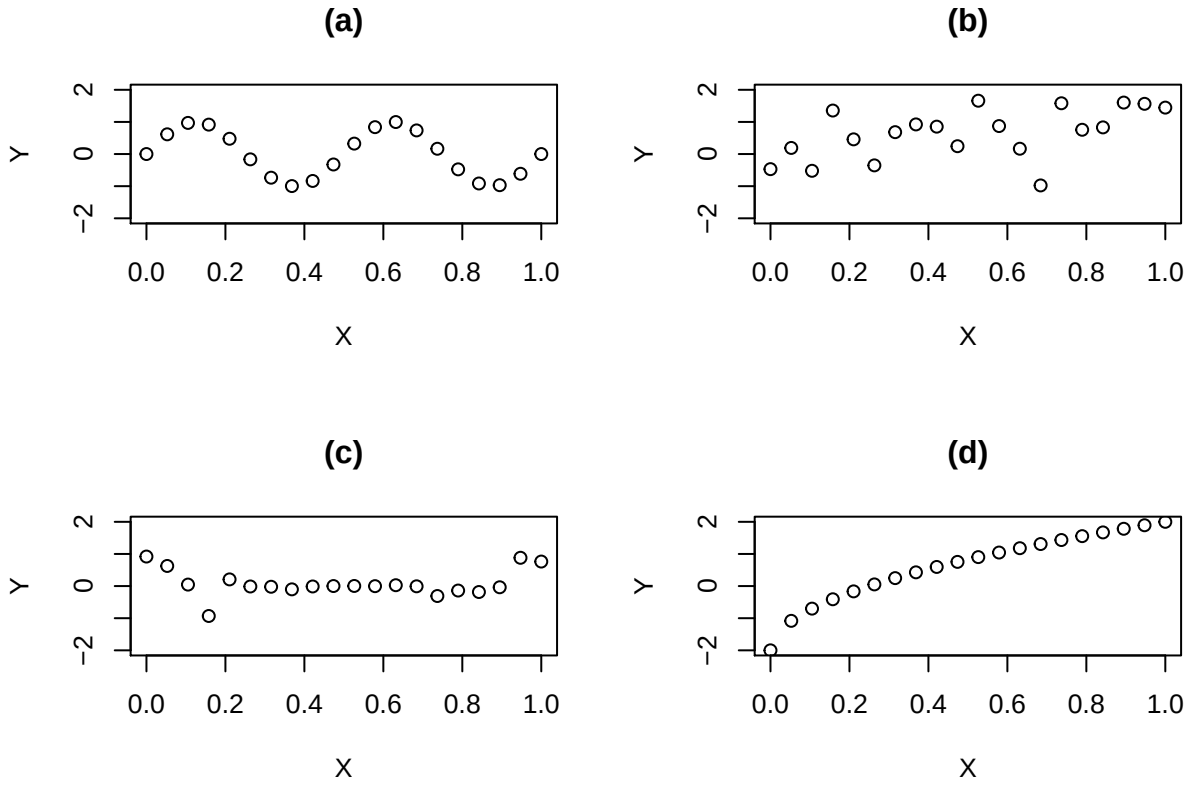
Homework 2: Solutions

Due: Thursday 2/13/20 by 8:30am

Rubric:

- Maximum of 2 points each for 1., 2., and 3., determined as follows:
  - 0 points for no solutions whatsoever or incomplete solutions;
  - 1 point for solutions provided for each part, but at least one incorrect solution;
  - 2 points for correct solutions to each part;
- Maximum of 3 points for 4., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for mostly correct answers but at least one substantial issue;
  - 3 points for nearly/exactly correct.

1. For each of these figures, indicate whether a functional or statistical relationship is depicted between Y and X.



a. Functional
b. Statistical
c. Statistical
d. Functional

2. Suppose we collected data on the age and shoe size of $n = 3$ people. Let Y_i refer to the i-th subject's shoe size, and X_i refer to subject i's age.

(a) Suppose the first two subjects were the same age, i.e. $X_1 = X_2$, but had different shoe sizes, $Y_1 \neq Y_2$. What feature of the simple linear regression model described in Equation 1.1 of the text is illustrated by this?

This illustrates the random errors ϵ_i of the simple linear regression model, which account for deviations of the observed responses Y_1 and Y_2 from the regression function $\beta_0 + \beta_1 X_1 = \beta_0 + \beta_2 X_2$.

(b) Suppose that we only collected data on high school students. If we assume the simple linear regression model described in Equation 1.1 of the text, what is the scope of the model?

The scope of the model is going to be the ages of high school students, from $X_i = 14$, or whatever the age of the youngest student sampled is, to $X_i = 18$, or whatever the age of the oldest student sampled is.

3. Suppose Instagram magically knew that every time the number of times user i purchases a product, denoted by Y_i, is related to the number of times the product has been advertised to user i, denoted by X_i, as follows:

$$Y_i = 1 + 2X_i + \epsilon_i$$

where ϵ_i is a random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = 0.1$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for all $i \neq j$) for $i = 1, \ldots n$.

(a) Which value corresponds to the intercept, β_0? In at most one sentence, interpret it, assuming that the scope of the model includes $X_i = 0$.

