Chapter Review

Problems

1. Let *kids* denote the number of children ever born to a woman, and let *educ* denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = \beta_0 + \beta_1 educ + u,$$

where *u* is the unobserved error.

   i. What kinds of factors are contained in *u*? Are these likely to be correlated with level of education?

   ii. Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

2. In the simple linear regression model $y = \beta_0 + \beta_1 x + u$, suppose that $\mathbf{E}(u) \neq 0$. Letting $\alpha_0 = \mathbf{E}(u)$, show that the model can always be rewritten with the same slope, but a new intercept and error, where the new error has a zero expected value.

3. The following table contains the *ACT* scores and the *GPA* (grade point average) for eight college students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

Student	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25

Student	*GPA*	*ACT*
7	2.7	25
8	3.7	30

i. Estimate the relationship between *GPA* and *ACT* using OLS; that is, obtain the intercept and slope estimates in the equation

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.$$

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the *GPA* predicted to be if the *ACT* score is increased by five points?

ii. Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

iii. What is the predicted value of *GPA* when $ACT = 20$?

iv. How much of the variation in *GPA* for these eight students is explained by *ACT*? Explain.

4. The data set BWGHT contains data on births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (*bwght*), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (*cigs*). The following simple regression was estimated using data on $n = 1{,}388$ *births*:

$$\widehat{bwght} = 119.77 - 0.514\, cigs$$

i. What is the predicted birth weight when $cigs = 0$? What about when $cigs = 20$ (one pack per day)? Comment on the difference.

ii. Does this simple regression necessarily capture a causal relationship between the child's birth weight and the mother's smoking habits? Explain.

iii. To predict a birth weight of 125 ounces, what would *cigs* have to be? Comment.

iv. The proportion of women in the sample who do not smoke while pregnant is about .85. Does this help reconcile your finding from part (iii)?

5. In the linear consumption function

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1 inc,$$

the (estimated) *marginal propensity to consume* (MPC) out of income is simply the slope, $\hat{\beta}_1$, while the *average propensity to consume* (APC) is $\widehat{cons}/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. Using observations for 100 families on annual income and consumption (both measured in dollars), the following equation is obtained:

$$\begin{aligned} \widehat{cons} &= -124.84 + 0.853\ inc \\ n &= 100, R^2 = 0.692. \end{aligned}$$

i. Interpret the intercept in this equation, and comment on its sign and magnitude.

ii. What is the predicted consumption when family income is $30,000?

iii. With *inc* on the *x*-axis, draw a graph of the estimated MPC and APC.

6. Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\begin{aligned} \widehat{\log(price)} &= 9.40 + 0.312 \log(dist) \\ n &= 135, R^2 = 0.162. \end{aligned}$$

i. Interpret the coefficient on $\log(dist)$. Is the sign of this estimate what you expect it to be?

ii. Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

iii. What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

7. Consider the savings function

$$sav = \beta_0 + \beta_1 inc + u, u = \sqrt{inc} \cdot e,$$

where *e* is a random variable with $\mathbf{E}(e) = 0$ and $\mathbf{Var}(e) = \sigma_e^2$. Assume that *e* is independent of *inc*.

i. Show that $\mathbf{E}(u|inc) = 0$, so that the key zero conditional mean assumption (Assumption SLR.4) is satisfied. [*Hint:* If *e* is independent of

inc, then $\mathbf{E}(e|inc) = \mathbf{E}(e)$.]

ii. Show that $\mathbf{Var}(u|inc) = \sigma_e^2 inc$, so that the homoskedasticity Assumption SLR.5 is violated. In particular, the variance of *sav* increases with *inc*. [*Hint:* $\mathbf{Var}(e|inc) = \mathbf{Var}(e)$ if *e* and *inc* are independent.]

iii. Provide a discussion that supports the assumption that the variance of savings increases with family income.

8. Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1, SLR.2, SLR.3, SLR.4 and SLR.5. The usual OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero (see Section 2-6).

i. Find $\mathbf{E}(\tilde{\beta}_1)$ in terms of the x_i, β_0, and β_1. Verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept (β_0) is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

ii. Find the variance of $\tilde{\beta}_1$. (*Hint:* The variance does not depend on β_0.)

iii. Show that $\mathbf{Var}(\tilde{\beta}_1) \leq \mathbf{Var}(\widehat{\beta}_1)$. [*Hint:* For any sample of data, $\sum_{i-1}^{n} x_i^2 \geq \sum_{i-1}^{n} (x_i - \overline{x})^2$, with strict inequality unless $\overline{x} = 0$.]

iv. Comment on the tradeoff between bias and variance when choosing between $\widehat{\beta}_1$ and $\tilde{\beta}_1$.

