Fitted Value
Gauss-Markov Assumptions
Heteroskedasticity
Homoskedasticity
Independent Variable
Intercept Parameter
Mean Independent
OLS Regression Line
Ordinary Least Squares (OLS)
Population Regression Function (PRF)
Predicted Variable
Predictor Variable
Regressand
Regression through the Origin
Regressor
Residual
Residual Sum of Squares (SSR)
Response Variable
R-squared
Sample Regression Function (SRF)
Semi-elasticity
Simple Linear Regression Model
Slope Parameter
Standard Error of $\hat{\beta}_1$
Standard Error of the Regression (SER)
Sum of Squared Residuals (SSR)
Total Sum of Squares (SST)
Zero Conditional Mean Assumption

PROBLEMS

2.1 In the simple linear regression model $y = \beta_0 + \beta_1 x + u$, suppose that $E(u) \neq 0$. Letting $\alpha_0 = E(u)$, 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.

2.2 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
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.

2.3 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.

2.4 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 $E(u) = 0$. 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?

2.5 Consider the savings function

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

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

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

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

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

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

(i) Show that $\hat{\beta}_1$ can be written as $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$ where $w_i = d_i/SST_x$ and $d_i = x_i - \bar{x}$.

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

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

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

(v) Do the algebra to simplify the expression in part (iv) to equation (2.58). [*Hint*: $\text{SST}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$.]

2.7 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*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$

$$n = 135, R^2 = 0.162.$$

(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?

2.8 (i) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ 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 $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_0$ and $\tilde{\beta}_0 = c_1\hat{\beta}_0$, 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 $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2\hat{\beta}_1$.

(iii) Now, let $\hat{\beta}_0$ and $\hat{\beta}_1$ 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 $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \log(c_1) + \hat{\beta}_0$.

(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 $\log(x_i)$?

2.9 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:

$$\widehat{cons} = -124.84 + 0.853\, inc$$

$$n = 100, R^2 = 0.692.$$

(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.

2.10 Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ 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 $E(\tilde{\beta}_1)$ 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 $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$. [*Hint:* For any sample of data, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, with strict inequality unless $\bar{x} = 0$.]

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

2.11 The data set BWGHT.RAW 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)?

COMPUTER EXERCISES

C2.1 The data in 401K.RAW are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if $mrate = 0.50$, then a \$1 contribution by the worker is matched by a 50¢ contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

(ii) Now, estimate the simple regression equation

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate,$$

and report the results along with the sample size and *R*-squared.

(iii) Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

(iv) Find the predicted *prate* when $mrate = 3.5$. Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in *prate* is explained by *mrate*? Is this a lot in your opinion?

C2.2 The data set in CEOSAL2.RAW contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, *ceoten* = 0)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

C2.3 Use the data in SLEEP75.RAW from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

(i) Report your results in equation form along with the number of observations and R^2. What does the intercept in this equation mean?

(ii) If *totwrk* increases by 2 hours, by how much is *sleep* estimated to fall? Do you find this to be a large effect?

C2.4 Use the data in WAGE2.RAW to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

(i) Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation equal to 15.)

(ii) Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in wage for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*?

(iii) Now, estimate a model where each one-point increase in *IQ* has the same percentage effect on *wage*. If *IQ* increases by 15 points, what is the approximate percentage increase in predicted *wage*?

C2.5 For the population of firms in the chemical industry, let *rd* denote annual expenditures on research and development, and let *sales* denote annual sales (both are in millions of dollars).

(i) Write down a model (not an estimated equation) that implies a constant elasticity between *rd* and *sales*. Which parameter is the elasticity?

(ii) Now, estimate the model using the data in RDCHEM.RAW. Write out the estimated equation in the usual form. What is the estimated elasticity of *rd* with respect to *sales*? Explain in words what this elasticity means.

C2.6 We used the data in MEAP93.RAW for Example 2.12. Now we want to explore the relationship between the math pass rate (*math10*) and spending per student (*expend*).

(i) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain.

(ii) In the population model

$$math10 = \beta_0 + \beta_1 \log(expend) + u,$$

argue that $\beta_1/10$ is the percentage point change in *math10* given a 10% increase in *expend*.

(iii) Use the data in MEAP93.RAW to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and *R*-squared.

(iv) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

(v) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this data set?

C2.7 Use the data in CHARITY.RAW [obtained from Franses and Paap (2001)] to answer the following questions:

(i) What is the average gift in the sample of 4,268 people (in Dutch guilders)? What percentage of people gave no gift?

(ii) What is the average mailings per year? What are the minimum and maximum values?

(iii) Estimate the model

$$gift = \beta_0 + \beta_1 mailsyear + u$$

by OLS and report the results in the usual way, including the sample size and *R*-squared.

(iv) Interpret the slope coefficient. If each mailing costs one guilder, is the charity expected to make a net gain on each mailing? Does this mean the charity makes a net gain on every mailing? Explain.

(v) What is the smallest predicted charitable contribution in the sample? Using this simple regression analysis, can you ever predict zero for *gift*?

Appendix 2A

Minimizing the Sum of Squared Residuals

We show that the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ do minimize the sum of squared residuals, as asserted in Section 2.2. Formally, the problem is to characterize the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ to the minimization problem

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0' - b_1 x_i)^2,$$

where b_0 and b_1 are the dummy arguments for the optimization problem; for simplicity, call this function $Q(b_0, b_1)$. By a fundamental result from multivariable calculus (see Appendix A), a necessary condition for $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the minimization problem is that the partial derivatives of $Q(b_0, b_1)$ with respect to b_0 and b_1 must be zero when evaluated at

$\hat{\beta}_0, \hat{\beta}_1$: $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_0 = 0$ and $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_1 = 0$. Using the chain rule from calculus, these two equations become

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

$$-2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

These two equations are just (2.14) and (2.15) multiplied by $-2n$ and, therefore, are solved by the same $\hat{\beta}_0$ and $\hat{\beta}_1$.

How do we know that we have actually minimized the sum of squared residuals? The first order conditions are necessary but not sufficient conditions. One way to verify that we have minimized the sum of squared residuals is to write, for any b_0 and b_1,

$$Q(b_0, b_1) = \sum_{i=1}^{n}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2$$

$$= \sum_{i=1}^{n}[\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2$$

$$= \sum_{i=1}^{n}\hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2\sum_{i=1}^{n}x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1)\sum_{i=1}^{n}x_i,$$

where we have used equations (2.30) and (2.31). The first term does not depend on b_0 or b_1, while the sum of the last three terms can be written as

$$\sum_{i=1}^{n}[(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2,$$

as can be verified by straightforward algebra. Because this is a sum of squared terms, the smallest it can be is zero. Therefore, it is smallest when $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$.