The value 1 corresponds to the intercept β_0, which can be interpreted as the expected number of times user i will purchase a product that has not been advertised to them at all.

(b) Which value corresponds to the slope β_1? In at most one sentence, interpret it.

The value 2 corresponds to the slope β_1, which can be interpreted as the expected increase in the number of products user i will purchase if they see one more additional advertisement.

(c) What do you expect the regression function to look like?

I expect that the regression function will look like a linear, increasing function of X_i, the number of times the product has been advertised to user i

(e) Based on the assumed model and the information provided, can we conclude that the number of times user i purchases a product Y_i is uncorrelated with the number of times user j purchases a product Y_j?

Yes.

(f) Based on the assumed model and the information provided, can we state the exact probability that a single value Y_i will be greater than 4 given that $X_i = 1$?

No.

4. Problem 1.28 from the `.pdf` version of the textbook. Requires use of the `crime` data that has been posted on the Homework page.

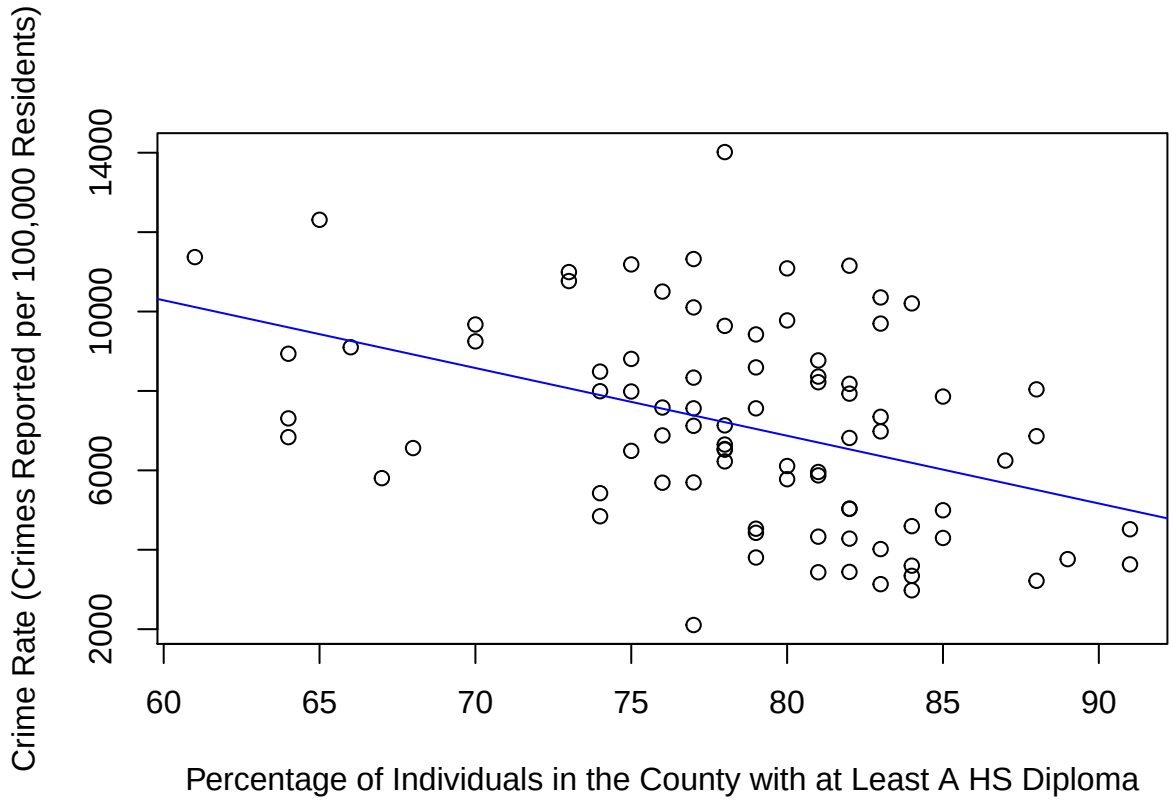
(a)

```
load("~/Dropbox/Teaching/STAT525-2020/stat525/content/homework/crime.RData")
X <- data$X
Y <- data$Y

X.bar <- mean(X)
Y.bar <- mean(Y)

b1 <- sum((X - X.bar)*(Y - Y.bar))/sum((X - X.bar)^2)
b0 <- Y.bar - b1*X.bar
```

```
plot(X, Y, xlab = "Percentage of Individuals in the County with at Least A HS Diploma",
     ylab = "Crime Rate (Crimes Reported per 100,000 Residents)")
abline(a = b0, b = b1, col = "blue")
```



(b)

```
y.hat <- b0 + b1*80
e10 <- Y[10] - b0 - b1*X[10]
s.sq <- sum((Y - b0 - b1*X)^2)/(length(Y) - 2)
```

We obtain (1) $b_1 = -170.58$ for the difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point, (2) $\hat{Y} = 6871.58$ for the mean crime rate last year in counties with high school graduation percentage $X = 80$, (3) $e_{10} = 1401.57$ for an estimate of ϵ_{10}, and (4) $s^2 = 5.5521118 \times 10^6$ for an estimate of σ^2.