PRINCIPLES OF ECONOMETRICS

5^{TH} EDITION

ANSWERS TO ODD-NUMBERED EXERCISES IN CHAPTER 8

EXERCISE 8.1

Under homoskedasticity $\operatorname{var}(e_i \mid \mathbf{x}_i) = \sigma_i^2 = \sigma^2$, which is an unvarying constant. Substitute this into the variance expression and simplify.

$$\operatorname{var}(b_2 \mid \mathbf{x}_i) = \left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1}\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\sigma_i^2\right]\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1}$$

$$= \left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1}\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\sigma^2\right]\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1}$$

$$= \sigma^2\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1}\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1}$$

$$= \sigma^2\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{-1} = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

The final expression is the conditional variance of the OLS estimator under SR1-SR5, as given in Equation (2.15) of *POE5*.

EXERCISE 8.3

(a) The combined error term is $e_{Ai} = \sum_{j=1}^{N_i} e_{ij}$. Its conditional variance is

$$\operatorname{var}\left(e_{Ai} \mid \mathbf{X}\right) = \operatorname{var}\left(\sum_{j=1}^{N_i} e_{ij} \mid \mathbf{X}\right) = \sum_{j=1}^{N_i}\operatorname{var}\left(e_{ij} \mid \mathbf{X}\right) + \sum_{j \neq s}\sum \operatorname{cov}\left(e_{ij}, e_{is} \mid \mathbf{X}\right) = N_i\sigma^2$$

(b) Now, $\overline{e}_i = \sum_{j=1}^{N_i} e_{ij} \Big/ N_i$ so that

$$\operatorname{var}\left(\overline{e}_i \mid \mathbf{X}\right) = \operatorname{var}\left(\sum_{j=1}^{N_i} e_{ij} \Big/ N_i \Big| \mathbf{X}\right) = \frac{1}{N_i^2}\operatorname{var}\left(\sum_{j=1}^{N_i} e_{ij} \mid \mathbf{X}\right) = \frac{1}{N_i^2}N_i\sigma^2 = \frac{\sigma^2}{N_i}$$

(c) The Bernoulli random variable $y_{ij} = 1$ with probability P_i, and $y_{ij} = 0$ with probability $1 - P_i$. Using methods from the probability primer $E\left(y_{ij}\right) = P_i$ and $\operatorname{var}\left(y_{ij}\right) = P_i\left(1 - P_i\right)$. $\overline{y}_i = \sum_{j=1}^{N_i} y_{ij} \Big/ N_i = p_i$. Given **X,** $\overline{e}_i$ has the same variance as $\overline{y}_i$. Then, again assuming outcomes on each farm and acre are uncorrelated,

$$\operatorname{var}\left(\overline{y}_i \mid \mathbf{X}\right) = \frac{1}{N_i^2}\operatorname{var}\left(\sum_{j=1}^{N_i} y_{ij} \mid \mathbf{X}\right) = \frac{1}{N_i^2}\left(\sum_{j=1}^{N_i}\operatorname{var}\left(y_{ij} \mid \mathbf{X}\right)\right)$$

$$= \frac{1}{N_i^2}N_iP_i\left(1 - P_i\right) = \frac{P_i\left(1 - P_i\right)}{N_i} = \operatorname{var}\left(\overline{e}_i \mid \mathbf{X}\right)$$

EXERCISE 8.5

(a) The values of x_{i2}, e_i and y_i are

x_{i2}	e_i	y_i
1	1	3
2	-1	2
3	0	4
4	6	11
5	-6	0

(b)

obs	y_i	$\hat{y}_i$	$\hat{e}_i$
1	3	3.4	-0.4
2	2	3.7	-1.7
3	4	4	0
4	11	4.3	6.7
5	0	4.6	-4.6

The least squares residuals for the first and fourth observations are -0.4 and 6.7. The sum of the least squares residuals is zero. In this example the sum of the true random errors is zero. In models that include an intercept the sum of the least squares residuals is always zero. The sum of the true random errors is not zero in general. An assumption for the simple linear model is that $E(e_i \mid \mathbf{X}) = E(e_i) = 0$. That is the population average of the random errors is zero.

(c) The *F*–statistic will have the *F*-distribution with 1 numerator degree of freedom and 3 denominator degrees of freedom if the null hypothesis of no heteroskedasticity is true. The 95th percentile of the *F*-distribution from Statistical Table 4 is 10.13, or 10.127964 using computer software. The *p*-value is 0.037. We reject the null hypothesis of homoskedasticity at the 5% level.

(d) The value of the LM test statistic is $NR^2 = 5(.8108) = 4.054$. The test statistic has the $\chi^2_{(1)}$ distribution if the null hypothesis of no heteroskedasticity is true. Using Statistical Table 3 the 95th percentile of the $\chi^2_{(1)}$ is 3.841. The *p*-value is 0.044. We reject the null hypothesis of homoskedasticity at the 5% level.

(e) The third observation is $\hat{e}_i = 0$ so $\hat{e}_i^2 = 0$ and $\ln(\hat{e}_i^2)$ is undefined for that observation. The software counts that as a "missing value" and discards the observation.

(f) The transformed observations for $i = 1$ and $i = 4$ are

$$y_1^* = 3 \qquad x_{11}^* = 1 \qquad x_{12}^* = 1 \qquad\qquad y_4^* = 1.637 \qquad x_{41}^* = 0.149 \qquad x_{42}^* = 0.595$$

EXERCISE 8.7

(a) In Appendix 7B of *POE5* it is shown that $\sum\left(d_i-\bar{d}\right)^2=N_0N_1/N$. Under conditional homoskedasticity $\operatorname{var}\left(b_2\,|\,\mathbf{d}\right)=\dfrac{\sigma^2}{\sum\left(d_i-\bar{d}\right)^2}=N\sigma^2/\left(N_0N_1\right)$.

(b) The difference estimator is $b_2=\bar{y}_1-\bar{y}_0$. The intercept estimator is $b_1=\bar{y}-b_2\bar{d}$, where $\bar{d}=\sum_{i=1}^{N}d_i/N=N_1/N$. The sample mean $\bar{y}$ can be rewritten as

$$\begin{aligned}\bar{y}&=\frac{\sum_{i=1}^{N}y_i}{N}=\frac{\sum_{d_i=1}y_i+\sum_{d_i=0}y_i}{N_1+N_0}=\frac{\sum_{d_i=1}y_i}{N_1+N_0}+\frac{\sum_{d_i=0}y_i}{N_1+N_0}\\&=\frac{N_1\left(\sum_{d_i=1}y_i/N_1\right)}{N_1+N_0}+\frac{N_0\left(\sum_{d_i=0}y_i/N_0\right)}{N_1+N_0}=\frac{N_1\bar{y}_1}{N_1+N_0}+\frac{N_0\bar{y}_0}{N_1+N_0}\end{aligned}$$

Then

$$\begin{aligned}b_1=\bar{y}-b_2\bar{d}&=\frac{N_1\bar{y}_1}{N_1+N_0}+\frac{N_0\bar{y}_0}{N_1+N_0}-\left(\bar{y}_1-\bar{y}_0\right)\frac{N_1}{N_1+N_0}\\&=\frac{N_1\bar{y}_1}{N_1+N_0}+\frac{N_0\bar{y}_0}{N_1+N_0}-\bar{y}_1\frac{N_1}{N_1+N_0}+\bar{y}_0\frac{N_1}{N_1+N_0}\\&=\frac{N_0\bar{y}_0}{N_1+N_0}+\frac{\bar{y}_0N_1}{N_1+N_0}=\bar{y}_0\end{aligned}$$

