0 + \mathbf{x}\beta + h(\mathbf{x})/2] > 0$. Thus, the sign of the partial effect is the same as the sign of

$$\beta_j + \frac{\partial h(\mathbf{x})}{\partial x_j},$$

which can be either positive or negative, irrespective of the sign of β_j.

(ii) The partial effect of x_j on $\text{Med}(y|\mathbf{x})$ is $\beta_j \exp(\beta_0 + \mathbf{x}\beta)$, which as the same sign as β_j. In particular, if $\beta_j < 0$, then increasing x_j decreases $\text{Med}(y|\mathbf{x})$. From part (i), we know that the partial effect of x_j on $\text{E}(y|\mathbf{x})$ has the same sign as $\beta_j + \delta_j / 2$ because $\partial h(\mathbf{x}) / \partial x_j = \delta_j$. If $\delta_j > -2\beta_j$ then the partial effect on $\text{E}(y|\mathbf{x})$ is positive.

(iii) If $h(\mathbf{x}) = \sigma^2$, then

$$\text{E}(y|\mathbf{x}) = \exp(\beta_0 + \mathbf{x}\beta + \sigma^2 / 2),$$

and so our prediction of *y* based on the conditional mean, for a given vector $\mathbf{x}$, is

$$\hat{\text{E}}(y|\mathbf{x}) = \exp(\hat{\beta}_0 + \mathbf{x}\hat{\beta} + \hat{\sigma}^2 / 2) = \exp(\hat{\sigma}^2 / 2)\exp(\hat{\beta}_0 + \mathbf{x}\hat{\beta}),$$

where the estimates are from the OLS regression of $\log(y_i)$ on a constant and x_{i1}, …, x_{ik}, including $\hat{\sigma}^2$, the usual unbiased estimator of σ^2.

The prediction based on $\text{Med}(y|\mathbf{x})$ is

$$\widehat{\text{Med}}(y|\mathbf{x}) = \exp(\hat{\beta}_0 + \mathbf{x}\hat{\beta}).$$

The prediction based on the mean differs by the multiplicative factor $\exp(\hat{\sigma}^2 / 2)$, which is always greater than unity because $\hat{\sigma}^2 > 0$. So the prediction based on the mean is always larger than that based on the median.

SOLUTIONS TO COMPUTER EXERCISES

C9.1 (i) To obtain the RESET *F* statistic, we estimate the model in Computer Exercise 7.5 and obtain the fitted values, say $\widehat{lsalary}_i$. To use the version of RESET in (9.3), we add $(\widehat{lsalary}_i)^2$ and $(\widehat{lsalary}_i)^3$ and obtain the *F* test for joint significance of these variables. With 2 and 203 *df*, the *F* statistic is about 1.33 and *p*-value $\approx$ .27, which means that there is not much concern about functional form misspecification.

(ii) Interestingly, the heteroskedasticity-robust *F*-type statistic is about 2.24 with *p*-value $\approx$ .11, so there is stronger evidence of some functional form misspecification with the robust test. But it is probably not strong enough to worry about.

C9.3 (i) If the grants were awarded to firms based on firm or worker characteristics, *grant* could easily be correlated with such factors that affect productivity. In the simple regression model, these are contained in *u*.

(ii) The simple regression estimates using the 1988 data are

$$\widehat{\log(scrap)} = .409 + .057\, grant$$
$$(.241) \quad (.406)$$
$$n = 54, \quad R^2 = .0004.$$

The coefficient on *grant* is actually positive, but not statistically different from zero.

(iii) When we add $\log(scrap_{87})$ to the equation, we obtain

$$\widehat{\log(scrap_{88})} = .021 - .254\, grant_{88} + .831 \log(scrap_{87})$$
$$(.089) \quad (.147) \quad (.044)$$
$$n = 54, \quad R^2 = .873,$$

where the year subscripts are for clarity. The *t* statistic for H_0: $\beta_{grant} = 0$ is $-.254/.147 \approx -1.73$. We use the 5% critical value for 40 *df* in Table G.2: -1.68. Because $t = -1.73 < -1.68$, we reject H_0 in favor of H_1: $\beta_{grant} < 0$ at the 5% level.

(iv) The *t* statistic is $(.831 - 1)/.044 \approx -3.84$, which is a strong rejection of H_0.

