Simple Linear Regression Problems

(No submission needed)

STAT-S 520

4/27/23

Instructions:

- You do not need to submit these questions, but you should try to solve them and understand them well before the final exam.
- Solutions to these problems have been shared as well (different file).

Questions:

1.

United Nations (Data file: `UN11` from package `alr4`) The data in the file `UN11` contains several variables, including `ppgdp`, the gross national product per person in U.S. dollars, and `fertility`, the birth rate per 1000 females, both from the year 2009. The data are for 199 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries. The data were collected from United Nations (2011). We will study the dependence of fertility on ppgdp.

a. Identify the predictor and the response.

b. Draw the scatterplot of `fertility` on the vertical axis versus `ppgdp` on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be plausible for a summary of this graph?

c. Draw the scatterplot of `log(fertility)` versus `log(ppgdp)` using natural logarithms. Does the simple linear regression model seem plausible for a summary of this graph?

d. Compute the simple linear regression model corresponding to the graph in part c.

e. Add the fitted line to the graph in part c.

f. Test the hypothesis that the slope is 0 versus the alternative that it is negative (a one-sided test). Give the significance level of the test and a sentence that summarizes the result.

g. Give the value of the coefficient of determination, and explain its meaning.

h. For a locality not in the data with `ppgdp = 1000`, obtain a point prediction and a 95% prediction interval for `log(fertility)`. If the interval `(a, b)` is a 95% prediction interval for `log(fertility)`, then a 95% prediction interval for fertility is given by `(exp(a), exp(b))`. Use this result to get a 95% prediction interval for `fertility`.

i. Obtain the residual plot for model in part d and determine if there are any violations to the assumptions of the model.

2.

We use the data `Sahlins.txt`. You can download this file from our Canvas page (same module as these questions)[1], were compiled by Sahlins (1972) from information presented in Scudder's (1962) report on the Gwenba valley of Central Africa. The data describe agricultural production in Mazulu village. The explanatory variable (Consumers/Gardener) is the ratio of consumers to productive individuals in each of 20 households, making suitable adjustments for the consumption requirements of different household members. The response variable (Acres/Gardener) is a measure of domestic-labor intensity, based on the amount of land cultivated by each household. Think of Consumers/Gardener as representing the relative consumption needs of the household, and Acres/Gardener as representing how hard each productive individual in the household works. Sahlins was interested in production, consumption, and redistribution of the social product in "primitive' ' communities.

a. Draw a scatterplot of Acres/Gardener (Y) versus Consumers/Gardener (X). What relationship, if any, do you discern in this plot –does the relationship appear to be positive or negative (or neither), linear or nonlinear, strong or weak? Is there anything else noteworthy about the data– for example, do any households appear to be unusual?

b. Analyze the data by regressing Acres/Gardener on Consumers/Gardener. In a society characterized by primitive communism, the social product of the village would be redistributed according to need, while each household would work in proportion to its capacity, implying a regression slope of zero. In contrast, in a society in which redistribution is purely through the market, each household should have to work in proportion to its consumption needs, suggesting a positive regression slope and an intercept of zero. Interpret the results of the regression in light of these observations. Examine and interpret the values of $\hat{\beta}_0, \hat{\beta}_1$, $\hat{\sigma}^2$. Do the results change if the fourth household is deleted? Plot the regression lines calculated with and without the fourth household on a scatterplot of the data. Does either regression do a good job of summarizing the relationship between Acres/Gardener and Consumers/Gardener? (see your response in part (a))

c. Find the standard errors of the intercept and slope. Can we conclude that the population slope is greater than zero? Can we conclude that the intercept is greater than zero? Obtain both, confidence intervals and perform hypothesis tests to answer these questions. Use some reasonable significance level (or corresponding confidence levels). Repeat these computations omitting the fourth household. Provide your conclusions for both scenarios.

d. Use the regression coefficients for the entire data (20 households). What do you expect to be the Acres/Gardener ratio for a household with a Consumers/Gardener ratio equal to 1.5. To answer this question, obtain an interval with a 98% confidence level. Would your answer change if instead you are asked to determine the mean Acres/Gardener ratio for all those households with a Consumers/Gardener ratio equal to 1.5? Explain why or why not.