The least squares fitted values are

$$\hat{y}_i=\begin{cases}b_1=\bar{y}_0 & \text{if } d_i=0\\ b_1+b_2=\bar{y}_1 & \text{if } d_i=1\end{cases}$$

Then the least squares residuals are

$$\hat{e}_i=\begin{cases}y_i-\bar{y}_0 & \text{if } d_i=0\\ y_i-\bar{y}_1 & \text{if } d_i=1\end{cases}$$

The sum of squared residuals is

$$\begin{aligned}\sum_{i=1}^{N}\hat{e}_i^2&=\sum_{d_i=1}\hat{e}_i^2+\sum_{d_i=0}\hat{e}_i^2\\&=\sum_{d_i=1}\left(y_i-\bar{y}_1\right)^2+\sum_{d_i=0}\left(y_i-\bar{y}_0\right)^2\\&=SST_1+SST_0\end{aligned}$$

Then $\hat{\sigma}^2=\sum_{i=1}^{N}\hat{e}_i^2/\left(N-2\right)=\left(SST_0+SST_1\right)/\left(N-2\right)$, and

$$\widehat{\operatorname{var}}\left(b_2\,|\,\mathbf{d}\right)=\frac{\hat{\sigma}^2}{\sum\left(d_i-\bar{d}\right)^2}=N\hat{\sigma}^2/\left(N_0N_1\right)=\frac{N}{N_0N_1}\frac{SST_0+SST_1}{N-2}$$

(c) Equation (2.14) is

$$\operatorname{var}(b_1 \mid \mathbf{d}) = \sigma^2 \left[\frac{\sum d_i^2}{N\sum\left(d_i - \bar{d}\right)^2} \right] = \sigma^2 \left[\frac{N_1}{N\dfrac{N_0 N_1}{N}} \right] = \frac{\sigma^2}{N_0}$$

Using results from above,

$$\widehat{\operatorname{var}}(b_1 \mid \mathbf{d}) = \frac{\hat{\sigma}^2}{N_0} = \frac{\left(SST_0 + SST_1\right)}{N_0\left(N-2\right)}$$

(d) Because the treatment and control groups are statistically independent,

$$\operatorname{var}\left(\bar{y}_1 - \bar{y}_0\right) = \operatorname{var}\left(\bar{y}_1\right) + \operatorname{var}\left(\bar{y}_0\right) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

An unbiased estimator of σ_1^2 is $\hat{\sigma}_1^2 = SST_1/\left(N_1 - 1\right)$ and similarly $\hat{\sigma}_0^2 = SST_0/\left(N_0 - 1\right)$ so that

$$\widehat{\operatorname{var}}(b_2 \mid \mathbf{d}) = \widehat{\operatorname{var}}\left(\bar{y}_1 - \bar{y}_0 \mid \mathbf{d}\right) = \frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_0^2}{N_0} = \frac{SST_1}{\left(N_1 - 1\right)N_1} + \frac{SST_0}{\left(N_0 - 1\right)N_0}$$

(e) Equation (8.9) becomes

$$\begin{aligned}
\widehat{\operatorname{var}}(b_2 \mid \mathrm{d}) &= \left[\Sigma\left(d_i - \bar{d}\right)^2\right]^{-1} \left\{\Sigma\left[\left(d_i - \bar{d}\right)^2 \left(\frac{N}{N-2}\right)\hat{e}_i^2\right]\right\} \left[\Sigma\left(d_i - \bar{d}\right)^2\right]^{-1} \\
&= \left[\frac{N_0 N_1}{N}\right]^{-1} \left(\frac{N}{N-2}\right) \left\{\Sigma\left[\left(d_i - \bar{d}\right)^2 \hat{e}_i^2\right]\right\} \left[\frac{N_0 N_1}{N}\right]^{-1} \\
&= \left[\frac{N_0 N_1}{N}\right]^{-1} \left(\frac{N}{N-2}\right) \left\{\sum\nolimits_{d_i=1}\left[\left(1-\bar{d}\right)^2 \hat{e}_i^2\right] + \sum\nolimits_{d_i=0}\left[\left(-\bar{d}\right)^2 \hat{e}_i^2\right]\right\} \left[\frac{N_0 N_1}{N}\right]^{-1} \\
&= \left[\frac{N_0 N_1}{N}\right]^{-1} \left(\frac{N}{N-2}\right) \left\{\frac{N_0^2}{N^2}\sum\nolimits_{d_i=1}\hat{e}_i^2 + \frac{N_1^2}{N^2}\sum\nolimits_{d_i=0}\hat{e}_i^2\right\} \left[\frac{N_0 N_1}{N}\right]^{-1} \\
&= \frac{N^2}{N_0^2 N_1^2}\left(\frac{N}{N-2}\right)\left\{\frac{N_0^2}{N^2}SST_1 + \frac{N_1^2}{N^2}SST_0\right\} = \left(\frac{N}{N-2}\right)\left(\frac{SST_1}{N_1^2} + \frac{SST_0}{N_0^2}\right) \\
&= \left(\frac{N}{N-2}\right)\left(\frac{SST_1/N_1}{N_1} + \frac{SST_0/N_0}{N_0}\right) = \left(\frac{N}{N-2}\right)\left(\frac{\tilde{\sigma}_1^2}{N_1} + \frac{\tilde{\sigma}_0^2}{N_0}\right)
\end{aligned}$$

The difference is that in part (d) the unbiased variance estimators were used, and here the method of moments estimator is used, as on the top of page 821, in *POE5*.

(f) It becomes $\widetilde{\operatorname{var}}\left(\bar{y}_1 - \bar{y}_0 \mid \mathbf{d}\right) = \dfrac{\tilde{\sigma}_1^2}{N_1} + \dfrac{\tilde{\sigma}_0^2}{N_0}$. This is the same estimator used in part (d) except the biased estimators for the variances are used.

EXERCISE 8.9

(a) The estimated coefficient is positive indicating that homes close to a major university have a higher expected (average) price, holding all else constant. The coefficient value of 26.6657 suggests that the average price of a home close to the university is $26,665.70 higher than that of an identical home not close to the university. The *t*-value is 2.68, which is greater than 2.58, the critical value for a test at the 1% level of significance; we reject the null hypothesis that the location effect is zero.

(b) (i) False (ii) False (iii) False (iv) False, however it is true most of the time (v) True (vi) False (vii) True

(c) The test statistic is NR^2 which has an approximate $\chi^2_{(6)}$ distribution in large samples if the null hypothesis of homoskedasticity is true. The 5% test critical value is 12.592. The value of the test statistic is $NR^2 = 500 \times 0.3028 = 151.4$. Thus, we reject the null hypothesis of homoskedasticity and accept the alternative that heteroskedasticity is present and somehow related to the model variables.

(d) The standard errors are robust in the sense that they are valid in large samples whether the errors are homoskedastic or heteroskedastic. The intercept, and the coefficient of *SQFT* have wider interval estimates using the robust standard errors, but the coefficient of *CLOSE* has a narrower interval estimate using the robust standard error. None of the coefficients change their statistical 5% significance.

(e) When dividing by *SQFT* the researcher is assuming $\operatorname{var}(e_i \mid \mathbf{X}) = \sigma^2 SQFT_i^2$. Yes there are some large differences. The estimated intercept becomes insignificant. The coefficient of *SQFT* falls from 13.3417 to 7.5803 but remains significant at the 1% level. The change in magnitude means that the estimate of the value of an additional 100 sqft of living area drops from $13,341.70 to $7,580.30. The coefficient on *AGE* changes sign but remains insignificant. The coefficient on *FIREPLACE* changes from insignificant to significant at the 5% level, and the value suggests that a fireplace increases the average value of house by $17,382.70. The coefficient of *TWOSTORY* changes from being insignificant to significant at the 10% level. The magnitude suggests that the average price of a two-story home is higher by $26,722.40.