(v) With the heteroskedasticity-robust standard error, the *t* statistic for $grant_{88}$ is $-.254/.142 \approx -1.79$, so the coefficient is even more significantly less than zero when we use the heteroskedasticity-robust standard error. The *t* statistic for H_0: $\beta_{\log(scrap_{87})} = 1$ is $(.831 - 1)/.071 \approx -2.38$, which is notably smaller than before, but it is still pretty significant.

C9.5 With *sales* defined to be in billions of dollars, we obtain the following estimated equation using all companies in the sample:

$$\widehat{rdintens} = 2.06 + .317\, sales - .0074\, sales^2 + .053\, profmarg$$
$$(0.63) \quad (.139) \quad (.0037) \quad (.044)$$
$$n = 32, \quad R^2 = .191, \quad \bar{R}^2 = .104.$$

When we drop the largest company (with sales of roughly $39.7 billion), we obtain

$$\widehat{rdintens} = 1.98 + .361\, sales - .0103\, sales^2 + .055\, profmarg$$
$$(0.72) \quad (.239) \quad (.0131) \quad (.046)$$
$$n = 31, \quad R^2 = .191, \quad \bar{R}^2 = .101.$$

When the largest company is left in the sample, the quadratic term is not statistically significant, even though the coefficient on the quadratic is less in absolute value than when we drop the largest firm. What is happening is that by leaving in the large sales figure, we greatly increase the variation in both *sales* and $sales^2$; as we know, this reduces the variances of the OLS

estimators (see Section 3.4). The *t* statistic on $sales^2$ in the first regression is about –2, which makes it almost significant at the 5% level against a two-sided alternative. If we look at Figure 9.1, it is not surprising that a quadratic is significant when the large firm is included in the regression; *rdintens* is relatively small for this firm even though its sales are very large compared with the other firms. Without the largest firm, a linear relationship between *rdintens* and *sales* seems to suffice.

(ii) With *sales* defined to be in billions of dollars, we obtain the following LAD estimated equation using all companies in the sample:

$$\widehat{rdintens} = \underset{(0.719)}{1.402} + \underset{(.159)}{.263\, sales} - \underset{(.0042)}{.0060\, sales^2} + \underset{(.051)}{.114\, profmarg}$$

$$n = 32, \quad R^2 = .098.$$

When we drop the largest company (with sales of roughly \$39.7 billion), we obtain

$$\widehat{rdintens} = \underset{(0.802)}{2.608} - \underset{(.267)}{0.2223\, sales} + \underset{(.0146)}{.0168\, sales^2} + \underset{(.0512)}{.0762 profmarg}$$

$$n = 31, \quad R^2 = .101.$$

When the largest company is left in the sample, the quadratic term is not statistically significant, even though the coefficient on the quadratic is less in absolute value than when we drop the largest firm. Here also, without the largest firm, a linear relationship between *rdintens* and *sales* seems to suffice.

(iii) From part (i) and (ii), we can say that OLS or LAD is more resilient to outliers.

C9.7 (i) 205 observations out of the 1,989 records in the sample have *obrate* > 40. (Data are missing for some variables, so not all of the 1,989 observations are used in the regressions.)

(ii) When observations with *obrat* > 40 are excluded from the regression in part (iii) of Computer exercise C7.8, we are left with 1,784 observations. The coefficient on *white* is about .129 (se ≈ .020). To three decimal places, these are the same estimates we got when using the entire sample (see Computer Exercise C7.8). Perhaps this is not very surprising since we only lost 205 out of 1,989 observations. However, regression results can be very sensitive when we drop over 10% of the observations, as we have here.

(iii) The estimates from part (ii) show that $\hat{\beta}_{white}$ does not seem very sensitive to the sample used, although we have tried only one way of reducing the sample.

C9.9 (i) The equation estimated by OLS is

$$\widehat{nettfa} = 21.198 - .270\,inc + .0102\,inc^2 - 1.940\,age + .0346\,age^2$$
$$(9.992)\quad (.075)\quad (.0006)\quad (.483)\quad (.0055)$$

$$+ 3.369\,male + 9.713\,e401k$$
$$(1.486)\quad (1.277)$$

$n = 9{,}275,\ R^2 = .202.$

The coefficient on *e401k* means that, holding other things in the equation fixed, the average level of net financial assets is about \$9,713 higher for a family eligible for a 401(k) than for a family not eligible.

(ii) The OLS regression of $\hat{u}_i^2$ on inc_i, inc_i^2, age_i, age_i^2, $male_i$, and $e401k_i$ gives $R^2_{\hat{u}^2} = .0374$, which translates into $F = 59.97$. The associated *p*-value, with 6 and 9,268 *df*, is essentially zero. Consequently, there is strong evidence of heteroskedasticity, which means that *u* and the explanatory variables cannot be independent [even though $E(u|x_1,x_2,\ldots,x_k) = 0$ is possible].