3.

In an 1857 article, the Scottish physicist James D. Forbes (1809–1868) discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water. He knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. Barometers in the middle of the nineteenth century were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure. Forbes collected data in the Alps and in Scotland. He measured at each location the atmospheric pressure `pres` in inches of mercury with a barometer and boiling point `bp` in degrees Fahrenheit using a thermometer. Boiling point measurements were adjusted for the difference between the ambient air temperature when he took the measurements and a standard temperature. The data for `n = 17` locales are reproduced in the file `Forbes`. (package `alr4`).

[1] The data and questions were constructed based on the supplementary material of "Applied Regression Analysis and Generalized Linear Models" 3rd Ed by Fox.

a. Draw the plot of `pres` versus `bp`, and comment relevant observations.

b. Compute the linear regression implied by this problem and interpret the intercept and slope obtained

c. Obtain a plot of residuals against fitted values and discuss your findings. Are there any violations to the model assumptions? If Yes, answer part d. If not, answer directly part e.

d. If violations to the model assumptions are present, try to determine if log transformations help alleviate these problems. Try using transformations of only the response, only the regressor, or both. Find the regressions for each one of these cases, find linear regressions and determine which residual plots are more appropriate.

e. Using the most appropriate model, find and interpret a 97% confidence interval for the slope.

f. Using the most appropriate model. What would be the atmospheric pressure if the boiling point of water is 200 F? Find a 94% prediction interval.

4.

Refer to problem 3. An alternative approach to the analysis of Forbes's experiments comes from the Clausius–Clapeyron formula of classical thermodynamics, which dates to Clausius (1850). According to this theory, we should find that

$$E(\text{pres}|\text{bp}) = \beta_0 + \beta_1 \frac{1}{\text{bpKelvin}}$$

where `bpKelvin` is boiling point in kelvin, which equals `255.37 + (5/9) x bp`. If we were to graph this mean function on a plot of `pres` versus `bpKelvin`, we would get a curve, not a straight line. However, we can estimate the parameters β_0 and β_1 using simple linear regression methods by defining `u1` to be the inverse of temperature in kelvin,

$$u_1 = \frac{1}{\text{bpKelvin}} = \frac{1}{(5/9)\text{bp} + 255.37}$$

and the mean function can be rewritten as

$$E(\text{pres}|\text{bp}) = \beta_0 + \beta_1 u_1$$

for which simple linear regression is suitable. The notation we have used is a little different, as the left side of the equation says we are conditioning on `bp`, but the variable `bp` does not appear explicitly on the right side of the equation, although of course the regressor `u1` depends on `bp`.

a. Draw the plot of `pres` versus `u1`, and verify that apart from case 12 the 17 points in Forbes's data fall close to a straight line. Explain why the apparent slope in this graph is negative when the slope in question 3 was positive.

b. Compute the linear regression and summarize your results.

c. We now have two possible models for the same data based on the regression obtained in 3b and the one obtained in 4b. To compare these two mean functions, draw the plot of the fitted values from 3b to those in 4b. On the basis of this plot, is it possible to prefer one approach over the other? Why?

d. (*This part won't be included in your final exam, but it's still interesting to consider here*) In his original paper, Forbes provided additional data collected by the botanist Joseph D. Hooker (1817–1911) on temperatures and boiling points measured often at higher altitudes in the Himalaya Mountains. The data for `n = 31` locations is given in the file `Hooker`. Find the linear regression similar the one found in 3b but this time using Hooker's data and compare them both. Are the slopes truly different? Use a t-test for comparing to parameters (as a difference) to answer this question and explain your results.