(f) The test statistic $NR^2 = 500 \times 0.0237 = 11.85$. The test degrees of freedom is 6 [the transformed variable *SQFT* = 1 and thus falls from the regression due to collinearity] so that the test 5% critical value is 12.592. Using this test we fail to reject the null hypothesis of homoskedasticity at the 5% level.

(g) It is prudent. The assumption of a "model" of heteroskedasticity is always open to debate. How are we to know that the assumed model $\operatorname{var}(e_i \mid \mathbf{X}) = \sigma^2 SQFT_i^2$ is correct or incorrect? While it is reasonable, and seems to have reduced the heteroskedasticity, the result of the test is that we fail to reject the null hypothesis, not that there is no heteroskedasticity remaining. Using robust standard errors changes little in terms of significance. Only the coefficient of *FIREPLACE* has its significance level fall from 5% to 10%. The robust GLS standard error for the variable *SQFT* is about twice as large as the usual GLS standard error, meaning that the interval estimate of the effect of *SQFT* is about twice as wide, changing from about [6.56, 8.60] to [5.66, 9.51].

EXERCISE 8.11

(a) The plot suggests the error variance is increasing with *ACRES*. The value of the Breusch-Pagan LM test for heteroskedasticity is $NR^2 = 44 \times 0.2068 = 9.0992$. The critical value is $\chi^2_{(0.99,1)} = 6.635$ so that we reject homoskedasticity based on this test. The *t*-value for the *ACRES* coefficient is $t = 2.024/0.612 = 3.307$ which is significant at the 1% level. Thus, based on these three pieces of information we suspect that heteroskedasticity is present and related to *ACRES*.

(b) This model presumes that $\text{var}\left(e_i \mid \mathbf{X}\right) = \sigma^2 ACRES_i^2$.

(c) The value of this variable is likely to be small, since in most cases farms consist of a good number of acres. The variable would exhibit little variation and contribute little explanatory power.

(d) If the transformation in (c) was successful we should find little evidence relating the squared residuals from the transformed model to *ACRES*. The coefficient is negative suggesting that the variation in the residuals now falls as *ACRES* increase, and while a 10% level of significance is not very impressive, the suggestion is that perhaps the transformation was too aggressive, and that we over-corrected. The BP test value $NR^2 = 44 \times 0.0767 = 3.3748$. The 5% critical values is $\chi^2_{(.95,1)} = 3.841$, so that using this test we have no evidence of heteroskedasticity in the transformed model.

(e) In the multiplicative model $\sigma_i^2 = \exp\left(\alpha_1 + \alpha_2 \ln\left(ACRES_i\right)\right) = \sigma^2 ACRES_i^{\alpha_2}$, or, using the estimates, $\hat{\sigma}_i^2 = \sigma^2 ACRES^{1.11}$. That is, this result suggests that the model used in (b) was overly aggressive, and that perhaps the error variance is proportional to *ACRES* rather than its square. Similarly, the residuals from the transformed model suggest $\tilde{\sigma}_i^2 = \sigma^2 / ACRES^{1.21}$; so that we have overcorrected in part (b), or (c).

(f) Yes, as we found in part (e), the transformation may not have eliminated heteroskedasticity. By using a robust standard error we are protecting our inferences from this error. Note that this exercise is using tests and procedures that are valid in large samples, and $N = 44$ is marginal at best.

EXERCISE 8.13

(a) The LM test statistic value is $NR^2 = 1100 \times 0.0344 = 37.84$. The test statistic critical value is $\chi^2_{(.95,4)} = 9.488$. Therefore we reject the null hypothesis of homoskedasticity and accept the alternative that heteroskedasticity is present.

(b) Using conventional standard errors, the *t*-value for the coefficient of *ADVANCED* is 1.87 with a two-tail *p*-value of 0.062. Using the robust standard error the *t*-value is 1.58 and it has a two-tail *p*-value of 0.114. Thus, we do not conclude that *ADVANCED* is a significant explanatory factor using the 5% level. While it is usually the case that robust standard errors are larger than conventional ones when heteroskedasticity is present, it is not true in this case. Using the robust standard error leads to a smaller *t*-value which means the robust standard error is larger.

(c) This regression is based on a model of multiplicative heteroskedasticity, $\sigma_i^2 = \exp\left(\alpha_1 + \alpha_2 INCOME_i + \alpha_3 ADVANCED_i\right)$. The variable *INCOME* is significant at the 1% level, and has a positive sign, suggesting that higher income households have a larger error variation.

(d) Using the FGLS estimates we estimate that a household with an advanced degree holder will spend an average of about \$9.01 more per month per person on entertainment. Using the coefficient *t*-value the two tail *p*-value is 0.055, suggesting that the effect of an advanced degree holder is not significant at the 5% level, despite the *t*-value 1.92 being larger than the *t*-value robust standard error, which is the correct comparison because of the evidence of heteroskedasticity found.

(e) The LM test degrees of freedom is 4, so that the 5% critical value is 9.488. Thus, we fail to reject the null hypothesis of homoskedasticity using this test. White's test includes all the squares and cross-products of the variables. There are a total of 16 terms, but 3 of the variables are indicator variables whose squares and cross-products are still 1 or 0. Thus the White test would include the 10 variables *INCOME*, *URBAN*, *COLLEGE*, *ADVANCED*, $INCOME^2$, $INCOME \times URBAN$, $INCOME \times COLLEGE$, $INCOME \times ADVANCED$, $URBAN \times COLLEGE$, $URBAN \times ADVANCED$. Note that $COLLEGE \times ADVANCED$ is not included because those two groups are mutually exclusive making $COLLEGE \times ADVANCED$ always equal to zero. The critical value for the White test is 18.307, thus we fail to reject homoskedasticity using the White test also.

(f) The coefficient of advanced in the log-linear model is 0.2315, which implies roughly that a household with an advanced degree holder will spend on average 23.15% more per person than a household not having an advanced degree holder. Using the exact calculation is Chapter 7.3.2 the estimated effect is 26.05%. Furthermore, in the log-linear model the effect of advanced is statistically significant at the 1% level using a two tail test, with critical value 2.58. The explanation is illustrated in Example 8.8 of *POE5*. For skewed variables like entertainment expenditure, the log-transformation "regularizes" the data and can in some cases reduce or eliminate heteroskedasticity as it has done in this case.

EXERCISE 8.15

(a) The regression function is

$$E\left(FTE_i\right) = \begin{cases} \beta_1 & d_i = 0 \\ \beta_1 + \beta_2 & d_i = 1 \end{cases}$$

The estimate of β_1 is the sample mean of the observations for the control group, PA. From the table the sample mean for these 77 observations is 21.16558, which is the value from the estimated regression. The estimate of $\beta_1 + \beta_2$ is the sample mean of the observations for the treatment group, NJ. From the table the sample mean for these 319 observations is 21.02743. The estimated value of β_2 is the difference 21.02743 − 21.16558 = −0.13815, which, apart from the rounding difference, is the slope estimate in the regression.