9. i. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be the intercept and slope from the regression of y_i on x_i, using *n* observations. Let c_1 and c_2, with $c_2 \neq 0$, be constants. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $c_1 y_i$ on $c_2 x_i$. Show that $\tilde{\beta}_1 = (c_1/c_2)\widehat{\beta}_0$ and $\tilde{\beta}_0 = c_1 \widehat{\beta}_0$, thereby verifying the claims on units of measurement in Section 2-4. [*Hint:* To obtain $\tilde{\beta}_1$, plug the scaled versions of *x* and *y* into (2.19). Then, use (2.17) for $\tilde{\beta}_0$, being sure to plug in the scaled *x* and *y and* the correct slope.]

ii. Now, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be from the regression of $(c_1 + y_i)$ on $(c_2 + x_i)$ (with no restriction on c_1 or c_2). Show that $\tilde{\beta}_1 = \widehat{\beta}_1$ and $\tilde{\beta}_0 = \widehat{\beta}_0 + c_1 - c_2 \widehat{\beta}_1$.

iii. Now, let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be the OLS estimates from the regression $\log(y_i)$ on x_i, where we must assume $y_i > 0$ for all *i*. For $c_1 > 0$, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $\log(c_1 y_i)$ on x_i. Show that $\tilde{\beta}_1 = \widehat{\beta}_1$ and $\tilde{\beta}_0 = \log(c_1) + \widehat{\beta}_0$.

iv. Now, assuming that $x_i > 0$ for all *i*, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of y_i on $\log(c_2 x_i)$. How do $\tilde{\beta}_0$ and $\tilde{\beta}_1$ compare with the intercept and slope from the regression of y_i on $\log(x_i)$?

10. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be the OLS intercept and slope estimators, respectively, and let $\overline{u}$ be the sample average of the errors (not the residuals!).

    i. Show that $\widehat{\beta}_1$ can be written as $\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$, where $w_i = d_i / \mathrm{SST}_x$ and $d_i = x_i - \overline{x}$.

    ii. Use part (i), along with $\sum_{i=1}^{n} w_i = 0$, to show that $\widehat{\beta}_1$ and $\overline{u}$ are uncorrelated. [*Hint*: You are being asked to show that $\mathrm{E}[(\widehat{\beta}_1 - \beta_1) \cdot \overline{u}] = 0$.]

    iii. Show that $\widehat{\beta}_0$ can be written as $\widehat{\beta}_0 = \beta_0 + \overline{u} - (\widehat{\beta}_1 - \beta_1)\overline{x}$.

    iv. Use parts (ii) and (iii) to show that $\mathrm{Var}(\widehat{\beta}_0) = \sigma^2/n + \sigma^2(\overline{x})^2/\mathrm{SST}_x$.

    v. Do the algebra to simplify the expression in part (iv) to equation (2.58).

    [*Hint*: $\mathrm{SST}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$.]

11. Suppose you are interested in estimating the effect of hours spent in an SAT preparation course (*hours*) on total SAT score (*sat*). The population is all college-bound high school seniors for a particular year.

    i. Suppose you are given a grant to run a controlled experiment. Explain how you would structure the experiment in order to estimate the causal effect of *hours* on *sat*.

    ii. Consider the more realistic case where students choose how much time to spend in a preparation course, and you can only randomly sample *sat* and *hours* from the population. Write the population model as

    $$sat = \beta_0 + \beta_1 hours = u$$

    where, as usual in a model with an intercept, we can assume $\mathrm{E}(u) = 0$. List at least two factors contained in *u*. Are these likely to have positive or negative correlation with *hours*?

    iii. In the equation from part (ii), what should be the sign of β_1 if the preparation course is effective?

    iv. In the equation from part (ii), what is the interpretation of β_0?

12. Consider the problem described at the end of Section 2-6: running a regression and only estimating an intercept.

    i. Given a sample $\{y_i: i = 1, 2, \ldots, n\}$, let $\tilde{\beta}_0$ be the solution to

    $$\min_{b_0} \sum_{i=1}^{n} (y_i - b_0)^2.$$

    Show that $\tilde{\beta}_0 = \overline{y}$, that is, the sample average minimizes the sum of squared residuals. (*Hint*: You may use one-variable calculus or you can show the result directly by adding and subtracting $\overline{y}$ inside the squared residual and then doing a little algebra.)

    ii. Define residuals $\tilde{u}_i = y_i - \overline{y}$. Argue that these residuals always sum to zero.