

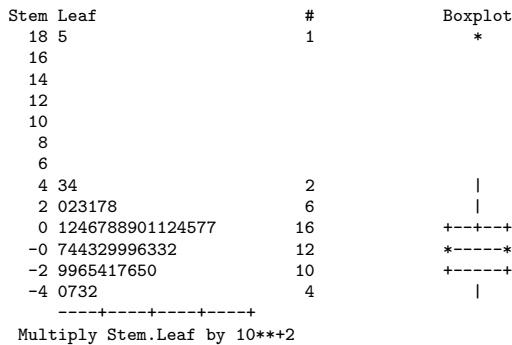
STA 6127: Solutions of Exercises for Chapter 14

- 2a. Percent with a high school education is not significant, controlling for the other four predictors. On its own, it is highly significant. The other four predictors together explain almost all the variability in murder rate that high school education explained.
- b. Only percent with high school education is removed.
- c. All variables are entered (order SI, HS, ME, WH, PO).
- d. After all five variables are entered, percent high school graduates is no longer significant and is removed.
- e. Stepwise and backward are identical. Forward also includes HS, which explains a lot of variability by itself or with SI alone in the model, but is unneeded when all four other predictors are in the model.

8. The variability of the residuals seems to increase as the predicted value increases. This suggests that the variance is higher when income is higher, in which case ordinary least squares does not provide the best fit. (See, for instance, the subsection of Section 14.5 on models assuming a gamma distribution.)

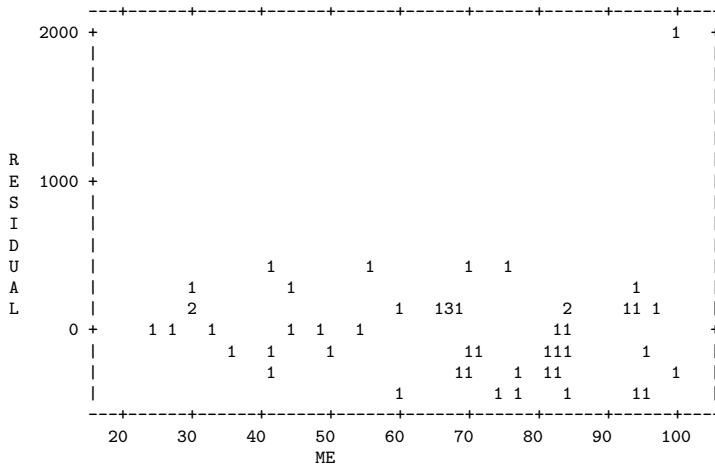
- 10b. The variability may tend to be greater when the predicted birth rate is higher.
- c. The studentized residuals for observations 11 and 12 stand out a bit, but it is not unusual to get a couple larger than 2 in absolute value in a sample of this size.
- d. $p/n = 3/23 = .13$, and observations 6 and 17 have values more than double this.
- e. Not particularly, since observations with large studentized residuals do not have overly large hat values, and observations with large hat values do not have overly large studentized residuals.
- f. Observations 6, 11, and 17.
- g. Observation 11.

12a. $\hat{Y} = -123.7 + 10.9X$.



There is one obvious outlier (D. C.).

b.



There is an obvious outlier at a high level of the predictor (This is D. C.).
c. $\hat{Y} = 25.6 + 8.1X$.

- 13a. D. C. has a studentized residual of 8.4.
b. No extreme values; Montana has .098.
c. D. C. since its studentized residual is so large, and its hat value (.06) is above average.
d. D. C. has DFFITS = 2.2, much larger than others.
e. D. C. has DFBETAS = -1.4 for intercept and 1.82 for coefficient of percent metropolitan, both much larger than for other variables.
f. Removing D. C. line changes from $\hat{Y} = -123.7 + 10.9X$ to $\hat{Y} = 25.6 + 8.1X$.

18a. No. Both X_1 and X_2 would be highly positively correlated with Y . However, $r_{X_1X_2}$ would be very close to 1.0, and once one of them is in the model, the other would be redundant for practical purposes. Thus, the partial effects might well be nonsignificant.
b. No, this test would probably have a very small P -value, since R^2 , like either r_{YX_1} or r_{YX_2} , would be large.
c. Select the model using X_2 alone as the predictor. It would be selected first, since it is most highly correlated with Y , but once it is in the model, X_1 would not make a significant contribution toward explaining additional variation, so it would not be added to the model.

21a. Convex (bowl-shaped), continuously increasing over the positive set of values for S .
b. (i) 54.0, (ii) 123.6, (iii) 206.6; changes in S^2 increase as S increases.
c. The minimum occurs at $S = -49.34/2(6.74) = -3.7$; \hat{Y} increases as S increases from -3.7, in particular over all positive S values.
d. The linear association is strong, since r^2 is large. The degree of nonlinearity is minor, since R^2 for the quadratic model is only slightly larger than r^2 for the linear model.

22e. (i) $2.63 + 16.7(2) + 4.168(2)^2 = 52.7$, (ii) 90.3, (iii) 136.1
f. This is caused by multicollinearity based on the very strong correlation between the predictor and its square. R^2 is only slightly higher for the quadratic model (.353 vs. .348), and the improvement is not significant. Thus, to assess the impact of number of bedrooms, one should drop the squared term from the model before doing the t test.

24a. $\hat{Y} = 35.88 - .266X$ (excluding Germany observation).
b. $\hat{Y} = 38.22 - .390X + .00133X^2$, a convex (bowl-shaped) function that is decreasing with slope $-.39$ when it crosses the Y -axis.
c. $X = .390/2(.00133) = 146$; yes, it is decreasing over the entire range.
d. No, the quadratic model does not provide an improved fit. Explained variance increases from $r^2 = .273$ only to $R^2 = .276$ (adjusted R^2 , discussed in Exercise 11.51, actually decreases from .241 to .210), and the t test for the quadratic effect has a P -value of .76, indicating that linearity is plausible. The inferential analysis is less relevant here, since the observations are not a random sample.

25. Excluding Germany, $\log(\hat{\mu}) = 3.6545 - .0112X$, or $\hat{Y} = 38.65(.9889)^X$. When X increases 10 units, \hat{Y} multiplies by $(.9889)^{10} = .89$; i.e., there is a 11% reduction in predicted birth rate.

28a. (iv) is the correct response, since the function increases up to its maximum value at $X_1 = .2/2(.001) = 100$.
b. (iii) is correct, since the slope for X_1 decreases from .07 when $X_2 = 0$ to .01 when $X_2 = 100$.

48a. $1000 * (1.10)^x$
b. approximately 8 years, since $(1.1)^8 = 2.1$; more precisely, setting $(1.1)^X = 2.0$ yields the solution $X = (\log 2)/(\log 1.1) = 7.3$.

- 49a. $(1.042)^{10}100,000 = 150,896$.
b. 50.9% growth over the decade.

50a. A 1.3% growth rate per year corresponds to a multiplicative effect after 10 years of $(1.013)^{10} = 1.138$, or 13.8%. Or, to find the yearly multiplicative factor corresponding to a 10-year multiplicative effect of 1.1377, set $1.1377 = \beta^{10}$, and solve for β ; then, $\log(1.1377) = 10 \log(\beta)$, so $\log(\beta) = \log(1.1377)/10 = .0129$, and $\beta = e^{.0129} = 1.013$.
b. If the growth rate is 1.3% per year, then after X years, the multiplicative effect is $(1.013)^X$.

52. b, c, d.

53. b, c, d.

55. a, d

56a. True, b. False, c. False.

57. c, e, g, j, a, i, k, h