(b) $$\widehat{\operatorname{var}}\left(b_2 \mid \mathbf{d}\right)=\left(\frac{N}{N-2}\right)\left(\frac{SST_0+SST_1}{N_0 N_1}\right)=\left(\frac{396}{394}\right)\left(\frac{5206.326+27462.572}{77\times 319}\right)=1.3367557$$

$$\operatorname{se}\left(b_2\right)=\sqrt{\widehat{\operatorname{var}}\left(b_2 \mid \mathbf{d}\right)}=\sqrt{1.3367557}=\ 1.1561815$$

(c) The test statistic is $F=\hat{\sigma}_0^2/\hat{\sigma}_1^2 \sim F_{(N_0-1,N_1-1)}$ if the null hypothesis is true. The calculated F is

$$F=68.50429/86.36029=0.7932383$$

The critical values for a two-tail test at the 1% level are $F_{(.005,76,318)}=.60774596$ and $F_{(.995,76,318)}=1.5523351$. The test statistic falls in the non-rejection region; we cannot reject the null hypothesis $\sigma_0^2=\sigma_1^2$.

(d) $$\text{HC1 } \widehat{\operatorname{var}}\left(b_2 \mid \mathbf{d}\right)=\frac{N}{N-2}\left(\frac{SST_0}{N_0^2}+\frac{SST_1}{N_1^2}\right)=\frac{396}{394}\left(\frac{5206.326}{77^2}+\frac{27462.572}{319^2}\right)=1.1538126$$

$$\text{robse}=\sqrt{1.1538126}=\ 1.0741567$$

This value is the same as shown in the regression output.

(e) $$\text{HCE2 } \widehat{\operatorname{var}}\left(b_2 \mid \mathbf{d}\right)=\frac{SST_0}{N_0\left(N_0-1\right)}+\frac{SST_1}{N_1\left(N_1-1\right)}=\frac{5206.326}{77(76)}+\frac{27462.572}{319(318)}=1.160388$$

$$\text{rob2}=\sqrt{1.160388}=1.0772131$$

This variance estimator is not unbiased in this regression application, whether $\sigma_0^2=\sigma_1^2$ or not, however it is consistent.

(f) $$\text{HCE0 } \widehat{\operatorname{var}}\left(b_2 \mid \mathbf{d}\right)=\left(\frac{SST_0}{N_0^2}+\frac{SST_1}{N_1^2}\right)=\left(\frac{5206.326}{77^2}+\frac{27462.572}{319^2}\right)=1.1479853$$

$$\text{rob0}=\sqrt{1.1479853}=1.0714407$$

Having dropped the correction factor, which is larger than 1, we find the estimated standard error slightly smaller.

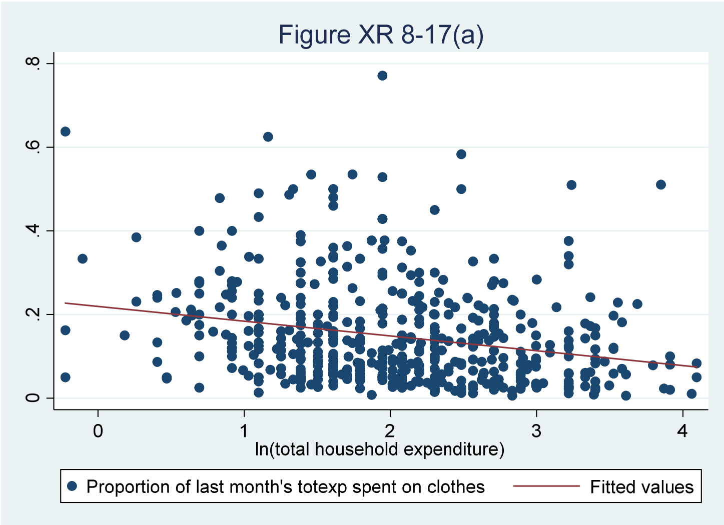
(g) $$\text{HCE3 } \widehat{\operatorname{var}}\left(b_2 \mid \mathbf{d}\right)=\left(\frac{\hat{\sigma}_0^2}{N_0-1}\right)+\left(\frac{\hat{\sigma}_1^2}{N_1-1}\right)=\frac{68.50429}{76}+\frac{86.36029}{318}=1.1729455$$

$$\text{rob3}=\sqrt{1.1729455}=1.0830261$$

Comparing the alternatives, HCE2 is the largest among the robust variances, although the conventional OLS variance estimator is largest overall. The smallest is HCE0.

EXERCISE 8.17

(a)



The estimates are reported in column (1) of Table XR8.17. The formula for the elasticity in the linear-log model $PCLOTHES = \beta_1 + \beta_2 \ln(TOTEXP) + e$ is

$$\varepsilon = \frac{dCLOTHES}{dTOTEXP} \times \frac{TOTEXP}{CLOTHES} = \frac{\beta_1 + \beta_2 \left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

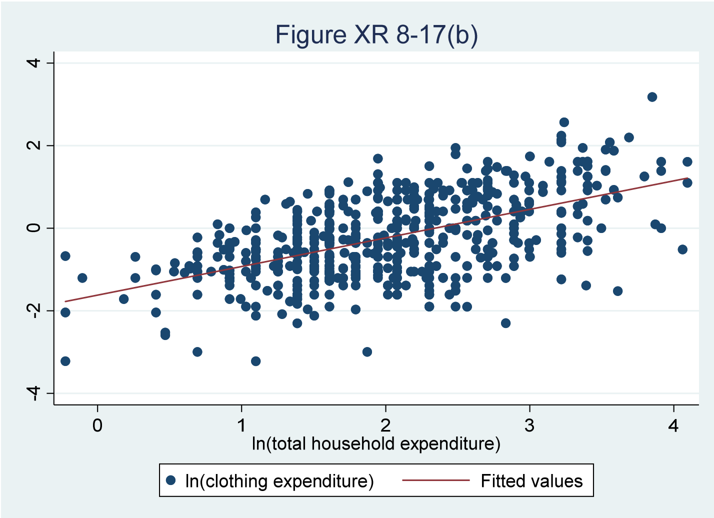
The sample mean of *TOTEXP* is 10.39265. Inserting the estimates for the unknown parameters we obtain an elasticity of 0.7413711. Using software for nonlinear functions of parameters allows us to compute a "delta method" standard error, 0.0503168, and a 95% interval estimate [0.6427521, 0.8399901]. The estimated elasticity is less than one, as expected for a necessary good.

Table XR8.17

	(1) part (a)	(2) part(b)	(3) part(c)
C	0.2193	-1.6186	0.3158
	(0.0136)	(0.0975)	(0.0953)
ln(*TOTEXP*)	-0.0353	0.6918	
	(0.0062)	(0.0448)	
TOTEXP			0.0978
			(0.0068)
N	556	556	556
F	31.9838	238.0263	205.0856

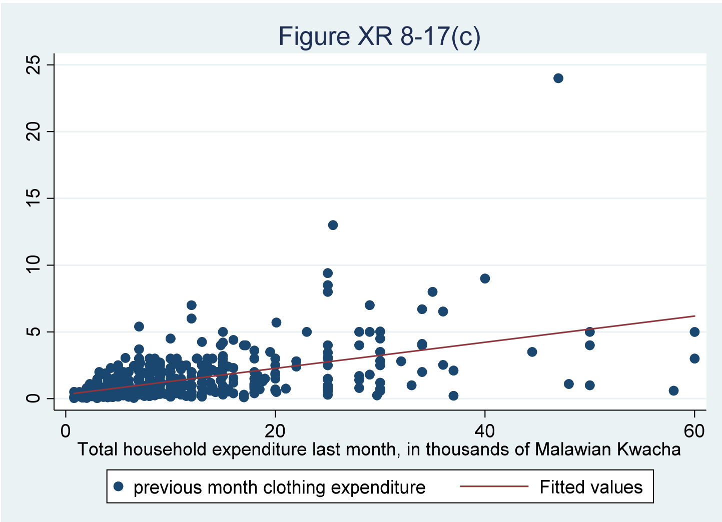
Standard errors in parentheses

(b)



The estimates for the log-log model $\ln(CLOTHES) = \alpha_1 + \alpha_2 \ln(TOTEXP) + v$ are in column (2) of Table XR8.17. The elasticity is $\hat{\alpha}_2 = 0.6918$. The 95% interval estimate is [0.6037309, 0.7798889]. The elasticity computed in part (a) is within the interval.