(iii) The equation estimated by LAD is

$$\widehat{nettfa} = 12.491 - .262\,inc + .00709\,inc^2 - .723\,age + .0111\,age^2$$
$$(1.382)\quad (.010)\quad (.00008)\quad (.067)\quad (.0008)$$

$$+ 1.018\,male + 3.737\,e401k$$
$$(.205)\quad (.177)$$

$n = 9{,}275$, Psuedo $R^2 = .109$.

Now, the coefficient on *e401k* means that, at given income, age, and gender, the median difference in net financial assets between families with and without 401(k) eligibility is about \$3,737.

(iv) The findings from parts (i) and (iii) are not in conflict. We are finding that 401(k) eligibility has a larger effect on mean wealth than on median wealth. Finding different mean and median effects for a variable such as *nettfa*, which has a highly skewed distribution, is not surprising. Apparently, 401(k) eligibility has some large effects at the upper end of the wealth distribution, and these are reflected in the mean. The median is much less sensitive to effects at the upper end of the distribution.

C9.11 (i) The regression gives $\hat{\beta}_{exec} = .085$ with $t = .30$. The positive coefficient means that there is no deterrent effect, and the coefficient is not statistically different from zero.

(ii) Texas had 34 executions over the period, which is more than three times the next highest state (Virginia with 11). When a dummy variable is added for Texas, its *t* statistic is –.32, which

is not unusually large. (The coefficient is large in magnitude, –8.31, but the studentized residual is not large.) We would not characterize Texas as an outlier.

(iii) When the lagged murder rate is added, $\hat{\beta}_{exec}$ becomes –.071 with $t = -2.34$. The coefficient changes sign and becomes nontrivial: each execution is estimated to reduce the murder rate by .071 (murders per 100,000 people).

(iv) When a Texas dummy is added to the regression from part (iii), its t is only –.37 (and the coefficient is only –1.02). So, it is not an outlier here, either. Dropping TX from the regression reduces the magnitude of the coefficient to –.045 with $t = -0.60$. Texas accounts for much of the sample variation in *exec*, and dropping it gives a very imprecise estimate of the deterrent effect.

C9.13 (i) The estimated equation, with the usual OLS standard errors in parentheses, is

$$\widehat{lsalary} = \underset{(0.26)}{4.37} + \underset{(.039)}{.165\, lsales} + \underset{(.049)}{.109\, lmktval} + \underset{(.014)}{.045\, ceoten} - \underset{(.0005)}{.0012\, ceoten^2}$$

$n = 177$, $R^2 = .343$.

(ii) There are eight observations with $|str_i| > 1.96$. Recall that if z has a standard normal distribution then $P(|z| > 1.96)$ is about .05. Therefore, *if* the str_i were random draws from a standard normal distribution, we would expect to see 5% of the observations having absolute value above 1.96 (or, rounding, two). Five percent of $n = 177$ is 8.85, so we would expect to see about nine observations with $|str_i| > 1.96$.

(iii) Here is the estimated equation using only the observations with $|str_i| \le 1.96$. Only the coefficients are reported:

$$\widehat{lsalary} = 4.14 + .154\, lsales + .153\, lmktval + .036\, ceoten - .0008\, ceoten^2$$

$n = 168$, $R^2 = .504$.

The most notable change is the much higher coefficient on *lmktval*, which increases to .153 from .109. Of course, all of the coefficients change.

(iv) Using LAD on the entire sample (coefficients only) gives

$$\widehat{lsalary} = 4.13 + .149\, lsales + .153\, lmktval + .043\, ceoten - .0010\, ceoten^2$$

$n = 177$, *Pseudo* $R^2 = .267$.

The LAD coefficient estimate on *lmktval* is closer to the OLS estimate on the reduced sample, but the LAD estimate on *ceoten* is closer to the OLS estimate on the entire sample.

(v) The previous findings show that it is not always true that LAD estimates will be closer to OLS estimates after "outliers" have been removed. However, it is true that, overall, the LAD estimates are more similar to the OLS estimates in part (iii). In addition to the two cases discussed in part (iv), note that the LAD estimate on *lmktval* agrees to the first three decimal places with the OLS estimate in part (iii), and it is much different from the OLS estimate in part (i).