(c)



The estimates for the linear regression model $CLOTHES = \gamma_1 + \gamma_2 TOTEXP + u$ are in column (3) of Table XR8.17.

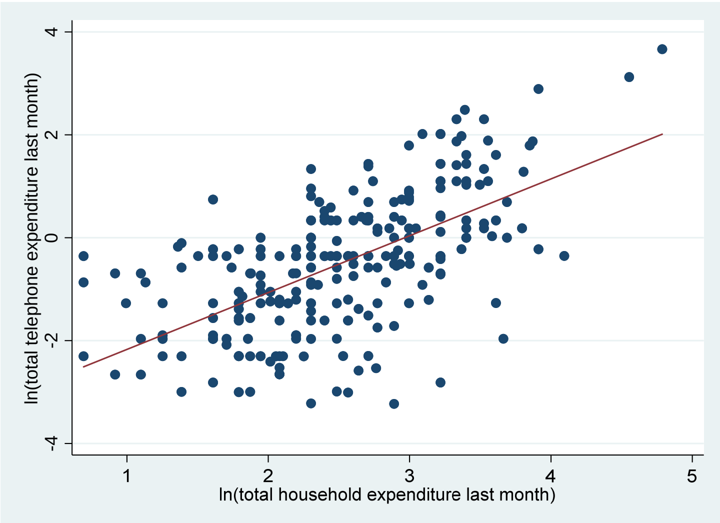
The elasticity at the means is $\hat{\varepsilon} = \hat{\gamma}_2 \dfrac{\overline{TOTEXP}}{\hat{\gamma}_1 + \hat{\gamma}_2 \overline{TOTEXP}} = 0.7628415$ with a "delta method" standard error of 0.0645408, leading to a 95% interval estimate of [0.6360668, 0.8896161].

(d) For the linear-log model $PCLOTHES = \beta_1 + \beta_2 \ln(TOTEXP) + e$, the Breusch-Pagan LM test statistic value is 3.24, with a *p*-value of 0.0717. For the log-log model $\ln(CLOTHES) = \alpha_1 + \alpha_2 \ln(TOTEXP) + v$ the value of the LM statistic is 12.82, with a *p*-value of 0.0003. For the linear model $CLOTHES = \gamma_1 + \gamma_2 TOTEXP + u$ the LM statistic value is 44.70 with a *p*-value of 0.0000. The linear-log model seems to be the only one where heteroskedasticity is not an issue.

(e) For the log-log model the elasticity interval estimate using the default robust standard error is [0.6005739, 0.783046]. For the linear model the elasticity interval becomes [0.5325294, 0.9931535]. For the log-log model the interval estimate is only slightly wider than using conventional OLS standard errors. For the linear model the interval estimate is much wider, which is to be expected given the strong heteroskedasticity in that model.

EXERCISE 8.19

(a)



There is a positive relationship between the log of telephone expenditure and the log of total household expenditure.

(b) In this log-log model $\ln(TELEPHONE) = \beta_1 + \beta_2 \ln(TOTEXP) + e$ the coefficient β_2 is the elasticity. The estimated values are

$$\widehat{\ln}(TELEPHONE) = -3.273096 + \underset{(0.09303)}{1.10419} \ln(TOTEXP)$$
$$\text{(se)}$$

We estimate that the elasticity of telephone expenditures with respect to total expenditures is 1.104. Because this value is larger than one, we would classify it as a luxury. The 95% interval estimate is [0.9209325, 1.287447], which does include some values less than one, so we could not reject the null hypothesis that the elasticity is 0.95 using a 2-tail test at the 5% level of significance.

(c) Regressing the squared OLS residuals on $\ln(TOTEXP)$ we obtain $R^2 = 0.0053$, and $NR^2 = 245 \times 0.0053 = 1.2948$. The critical value is $\chi^2_{(.95,1)} = 3.841458$. Thus, we fail to reject the null hypothesis of homoskedasticity at the 5% level. For the White test we regress the squared residuals on $\ln(TOTEXP)$ and its square. We obtain $R^2 = 0.0148$ and $NR^2 = 245 \times 0.0148 = 3.632$. The critical value is $\chi^2_{(.95,2)} =$ 5.99146. Thus we fail to reject the null hypothesis of homoskedasticity at the 5% level.

(d) The estimated model is

$$\widehat{PTELEPHONE} = 0.0474438 + 0.0133659 \ln(TOTEXP)$$
$$\text{(se)} \qquad\qquad (0.007289)$$

The critical value for a one tail test is $t_{(.95,243)} =$ 1.6511. The *t*-statistic value is 1.83 for a right tail test, so we conclude that there is a positive relationship between the proportion of telephone expenditures and total expenditures.

(e) The elasticity in this model is

$$\varepsilon = \frac{dFOOD}{dTOTEXP} \times \frac{TOTEXP}{FOOD} = \frac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

The median of *TOTEXP* is 12, giving an elasticity of $\hat{\varepsilon} = 1.165713$.

The 95% interval estimate, using the standard error 0.0915798 and critical values ±1.96 is

[0.9862202, 1.345206]

The point estimate of the elasticity is slightly larger; the interval estimate shifted rightward.

(f) Regressing the squared residuals on $\ln(TOTEXP)$ yields $NR^2 = 245 \times 0.0044 =$ 1.0898. The critical value is $\chi^2_{(.95,1)} = 3.841458$. Thus we fail to reject homoskedasticity. Using the White test we regress the squared residuals on $\ln(TOTEXP)$ and its square and obtain $NR^2 = 245 \times 0.0617 = 15.1125$. Thus, using the White test with critical value $\chi^2_{(.95,2)} =$ 5.99146, we do reject homoskedasticity at the 5% level.

(g) Estimating the auxiliary equation for the skedastic equation we obtain

$$\ln\left(\hat{e}^2\right) = -6.946123 + 0.2489945 \ln(TOTEXP)$$
$$\text{(se)} \qquad\qquad (0.16599)$$

The estimated FGLS model is

$$\widehat{PTELEPHONE}_i = 0.0672727 + 0.0052903 \ \ln(TOTEXP_i)$$
$$\text{(se)} \qquad\qquad (0.0071998)$$

We fail to conclude that there is a positive relation between the proportion of expenditures and the log of total expenditures. While not required, the elasticity based on the FGLS estimates is 1.066 with 95% interval estimate [0.8912, 1.2404].

(h) Using robust standard errors, the model becomes

$$\widehat{PTELEPHONE}_i = 0.0672727 + 0.0052903\ \ln\left(TOTEXP_i\right)$$
$$(\text{se}) \qquad\qquad (0.0099311)$$

(i) Using a log-log model, or the proportion-log model, the estimated elasticity of phone expenditures with respect to total expenditures is close to one.

EXERCISE 8.21

(a) $COKE_{ij}$ is a binary variable which assigns 1 if the shopper buys coke and zero otherwise. Therefore, the total number of shoppers who buy coke in store *i* is given by $\sum_{j=1}^{N_i} COKE_{ij}$ and the proportion will be given by $\frac{1}{N_i}\sum_{j=1}^{N_i} COKE_{ij}$, which is $\overline{COKE_{i\bullet}}$.

(b) $$E\left(\overline{COKE_{i\bullet}}\right) = \frac{1}{N_i} E\left(\sum_{j=1}^{N_i} COKE_{ij}\right) = \frac{1}{N_i}\sum_{j=1}^{N_i} E\left(COKE_{ij}\right) = \frac{1}{N_i} N_i P_i = P_i$$

$$\text{var}\left(\overline{COKE_{i\bullet}}\right) = \frac{1}{N_i^2}\text{var}\left(\sum_{j=1}^{N_i} COKE_{ij}\right) = \frac{1}{N_i^2}\sum_{j=1}^{N_i}\text{var}\left(COKE_{ij}\right) + \text{zero covariance terms}$$

$$= \frac{1}{N_i^2}\sum_{j=1}^{N_i} P_i(1-P_i) = \frac{N_i}{N_i^2} P_i(1-P_i) = \frac{P_i(1-P_i)}{N_i}$$

(c) P_i is the population proportion of customers in store *i* who purchase Coke. We can think of it as the proportion evaluated for a large number of customers in store *i*, or the probability that a customer in store *i* will purchase Coke. We can write

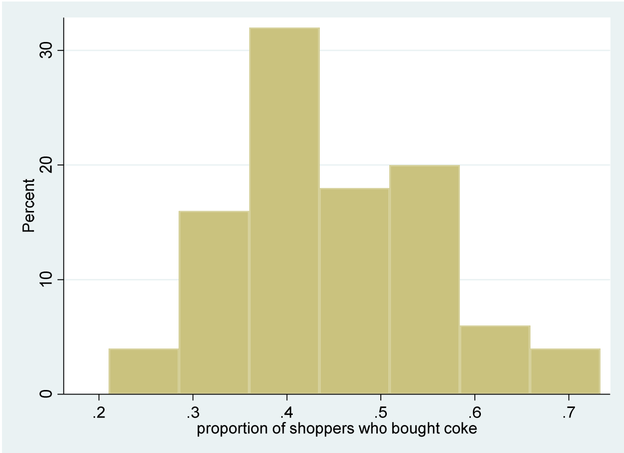
$$P_i = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i$$

(d)

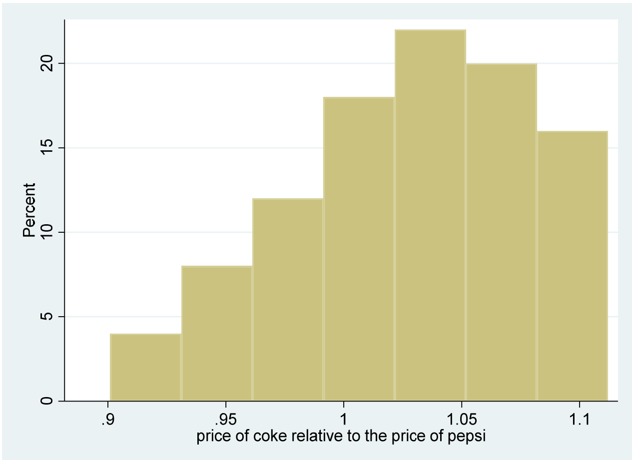
stats	$\overline{COKE}$	*PRATIO*
mean	.4485044	1.02694
Std. dev.	.1084821	.0505716
CV	24.18753	4.92449

We find that the mean of $\overline{COKE}$ is 0.449, or 44.9% of customers purchase Coke rather than Pepsi. The price ratio shows that the average price ratio of Coke to Pepsi is slightly larger than one, meaning that Coke, on average, is slightly more expensive. The coefficient of variation for $\overline{COKE}$ is 24.19% and for *PRATIO* is 4.92%. We prefer a larger variation in *PRATIO* as that helps identify the coefficient and reduce its variance. The variation in

$\overline{COKE}$ is the total variation that includes both the systematic portion and the random error portion.



The histogram for the proportion of shoppers who bought Coke shows quite a bit of variation from store to store, from a low of about 20% to a max of over 70%



The histogram of relative prices shows that in most stores the price of Coke is higher than the price of Pepsi.

(e) The estimated equation is

$$\hat{P} = 0.5195815 - 0.0659PRATIO + 0.08571DISP\_COKE - 0.1097DISP\_PEPSI$$
$$(\text{robse}) \quad (0.3886) \quad (0.0397) \quad (0.04073)$$

The logical alternative hypothesis is $\beta_3 \neq -\beta_4$, implying a two-tail test. The 5% critical values are $\pm t_{(0.975,46)} = \pm 2.0129$. The variance of the coefficient of *DISP_COKE* is 0.00157808, the variance of the coefficient of *DISP_PEPSI* is 0.00165897, the covariance is −0.00081794, and the test value is $t = -0.60$, which is not significant. We cannot conclude that the displays of Coke and Pepsi have effects of different magnitudes.

(f) The estimated model is

$$\widehat{COKE_{i\bullet}} = 0.5182 - 0.0679 PRATIO_i + 0.0976\, DISP_i$$
$$\text{(se)} \qquad (0.3094) \qquad (0.0387)$$

The value of the White test statistic is 6.7178 compared to the critical value $\chi^2_{(.95,5)} = 11.07$; thus, using this test we fail to reject homoskedasticity. The test using the candidate variable N_i yields $NR^2 = 50 \times 0.0933 = 4.6655$ which is greater than the critical value $\chi^2_{(.95,1)} = 3.8415$; thus we correctly conclude that heteroskedasticity is present. See part (b).

(g) The fitted values *p* have summary statistics:

variable	mean	Std. dev.	max	min
p	.4485044	.0407753	.5546912	.3471293

The estimated variance for each of the stores is calculated using $\widehat{\text{var}}\left(\overline{COKE_{i\bullet}}\right) = \dfrac{p_i(1-p_i)}{N_i}$.

(h) The fitted FGLS model is

$$\widehat{COKE_{i\bullet}} = 0.5617 - 0.1109 PRATIO_i + 0.08954\, DISP_i$$
$$\text{(se)} \qquad (.2983) \qquad (0.0380)$$

These estimates are similar to those in part (f). The coefficient of *PRATIO* is still insignificant. In part (d) we observed that *PRATIO* did not vary much relative to its mean, so that it is difficult to obtain a very precise estimate.

EXERCISE 8.23

(a) The OLS estimates are reported in column (1) of Table XR8.23a. The estimated coefficient of *UNION* is positive, and its value 0.0719 suggests that the expected wage of union members is 7.19% higher than non-union members holding other factors equal. The *t*-statistic value is 5.18 and the test critical value, for a one-tail test, is $t_{(0.995,9787)} = 2.576$. Thus, we can conclude that the effect of union membership on average wages is positive.

(b) The auxiliary regression estimates from regressing $\hat{e}_i^2$ on *UNION* and *METRO* are in column (1) of Table XR8.23b. The test value is $NR^2 = 21.175743$ and the 99th percentile of the chi-square distribution with 2 degrees of freedom is 9.210, so we reject the null hypothesis of homoskedasticity at the 1% level of significance.

(c) These auxiliary regression results are in columns (2) and (3) of Table XR8.23b, respectively. Both coefficients are positive suggesting the being a union member or living in a metropolitan area is positively correlated with the size of the error variance. Using unadjusted OLS standard errors, the two coefficients are significant at the 5% or better level.

Table XR8.23a

	(1) OLS	(2) OLS ROB	(3) FGLS	(4) FGLS ROB
C	0.7176***	0.7176***	0.7490***	0.7490***
	(0.057846)	(0.056102)	(0.056123)	(0.056300)
EDUC	0.1330***	0.1330***	0.1315***	0.1315***
	(0.003673)	(0.003683)	(0.003625)	(0.003714)
EXPER	0.0510***	0.0510***	0.0518***	0.0518***
	(0.002528)	(0.002575)	(0.002427)	(0.002455)
$EXPER^2$	-0.0005***	-0.0005***	-0.0005***	-0.0005***
	(0.000027)	(0.000028)	(0.000026)	(0.000027)
EXPER×EDUC	-0.0013***	-0.0013***	-0.0014***	-0.0014***
	(0.000132)	(0.000139)	(0.000125)	(0.000133)
FEMALE	-0.1698***	-0.1698***	-0.1731***	-0.1731***
	(0.009477)	(0.009433)	(0.009387)	(0.009367)
BLACK	-0.1114***	-0.1114***	-0.1106***	-0.1106***
	(0.016845)	(0.015987)	(0.016575)	(0.016018)
UNION	0.0719***	0.0719***	0.0818***	0.0818***
	(0.013898)	(0.013334)	(0.012826)	(0.013368)
METRO	0.1121***	0.1121***	0.1075***	0.1075***
	(0.012241)	(0.011547)	(0.011516)	(0.011441)
SOUTH	-0.0352**	-0.0352*	-0.0366**	-0.0366**
	(0.013566)	(0.013902)	(0.013457)	(0.013868)
MIDWEST	-0.0568***	-0.0568***	-0.0540***	-0.0540***
	(0.014031)	(0.013697)	(0.013836)	(0.013703)
WEST	0.0001	0.0001	0.0008	0.0008
	(0.014321)	(0.014492)	(0.014148)	(0.014465)
N	9799	9799	9799	9799
SSE	2097.192	2097.192	2044.156	2044.156
K-1	11	11	11	11
N-K	9787	9787	9787	9787
$\hat{\sigma}$	.4629076	.4629076	.4570169	.4570169

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(d) From column (3) of Table XR8.23b, the estimated skedastic function is

$$h = \exp\left(0.0511EDUC - 0.2316UNION + 0.1301METRO\right)$$

The FGLS estimates are reported in column (3) of Table XR8.23a. The OLS estimates with robust standard errors are in column (2) of Table XR 8-23b. We see that the FGLS estimate of the coefficient of *UNION* is slightly larger than the OLS estimate, 0.0818 compared to

0.0719. The FGLS estimate's standard error is slightly smaller than the OLS robust standard error, 0.012826 compared to 0.013334. The sample is so large the differences are negligible.

While not required, the FGLS estimates with robust standard errors are given in column (4) of Table XR823a.

Table XR8.23b

	(1) BP1	(2) BP2	(3) Mult
C	0.1866^{***}	0.0480^{*}	-3.6819^{***}
	(0.008705)	(0.020190)	(0.124355)
UNION	-0.0264^{*}	-0.0314^{**}	-0.2316^{***}
	(0.010874)	(0.010863)	(0.066907)
METRO	0.0380^{***}	0.0301^{**}	0.1301^{*}
	(0.009545)	(0.009574)	(0.058971)
EDUC		0.0103^{***}	0.0511^{***}
		(0.001348)	(0.008305)
N	9799	9799	9799
R^2	.002161	.0080151	.0056161

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

EXERCISE 8.25

(a) The coefficient estimates are in column (1) of Table XR8.25. The estimated coefficient of *PUBLIC* is 0.8982 with standard error 0.5419. The *t*-value is 1.65747. For a one tail test of significance at the 1% level the *t*-critical value, from Statistical Table 2, is 2.326, so that we cannot conclude that the coefficient on public is positive.

(b) Regressing the squared residuals on all the variables we obtain an $F = 1.09$ with a *p*-value of 0.3673. Thus, at the 1% level, we cannot reject the null hypothesis of homoskedasticity.

(c) The estimates are in Table XR8.25, columns (2) and (3). Equation (7.37) for a model with a single regressor is $\hat{\tau}_{ATE} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\bar{x} = (\hat{\alpha}_1 + \hat{\beta}_1\bar{x}) - (\hat{\alpha}_0 + \hat{\beta}_0\bar{x}) = \hat{y}_1 - \hat{y}_0$, which is the difference in the fitted values from the two separate regressions evaluated at the variable means. For the regression when $PUBLIC = 1$, the fitted value at the means is 3.080441. When $PUBLIC = 0$, the fitted value is 1.958469, so that the average treatment effect estimate is $\hat{\tau}_{ATE} = 3.080441 - 1.958469 = 1.1219717$.

(d) These estimates are in Table XR8.25, column (4). We have added the letter *D* to the end of the variable name to indicated "deviations." The coefficient on *PUBIC* is unchanged because subtracting the means does not affect slope coefficients. However, note that the robust standard error is smaller. The *t*-value is now 2.7326 which is greater than the 1% critical value 2.326. We conclude that there is a positive association between public assistance and the average number of doctor visits.

Table XR8.25

	(1) ols	(2) public=1	(3) public=0	(4) part(d)	(5) part(e)
C	4.0479***	5.1819***	2.7108	2.9867***	2.7108*
	(1.220396)	(1.123720)	(1.468209)	(0.157462)	(1.265039)
PUBLIC	0.8982				1.1220**
	(0.541906)				(0.368714)
FEMALE	0.3444	0.3681	0.0668		0.0668
	(0.320736)	(0.355009)	(0.537651)		(0.490364)
HHKIDS	-1.0387**	-1.1407**	-0.3865		-0.3865
	(0.344368)	(0.382255)	(0.579305)		(0.570940)
MARRIED	0.2711	0.2655	0.5134		0.5134
	(0.398090)	(0.443093)	(0.656534)		(0.492424)
SELF	-0.6360	-0.7477	-0.3936		-0.3936
	(0.707052)	(0.878614)	(0.766199)		(0.700470)
EDUC2	-0.1074	-0.1131	-0.0733		-0.0733
	(0.075629)	(0.089818)	(0.091953)		(0.093532)
HHNINC2	-0.0002	-0.0002	-0.0000		-0.0000
	(0.000103)	(0.000122)	(0.000120)		(0.000099)
PUBLICD				0.8982**	
				(0.328696)	
FEMALED				0.3444	
				(0.311839)	
HHKIDSD				-1.0387**	
				(0.327955)	
MARRIEDD				0.2711	
				(0.470102)	
SELFD				-0.6360	
				(0.552749)	
EDUC2D				-0.1074	
				(0.062106)	
HHNINC2D				-0.0002	
				(0.000086)	
FEMALEP					0.3014
					(0.599148)
HHKIDSP					-0.7541
					(0.677790)
MARRIEDP					-0.2480
					(0.720635)
SELFP					-0.3540
					(1.013865)
EDUC2P					-0.0398
					(0.120072)
HHNINC2P					-0.0002
					(0.000145)
N	1200	1063	137	1200	1200
SSE	35465.61	34239.66	1182.555	35465.61	35422.22
K-1	7	6	6	7	13
N-K	1192	1056	130	1192	1186
$\hat{\sigma}$	5.454634	5.694201	3.016054	5.454634	5.465068

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(e) The estimates are in column (5) of Table XR8.25. We have added a *P* to the end of each variable name to indicate that the variable is interacted with *PUBLIC.* The average treatment effect is the coefficient of the indicator variable *PUBLIC.* An advantage of this approach is that we can estimate the average treatment effect and its standard error at the same time. The *t*-value is 3.0429 which is significant at the 1% level, and thus for the 5% level, for a one-tail test.

EXERCISE 8.27

(a) The estimates are in Table XR8.27 column (1). In this linear-log model, we estimate a 1% increase in population will increase the expected number of medals won by 0.027643. Similarly, for a 1% increase in GDP, we estimate an increase in the expected number of medals won by 0.042705.

Table XR8.27

	(1) ols	(2) LM test	(3) ols rob
C	-13.1527*	-482.7027	-13.1527
	(5.974395)	(318.273535)	(6.709098)
LPOP	2.7643	133.5969	2.7643
	(2.069901)	(110.269713)	(1.789073)
LGDP	4.2705*	112.3777	4.2705*
	(1.717726)	(91.508324)	(1.779734)
N	63	63	63
R^2	.2753694	.1320799	.2753694

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) The regression of the squared residuals on the variables is in column (2) of Table XR8.27. The *F*-test statistic is 4.57, with a *p*-value of 0.0143. Using this test we reject the null hypothesis of homoskedasticity at the 5% level. The value of the LM statistic is $NR^2 = 63(0.1321) = 8.321036$. Using the $\chi^2_{(2)}$ distribution the *p*-value is 0.01559948. Using this test we reject the null hypothesis of homoskedasticity at the 5% level.

(c) The estimates with robust standard errors are in column (3) of Table XR8.27. The estimated coefficient is 4.270461 with robust standard error 1.779734, yielding a *t*-value of 2.399494. For a right tail test the 10% critical value is $t_{(0.90,60)} = 1.2958211$. The 5% critical value is $t_{(0.95,60)} = 1.6706489$. Thus, we conclude that there is a positive relationship between GDP and the expected number of medals won at either the 5% or 10% level of significance.

(d) The estimated coefficient is 2.764299 with robust standard error 1.789073, yielding a *t*-value of 1.5451. Using the test critical values from part (c), we conclude that there is a positive relationship between the population size and the expected number of medals won at the 10% level of significance, but not at the 5% level.

(e) The estimated expected number of medals won is

$$E\left(MEDALS \mid GDP=1010, POP=58\right) = -13.15271 + 2.764299\ln(58) + 4.270461\ln(1010)$$
$$= 27.61335$$

The standard error of the linear combination of the OLS coefficients is 6.756285. Using critical value $t_{(.975,60)} = 2.0002978$, the 95% interval estimate is [14.09877, 41.12794].

(f) Because 20 is inside the interval estimate from part (d), we will not reject the null hypothesis using a two-tail test at the 5% level. The *p*-value is 0.2643.

EXERCISE 8.29

(a) The estimates are in column (1) of Table XR8.29.

Table XR8.29

	(1) OLS	(2) attr=1	(3) attr=0	(4) part(e)	(5) part(f)
C	5.7525^{***}	7.1433^{***}	5.3625^{***}	5.3625^{***}	5.3625^{***}
	(0.0913)	(0.2567)	(0.0963)	(0.0989)	(0.0877)
BAR	0.2161^{***}	-0.5505^{***}	0.4979^{***}	0.4979^{***}	0.4979^{***}
	(0.0786)	(0.1973)	(0.0841)	(0.0864)	(0.0792)
STREET	-0.2621^{***}	-1.1424^{***}	0.0242	0.0242	0.0242
	(0.0794)	(0.2333)	(0.0842)	(0.0864)	(0.0771)
SCHOOL	0.1638^{***}	0.4518^{***}	0.0939^{***}	0.0939^{***}	0.0939^{***}
	(0.0238)	(0.0667)	(0.0248)	(0.0255)	(0.0252)
AGE	-0.0210^{***}	-0.0411^{***}	-0.0177^{***}	-0.0177^{***}	-0.0177^{***}
	(0.0015)	(0.0050)	(0.0015)	(0.0015)	(0.0013)
RICH	0.2924^{***}	0.6096^{***}	0.2599^{***}	0.2599^{***}	0.2599^{***}
	(0.0304)	(0.1227)	(0.0303)	(0.0311)	(0.0296)
ALCOHOL	0.2403^{***}	-0.1027	0.3220^{***}	0.3220^{***}	0.3220^{***}
	(0.0358)	(0.1158)	(0.0366)	(0.0375)	(0.0386)
ATTRACTIVE	0.2394^{***}			1.7808^{***}	1.7808^{***}
	(0.0316)			(0.2438)	(0.2444)
ABAR				-1.0484^{***}	-1.0484^{***}
				(0.1918)	(0.1580)
ASTREET				-1.1666^{***}	-1.1666^{***}
				(0.2201)	(0.2194)
ASCHOOL				0.3579^{***}	0.3579^{***}
				(0.0632)	(0.0721)
AAGE				-0.0234^{***}	-0.0234^{***}
				(0.0046)	(0.0047)
ARICH				0.3497^{***}	0.3497^{**}
				(0.1109)	(0.1391)
AALCOHOL				-0.4247^{***}	-0.4247^{***}
				(0.1073)	(0.1180)
N	3016	416	2600	3016	3016
R^2	.3250073	.3361243	.3313034	.3570282	.3570282
SSE	1041.964	179.5834	812.9513	992.5347	992.5347
$\hat{\sigma}$	.5885556	.6626305	.5599265	.5749996	.5749996

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(a) When clients are met in a bar, we estimate expected price to be approximately 21.61% higher; when they are met in the street we estimate expected price to be approximately 26.21% lower; when the sex worker has completed secondary school we estimate the expected price will be 16.38% higher; with each additional year of age we estimate an expected price that is 2.10% lower; if the client is rich we estimate the expected price will be 29.24% higher; if the client had consumed alcohol we estimate the expected price will be 24.03% higher; and if the sex worker is attractive we estimate a 23.94% higher expected price. All the coefficients are significant at the 1% level or better.

(b) The $R^2 = 0.0136$ from the regression of the squared residuals on *ATTRACTIVE*. There are N = 3,016 observations, so that $NR^2 = 41.129269$. The 99th percentile of the chi-square distribution with one degree of freedom is 6.6348966, so we reject the null hypothesis of homoskedasticity.

(c) These regressions are in columns (2) and (3) of Table XR8.29. The Goldfeld-Quandt test statistic is $GQ = \hat{\sigma}_1^2 / \hat{\sigma}_0^2 = 1.400$, where the subscript 1 indicates that *ATTRACTIVE* = 1 and the subscript 0 that *ATTRACTIVE* = 0. The 2.5 and 97.5 percentiles of the $F_{(409,2593)}$ distribution are 0.85929 and 1.15445, respectively. Thus, at the 5% level, we reject the null hypothesis of homoskedasticity based on this test.

(d) There are noticeable differences in the coefficient estimates for *BAR* (sign and magnitude), *STREET* (magnitude), *SCHOOL* (magnitude), *AGE* (magnitude), *RICH* (magnitude), and *ALCOHOL* (magnitude). The regression including the original variables and variables interacted with *ATTRACTIVE* are in column (4) of Table XR8.29. To indicate the interaction variables we have preceded the variable name with "*A*", such as *ASCHOOL*. This is a valid Chow test under homoskedasticity. The value of the joint test statistic for the coefficients of attractive and all the interaction variables is F = 29.95. The 99th percentile of the *F*-distribution with 7 numerator and 3002 denominator degrees of freedom is 2.643, so we reject the null hypothesis that the regression coefficients for attractive sex workers are the same as workers not deemed attractive.

(e) The regression $R^2 = 0.0046$ and the LM test statistic value is 13.748, which has a *p*-value of 0.0002. Thus we reject homoskedasticity and this casts doubt on our result in part (d).

(f) The regression with robust standard errors is in Table XR8.29 column (5). The standard errors are somewhat different, with some being larger and some being smaller. The value of the joint test statistic of *ATTRACTIVE* and all the interaction variables is 36.98. The 99th percentile of the *F*-distribution with 7 numerator and 3002 denominator degrees of freedom is 2.643, so we reject the null hypothesis that the regression coefficients for attractive sex workers are the same as workers not deemed attractive. This *F*-test is valid in large samples, but it cannot be represented as a Chow test that uses restricted and unrestricted sums of squared errors.